Universidade Federal de Minas Gerais

Instituto de Ciências Biológicas

Departamento de Bioquímica e Imunologia

Programa Interunidades de Pós-Graduação em Bioinformática

Tese de Doutorado

Ferramentas para análise filogenética e de distribuição taxonômica de genes ortólogos

Autor: Tetsu Sakamoto

Orientador: Dr. José Miguel Ortega

BELO HORIZONTE – MG

2016

TETSU SAKAMOTO

Ferramentas para análise filogenética e de distribuição taxonômica de genes ortólogos

Tese apresentada ao Programa Interunidades de Pós-Graduação em Bioinformática do Instituto de Ciências Biológicas da Universidade Federal de Minas Gerais como requisito para a obtenção do título de Doutor em Bioinformática.

Orientador: Dr. José Miguel Ortega

BELO HORIZONTE – MG

2016

043 Sakamoto, Tetsu.
Ferramentas para análise filogenética e de distribuição taxonômica de genes ortólogos [manuscrito] / Tetsu Sakamoto. – 2016.

112 f. : il. ; 29,5 cm.

Orientador: Dr. José Miguel Ortega.
Tese (doutorado) – Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas.

1. Evolução (Biologia). 2. Bioinformática - Teses. 3. Filogenia - Teses. 4. Genômica comparativa. 5. Árvore taxonômica. I. Ortega, José Miguel. II. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. III. Título.

CDU: 575.112

Universidade Federal de Minas Gerais
Instituto de Ciências Biológicas
Programa Interunidades de Pós-Graduação em Bioinformática da UFMG
Avenida Presidente Antônio Carlos, 6627 – Pampulha
31270-901 - Belo Horizonte – MG
Endereço eletrônico: bioinfo@icb.ufmg.br 55 31 3409-2554

"Ferramentas para análise filogenética e de distribuição taxonômica de genes ortólogos"

Tetsu Sakamoto

Tese aprovada pela banca examinadora constituída pelos Professores:

Prof. José Miguel Ortega - Orientador
UFMG

Prof. Francisco Prosdocimi
UFRJ

Prof. Mauro Castro
UFPR

Prof. Lucas Bleicher
UFMG

Prof. Aristóteles Góes Neto
UFMG

Belo Horizonte, 29 de julho de 2016.

À minha amada esposa Carla, com quem compartilho todas as minhas alegrias e que me ajuda a alcançar os patamares mais elevados.

AGRADECIMENTOS

Muitas mãos se empenharam para a realização deste trabalho. Por isso, existem várias pessoas, desde as mais íntimas até as desconhecidas, a quem devo os meus sinceros agradecimentos. Mas não posso deixar de agradecer primeiramente às mãos do D-us Único e sempre presente, pois sem o Seu auxílio e sem a Sua permissão, este projeto de doutorado não teria sido concluído. Hoje eu entendo que foi Ele quem orquestrou os mais variados acontecimentos, livrou-me e me fez encontrar com pessoas por quem sinto gratidão.

Assim, agradeço à D-us que desde muito tempo vem cuidando da minha vida. Agradeço a Ele por toda a base e o suporte para a formação do meu caráter que Ele me forneceu e vem me fornecendo através da minha família, que antes era pequena e se concentrava em São Paulo, mas agora se tornou numerosa e se estende até as terras de Minas Gerais. Agradeço à D-us pela vida da minha esposa, que sempre me apoiou e lutou ao meu lado durante esta jornada. Agradeço a D-us por ter me trazido até Belo Horizonte e mais especificamente até o Laboratório de Biodados - UFMG, onde recebi muitos aprendizados e aperfeiçoamentos na minha formação como pesquisador durante esses quatro anos de doutorado. Agradeço à D-us pela vida das pessoas com quem convivi. Pelo meu orientador, cujo zelo com seus alunos é raro de ser encontrado, e também por meus colegas de laboratório, que tornavam o ambiente de trabalho mais agradável por meio de discussões de trabalho e conversas descontraídas. Agradeço à D-us por cada aula, cada curso e cada evento que eu pude participar, pois todos eles foram de grande importância na minha formação acadêmica. Agradeço também à D-us pelo auxílio financeiro que Ele me proporcionou através das agências financiadoras FAPEMIG e CAPES, pois assim foi provido o sustento de cada dia.

Por fim, agradeço à D-us por ter me trazido até esta etapa e ter me permitido conhecer todas essas pessoas maravilhosas e compartilhar com elas momentos marcantes em minha vida. Que a vida de cada uma seja engrandecida por D-us com bons frutos e que a minha memória não se esqueça de seus atos de bondade. A todos que participaram direta ou indiretamente deste trabalho direciono os meus sinceros agradecimentos.

RESUMO

Estudar as relações evolutivas entre os organismos é um tema que fascina a ciência há vários anos e o seu modo de analisar e inferir a história evolutiva vem se modificando conforme surgem os adventos tecnológicos e novos tipos de dados ao longo da história. O que era, a princípio, analisado por meio de dados morfológicos e de desenvolvimento embrionário, agora pode ser pesquisado com dados moleculares, como sequências nucleotídicas ou de aminoácidos, uma abordagem que domina o campo da filogenia desde a criação da tecnologia de sequenciamento automático de DNA em meados dos anos 80. Durante esse tempo, os métodos de filogenia molecular vêm recebendo aperfeiçoamentos significativos, mas ainda não resolvem todas as lacunas presentes em vários pontos da árvore da vida. Os métodos de sequenciamento em larga-escala vieram para auxiliar na inferência da história evolutiva, mas eles também impõem novos desafios metodológicos, já que muitos procedimentos comuns da filogenia molecular não conseguem lidar com grande volume de sequências geradas por estas máquinas. Neste contexto, o presente trabalho teve como objetivo gerar ferramentas que atentem com as atuais demandas da filogenia molecular.

O presente trabalho gerou três aplicativos: (1) TaxOnTree, (2) HyperTriplets e (3) ELDOgraph. A TaxOnTree é uma ferramenta que incorpora os dados taxonômicos em uma árvore filogenética. A partir das árvores geradas por esta ferramenta, o usuário pode ter um acesso fácil e rápido a todas as informações taxonômicas das amostras presentes na árvore, assim como a relação evolutiva entre elas, tendo como base a árvore taxonômica do NCBI Taxonomy. O HyperTriplets é uma ferramenta que processa várias árvores de genes e gera uma superárvore a partir da análise de tripletos. Abordagens que preenchem lacunas do método de superárvore, como a restrição do uso de árvores de genes contendo parálogos na análise e a não inclusão de dados de distância nos ramos da superárvore, foram desenvolvidas e implementadas nesta ferramenta. Por fim, o ELDOgraph é uma ferramenta que utiliza as distâncias filogenéticas para realizar análises comparativas entre as espécies presentes na árvore. Esta ferramenta oferece ao usuário uma forma de verificar e visualizar as proximidades evolutivas diferente da abordagem das árvores filogenéticas. As três ferramentas estão disponíveis no endereço http://biodados.icb.ufmg.br/taxonphylotools.

Palavras chave: evolução, NCBI Taxonomy, superárvore, genômica comparativa.

ABSTRACT

Studying the evolutionary relationships between organisms is a subject that has fascinated scientists for a long time and the approaches to analyze and infer the evolutionary history has been modified as new technologies and new data types emerge. At first, the morphological and embryogenic data were the main source for phylogenetic analysis, but molecular data, such as nucleotide or amino acid sequences, have ruled the field of phylogenetic studies since the introduction of automated DNA sequencer in the mid ‘80s. Although molecular phylogenetic had received significant methodological improvements during this time, there are still several unresolved parts in the tree of life. Development and popularization of large-scale sequencing technologies came to help in elucidating the gaps in the evolutionary history, but they also impose new methodological challenges, since several common procedures in molecular phylogeny do not deal with a large amount of sequences generated by these machines. In this context, the objectives of the present work were to develop molecular phylogenetic tools that meet the current demands in molecular phylogeny.

We developed three applications in this work: (1) TaxOnTree, (2) HyperTriplets and (3) ELDOgraph. TaxOnTree is a tool that includes taxonomic data into a phylogenetic tree. The trees generated by this tool provide users with an easy and fast way to access all taxonomic information of the samples in the tree, as well as the evolutionary relationship between them based on the taxonomic tree from NCBI Taxonomy. The HyperTriplets is a tool that reads and processes several gene trees to reconstruct a supertree based on analysis of triplets. Approaches that fill some gaps in the supertree methodology, as the restriction on using gene trees with paralogs and the lack of distance values in the supertree branches, were developed and implemented in this tool. Finally, the ELDOgraph is a tool that extracts the phylogenetic distances in several gene trees to perform a comparative analysis amongst the sampled species. This tool provides a new approach to verify and visualize the evolutionary proximity in a distinct way from the tree structure approach. All three tools are available at http://biodados.icb.ufmg.br/taxonphylotools.

Keywords: evolution, NCBI Taxonomy, supertree, comparative genomics.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

OTU	Unidade Taxonômica Operacional
HTU	Unidade Taxonômica Hipotética
UPGMA	Unweighted Pair Group Method with Arithmetic Mean
MP	Máxima Parcimônia
MV	Máxima Verossimilhança
IB	Inferência Bayesiana
MCMCMC	Metropolis-coupled Markov chain Monte Carlo
NCBI	National Center for Biotechnology Information
GOLD	Genomes Online Database
Uniprot	Universal Protein Resource
COG	Clusters of Orthologous Groups
eggNOG	evolutionary genealogy of genes: Non-supervised Orthologous Groups
KEGG	Kyoto Encyclopedia of Genes and Genomes
HOG	Hierarchical Orthologous Groups
OMA	Orthologous MAtrix
MRP	Matrix Representation using Parsimony
MRC	Matrix Representation using Compatibility
ELDO	External Least Divergent Ortholog
LDO	Least Divergent Ortholog
BLAST	Basic Local Alignment Search Tool
RefSeq	Reference Sequence
HSP	High-scoring Segment Pair
NHX	New Hampshire X Format
LCA	Lowest Common Ancestor (menor ancestral comum)
GeneID	Identificador de gene
GI	Identificador de sequência do NCBI

jobID	Identificador da tarefa do TaxOnTree
SVG	Scalable Vector Graphics
iToL	Interactive Tree Of Life
T(I)LI	Triplet (Inference and) Local Inconsistency
Q(I)LI	Quartet (Inference and) Local Inconsistency
PANTHER	Protein ANalysis THrough Evolutionary Relationships

SUMÁRIO

1. INTRODUÇÃO

Estudar as relações evolutivas entre indivíduos ou espécies é um tema que possui uma longa data no meio científico. A ideia de desenhar e esquematizar a evolução biológica por meio de árvores filogenéticas provavelmente derivou do sistema de classificação desenvolvido por Linnaeus no século 18, que organiza os organismos conhecidos em uma série hierárquica de categorias taxonômicas (LINNÉ; SALVIUS, 1758). Este esquema foi comparado a uma árvore da vida e contribuiu para que, nos anos posteriores, naturalistas como Charles Darwin pudessem perceber por meio de uma estrutura em árvore que todos os organismos descendem de um mesmo ancestral (DARWIN; DE BEER, 1956).

Desde então, as relações entre as espécies vêm sendo estudadas através de estruturas de árvores. A sua forma de representação não sofreu variações significativas ao longo dos anos, mas os métodos para a sua construção acompanharam as mudanças das técnicas e dos tipos de dados gerados por elas para os estudos comparativos. Dados mais acessíveis como características morfológicas e anatômicas foram os primeiros a serem utilizados e são analisados até hoje, dada a sua importância em inferir a relação filogenética entre fósseis e em estimar as taxas e as épocas em que ocorreram os processos macroevolutivos (WIENS, 2004). No entanto, com o surgimento e o advento das técnicas de biologia molecular, dados moleculares se tornaram os principais alvos de estudos evolutivos tanto para inferir a evolução das espécies quanto dos genes. A filogenia baseada em dados moleculares, referida como filogenia molecular, iniciou-se com dados derivados de experimentos imunológicos, como a medição da quantidade de reações cruzadas que ocorrem entre um anticorpo específico para uma proteína de um organismo com a mesma proteína de organismos diferentes (NUTTALL, 1904), e de hibridação DNA-DNA (SIBLEY; AHLQUIST, 1984). Mas desde o advento das técnicas de sequenciamento de DNA em meados dos anos 80, as sequências de biomoléculas, como de DNA ou de proteína, têm sido as principais fontes de dados para os estudos filogenéticos e vêm auxiliando na elucidação de várias questões biológicas, como aquelas relacionadas às áreas da sistemática, forense e epidemiológica (HARTFIELD; MURALL; ALIZON, 2014; HILLIS; HUELSENBECK; CUNNINGHAM, 1994).

1.1. Estrutura da árvore filogenética

As árvores filogenéticas podem ser definidas como grafos que possuem uma estrutura hierárquica. Nestas árvores, os nós são denominados de unidades taxonômicas, que podem representar, dependendo dos dados analisados, espécies, populações, genes ou proteínas. Os nós são classificados em terminais (folhas), quando estes se encontram na extremidade da árvore, ou em internos, quando destes partem um ou mais ramos descendentes. Os nós terminais representam as próprias amostras utilizadas para a inferência da árvore e, por isso, são também denominados de unidades taxonômicas operacionais (OTUs), que por sua vez correspondem à unidade básica (espécie, população, gene ou proteína) a ser estudada e comparada (SOKAL, 1966). Os nós internos representam os eventos evolutivos que retratam a divergência da unidade taxonômica em análise, como os eventos de especiação, caso a unidade taxonômica seja população, ou eventos de duplicação de genes, caso a unidade taxonômica seja gene ou proteína. Como a determinação dos nós internos são produtos de uma inferência filogenética, denominamos-os também de unidades taxonômicas hipotéticas (HTUs). Os ramos são elementos que conectam os nós e o seu tamanho representa o tempo estimado da relação evolutiva entre as unidades taxonômicas. Quanto menor a distância dos ramos entre as unidades taxonômicas em comparação, mais próximas elas se encontram evolutivamente.

As árvores filogenéticas ainda podem assumir diferentes conformações dependendo da disposição dos ramos ao longo da árvore. A estas diferentes conformações que uma árvore pode assumir denominamos de topologias. O conceito da topologia é de grande importância nos estudos filogenéticos por ela representar a base de toda a interpretação das histórias evolutivas entre as amostras em análise. Diferentes topologias implicam em diferentes eventos evolutivos, e cabe aos programas de filogenia determinar aquela topologia que melhor se adequa aos dados fornecidos pelo usuário.

1.2. Métodos de inferência da árvore filogenética com dados moleculares

As análises filogenéticas a partir de dados de sequências de DNA ou de proteína se iniciam com um alinhamento de sequências ortólogas e podem ser divididos de acordo com o tipo de dados utilizados para a inferência. Uma parte dos métodos gera

a partir dos dados de alinhamento uma matriz de distância, que fornece uma medida de dissimilaridade entre os pares de sequências presentes no alinhamento. Uma medida simples de dissimilaridade é a porcentagem de sítios em que duas sequências diferem uma da outra ("p-distance"). Outras medidas mais complexas utilizam modelos evolutivos que consideram as múltiplas mudanças que podem ocorrer em um sítio quando as sequências comparadas são muito divergentes entre elas. Os métodos baseados em matrizes de distância caracterizam-se por ser computacionalmente rápidos e por fornecerem ao usuário uma única árvore. O método da média aritmética não ponderada (UPGMA, "Unweighted Pair Group Method with Arithmetic Mean") (SOKAL, 1958) e o método de agrupamento de vizinhos (do inglês "Neighbor-Joining") (SAITOU; NEI, 1987) são exemplos mais conhecidos desta abordagem.

A outra parte dos métodos de filogenia molecular utiliza algoritmos baseados em critérios de otimização. Este método avalia várias topologias da árvore e escolhe aquela que apresenta a melhor pontuação de otimização, ou seja, seleciona a árvore que melhor se ajusta aos dados do alinhamento. Entre os métodos que utilizam esta abordagem, pode ser citado o método da máxima parcimônia (MP), o método da máxima verossimilhança (MV) e a inferência bayesiana (IB). No método da MP, as diferentes topologias das árvores são pontuadas de acordo com o número de mudança dos caracteres (substituições) necessárias para que os dados se ajustem na topologia. O algoritmo escolherá aquela topologia que explique os dados do alinhamento com um menor número de substituições. O método da MV avalia as diferentes topologias das árvores de forma probabilística. A probabilidade das topologias das árvores é calculada baseada em um modelo de substituição e o algoritmo determinará aquela árvore que apresenta maior valor desta probabilidade. Na IB, a inferência da árvore também é baseada no cálculo da probabilidade, mas difere um pouco do conceito aplicado no MP e no MV uma vez que a sua análise produz múltiplas árvores. Uma IB se inicia definindo alguns parâmetros *a priori*, como o modelo de substituição, tamanho dos ramos e a topologia das árvores, para obter uma distribuição de probabilidades *a priori* de um conjunto de árvores. Posteriormente, dados de alinhamento são coletados e neles são aplicados os modelos evolutivos estocásticos e o teorema de Bayes para atualizar as probabilidades *a priori* em probabilidades *a posteriori*. As diferentes topologias que a árvore pode assumir são amostradas na IB utilizando o algoritmo de MCMCMC (do inglês "Metropolis-coupled Markov chain Monte Carlo"). Quando os dados são

informativos, a maior parte da distribuição de probabilidade *a posteriori* se concentrará em um conjunto de árvores.

1.3. Filogenia molecular e a classificação taxonômica

As inferências evolutivas realizadas a partir de dados moleculares permitem uma inferência sobre as relações evolutivas entre as espécies. Vários resultados obtidos a partir da filogenia molecular têm auxiliado na estruturação e na atual concepção da classificação taxonômica das espécies.

A classificação taxonômica das espécies envolve a criação de grupos taxonômicos e a organização das espécies dentro destes grupos. Este tipo de classificação normalmente é baseado nas comparações das características morfológicas compartilhadas entre os organismos. O Reino Eukaryota, um exemplo de um grupo taxonômico, possui este nome por ele reunir organismos que possuem o seu material genético confinado no envelope nuclear. A classificação hierárquica reflete a história evolutiva das espécies, já que duas espécies que compartilham mais características em comum costumam compartilhar um ancestral comum mais recente. No entanto, em alguns pontos dessa classificação podem sugerir dúvidas quanto à origem de um grupo taxonômico ou sobre a classificação de um organismo dentro de um grupo. Antigamente, o grupo taxonômico denominado Archonta reunia os morcegos, os lêmures voadores, os primatas e os escandêncios. A classificação dos morcegos dentro desse grupo era baseado na sua semelhança com os lêmures voadores em relação ao aspecto do voo. No entanto, análises de filogenia molecular utilizando genes mitocondriais revelaram que o morcego era mais próximo dos carnívoros, perissodáctilos e cetartiodáctilos (PUMO *et al.*, 1998), o que levou ao reposicionamento da ordem Chiroptera dentro da superordem Laurasiatheria.

O NCBI Taxonomy (SAYERS *et al.*, 2009) é um banco de dados curado que reúne os nomes e as classificações de todos os organismos que possuem alguma sequência depositada no GenBank. O número de espécies catalogado neste banco supera 360 mil (consultado em julho de 2016). Todas as espécies estão organizadas em uma estrutura hierárquica contendo grupos taxonômicos que possuem um consentimento na comunidade científica e nas análises filogenéticas. Para que a estrutura hierárquica esteja de acordo com as análises filogenéticas, os grupos taxonômicos de hierarquias superiores devem formar grupos monofiléticos na árvore. O grupo taxonômico

denominado Reptilia é um grupo que reúne todas as espécies de lagarto, cobras, tartarugas e crocodilos. No entanto, o consenso atual é que as aves encontram-se posicionadas como o grupo irmão dos crocodilos na árvore de espécie. Como a presença desse grupo quebraria o princípio filogenético desta estrutura por não englobar as aves, o grupo Reptilia foi removido do banco do NCBI Taxonomy. Em outras palavras, o grupo Reptilia não foi acrescido ao banco por ser um grupo parafilético. Ao invés disso, um novo grupo, denominado Sauropsida, foi criado para englobar as aves e répteis (The NCBI Handbook, 2013).

Devido às diversas aplicabilidades oferecidas pelos dados de sequenciamento, existe um grande esforço para que um maior número de espécies tenham alguns de seus genes sequenciados. Atualmente, existem mais de 1,6 milhões de espécies catalogadas (ROSKOV *et al.*, 2016) e a estimativa é que a maior parte das espécies existente ainda estão à espera de serem descobertos (COSTELLO; WILSON; HOULDING, 2012; MORA *et al.*, 2011). Dentre as espécies já catalogadas, 22% delas possuem pelo menos uma sequência depositada no banco de dados do NCBI.

1.4. Árvore de gene e árvore de espécie

Existe uma grande distinção sobre o conceito de árvore de espécie e árvore de gene. A árvore de espécie se refere a uma árvore de um conjunto de espécies que possui uma topologia verdadeira que reflete os passos evolutivos tomados pelas espécies em análise. Já a árvore de genes se refere a uma árvore de um conjunto de genes homólogos, inferidas a partir de dados moleculares. Apesar da inferência da árvore de espécie poder ser baseada na inferência da árvore a partir de sequências de genes, uma árvore de genes pode não refletir a uma árvore de espécie (NICHOLS, 2001). Uma diferença clara encontrada entre essas duas árvores é vista quando a árvore de gene demonstra um evento específico na evolução dos genes. Enquanto que, numa árvore de espécie, cada espécie é representada apenas uma única vez, na árvore de genes, cada espécie pode ser representada por uma amostra, mais de uma amostra, ou por nenhuma amostra. Cada uma dessas situações encontradas em uma árvore de gene representa um evento da evolução particular ao gene em análise, que denota uma duplicação, quando a árvore apresenta mais de uma amostra de uma espécie, ou deleção de gene, quando a árvore não apresenta uma amostra de uma espécie.

Outros fatores que influenciam na incongruência entre a árvore de genes e a árvore de espécie são as variações nas taxas de mutações dos genes ao longo da evolução, o fluxo gênico entre linhagens após o evento de duplicação e eventos de recombinação ao longo do genoma (DEGNAN; ROSENBERG, 2009). Como cada gene utilizado na inferência filogenética pode contar diferentes histórias evolutivas para as espécies em análise, a análise de múltiplos genes se tornou uma boa forma de estimar a árvore de espécie (DELSUC; BRINKMANN; PHILIPPE, 2005).

1.5. Dados genômicos e filogenômica

A revolução nos métodos de sequenciamento de DNA tem possibilitado o sequenciamento completo de genomas simples, como de vírus e organismos unicelulares, e complexos, como de animais, plantas e fungos, além de possibilitar uma nova abordagem para as inferências das árvores filogenéticas. Com a disponibilidade de sequências genômicas de vários organismos, os estudos da filogenia molecular podem agora não só depender da inferência filogenética a partir de um ou poucos genes, mas do conjunto completo de genes presentes nas amostras em análise. O termo filogenômica foi proposto para descrever uma subárea da filogenia molecular que utiliza os dados em escala genômica para realizar os estudos comparativos entre as amostras.

Um repositório importante para os estudos de filogenômica é o próprio banco de dados de genoma do NCBI, que armazena genomas de cerca de 11 mil espécies distintas (consultado em julho de 2016), sendo a maior parte de bactérias (8.593 espécies com genoma sequenciado), seguido de eucariotos (1.957 espécies) e arqueas (573 espécies). Dados de genoma também podem ser encontrados em outros repositórios, como no GOLD (REDDY *et al.*, 2014) e no Ensembl (HERRERO *et al.*, 2016). Além da disponibilidade das sequências de genoma, os acessos destes bancos ainda contam com as anotações dos genes, que são as regiões do genoma mais empregadas nas análises de filogenia molecular. O conjunto completo de proteínas encontradas em um organismo pode ser também obtido nestes bancos e ainda no banco de dados do Uniprot (CONSORTIUM, 2015).

Um tema de grande interesse na área da filogenômica é a determinação dos grupos de ortólogos, já que em todas as análises filogenéticas as comparações entre as sequências devem ser sempre realizadas entre sequências ortólogas, ou seja, sequências que compartilham uma sequência ancestral comum. A determinação dos grupos de

ortólogos é um tema bastante debatido, e por isso existem vários software que realizam esta análise a partir de dados genômicos. Entre os programas podemos citar o OrthoDB (KRIVENTSEVA *et al.*, 2015), OrthoMCL (CHEN, FENG *et al.*, 2006), InParanoid (SONNHAMMER; ÖSTLUND, 2014), COG (TATUSOV *et al.*, 2003), EggNOG (POWELL *et al.*, 2013), KEGG orthology (KANEHISA *et al.*, 2015), HOG (ALTENHOFF *et al.*, 2013) e OMA (ALTENHOFF *et al.*, 2015). Os grupos de ortólogos presentes em um desses bancos podem ser utilizados de forma conjunta (supermatriz) ou independente (superárvore) para a inferência da árvore de espécie (vide próximo tópico).

Tendo em mente que a inferência de um grande número de árvores exige um grande poder computacional e que este nem sempre é disponível para toda comunidade científica, alguns grupos de pesquisa focaram em organizar e distribuir bancos de dados contendo um grande volume de árvores filogenéticas inferidas a partir de métodos considerados de alta complexidade e acurácia. O PhylomeDB (HUERTA-CEPAS *et al.*, 2008), TreeFam(SCHREIBER *et al.*, 2014), OrthoMaM (DOUZERY *et al.*, 2014) e TreeBase (PIEL *et al.*, 2000) são exemplos de repositórios de árvores filogenéticas disponíveis.

A disponibilidade de um grande volume de dados aplicáveis aos métodos de filogenia molecular têm gerado boas expectativas em elucidar várias questões e debates acerca de alguns pontos não resolvidos da árvore de espécie. Ao mesmo tempo, lidar com grande volume de dados exige novos tipos de tratamento nas sequências e novas abordagens para a inferência da árvore de espécie, como será discutido a seguir.

1.6. Abordagens para construção de árvores de espécie com dados ômicos

Com a disponibilidade de sequências provenientes de vários genes, a sua análise simultânea oferece uma maior resolução sobre a história evolutiva das amostras em análise. A forma de analisar esses dados ainda é um tema muito debatido (PHILIPPE *et al.*, 2005; QUEIROZ; GATESY, 2007; VON HAESELER, 2012), mas existem duas principais abordagens que utilizam os dados de vários genes para a reconstrução de uma árvore de espécie: supermatriz e superárvore.

A supermatriz (QUEIROZ; GATESY, 2007) é uma abordagem que utiliza diretamente os dados de alinhamento para a reconstrução da árvore de espécie. Este método, também é referido como o método de concatenação e consiste, primeiramente,

em combinar os dados de alinhamento de sequência de vários genes em um único arquivo. Este arquivo, denominado de supermatriz, é, posteriormente, submetido a uma metodologia de inferência filogenética (por exemplo, a inferência bayesiana ou o método da máxima verossimilhança), que realizará uma análise simultânea de todos os sítios concatenados para a reconstrução da árvore.

Já a superárvore (OLAF R. P. BININDA-EMONDS; JOHN L. GITTLEMAN; STEEL, 2002; SANDERSON; PURVIS; HENZE, 1998) é uma abordagem que utiliza um conjunto de árvores filogenéticas construídas de forma independente a partir de diferentes genes. Os métodos baseados em superárvore reconstruirão, a partir destas árvores de genes, uma árvore consenso que é interpretado como uma árvore de espécie. Entre os algoritmos e software mais comumente utilizados podem ser citados o MRP (do inglês "Matrix Representation using Parsimony") (RAGAN, 1992) e o MRC (do inglês "Matrix Representation using Compatibility") (ROSS; RODRIGO, 2004).

Ambos os métodos foram largamente utilizados na inferência das árvores de espécies em diferentes clados taxonômicos, como de mamíferos (BININDA-EMONDS *et al.*, 2007; SONG *et al.*, 2012), artrópodes (REGIER *et al.*, 2010) e procariotos (LANG; DARLING; EISEN, 2013). A abordagem da supermatriz tem sido descrita como o método preferido para as análises filogenômicas (PHILIPPE *et al.*, 2005), já que a árvore resultante apresenta valores de distância filogenética e uma análise estatística que considera os modelos evolutivos para a verificação dos suportes dos ramos. Apesar dos métodos que geram superárvores conferirem uma boa inferência sobre as relações evolutivas entre as amostras, em geral elas apresentam apenas a topologia da árvore final na forma de um cladograma. Além disso, as medidas de suporte dos ramos apresentadas em uma superárvore geram dúvidas quanto à sua interpretação no contexto evolutivo, uma vez que elas são específicas ao método e não possuem o mesmo princípio dos valores de suporte apresentados pelas análises convencionais de filogenia.

Apesar das medidas de suporte serem alvos de muitas críticas para a análise de superárvore, a análise de supermatriz vem enfrentando uma forte crítica também acerca dos métodos comumente utilizados para a obtenção dos seus valores de suporte. Quando se submete os dados de alinhamento com muitos sítios aos métodos de reconstrução da árvore filogenética, como da máxima verossimilhança ou inferência bayesiana, quase todos os ramos da árvore gerada apresentarão valores de suportes estatísticos próximos de 100%, independente dos dados apresentados (SALICHOS;

ROKAS, 2013). Isto ocorre pelo fato das análises do suporte dos ramos utilizados nestas abordagens (*bootstrapping* ou probabilidade *a posteriori*) terem sido, primariamente, desenvolvidas para verificar a robustez da inferência quando os dados são limitantes. Como o número de dados faltantes é mínimo quando consideramos um alinhamento que representa quase toda a extensão do genoma/proteoma das amostras, o valor alto do suporte dos ramos pode ser verificado mesmo na presença de dados conflitantes ou erros sistemáticos no alinhamento.

De certo, a questão sobre o método mais apropriado para a inferência da árvore de espécie utilizando os dados genômicos ainda se encontra em aberto e ainda possui espaço para o desenvolvimento de um novo software que utilize uma das duas abordagens citadas.

2. JUSTIFICATIVA E OBJETIVO

Os métodos utilizados na filogenia molecular vêm se adaptando conforme os dados disponíveis para a sua análise. O crescente aumento do número de sequências de diferentes organismos é de grande valia para os estudos filogenéticos, no entanto, impõe novos desafios tanto no âmbito metodológico, como também na visualização da árvore e no acesso aos dados referentes às amostras.

Entre as principais demandas dos filogenistas para aqueles que desenvolvem programas na área são de ferramentas que sejam capazes de (1) inferir árvores filogenéticas confiáveis a partir de um grande volume de dados de sequenciamento e de (2) facilitar a visualização e o acesso às informações e às anotações das amostras presentes na árvore. O presente projeto foi executado de forma a desenvolver softwares na área de filogenia molecular que contribuam com estas demandas. Durante o desenvolvimento do trabalho, ao explorar árvores gênicas onde certas amostras discordavam da árvore de espécie esperada, e com a capacidade que tínhamos de identificar programaticamente os clados que apresentavam genes mais próximos, nos envolvemos no desenvolvimento de uma maneira de apresentar proximidades gênicas independentemente da árvore de espécie, uma demanda da genômica comparativa que se choca com a necessidade de solução no formato de uma única árvore.

O objetivo deste trabalho é criar ferramentas que manipulem eficientemente os dados do NCBI Taxonomy, aplicando-os a análises filogenéticas e de genômica comparativa.

3. MATERIAIS E MÉTODOS

3.1. Linguagem de programação

Os programas desenvolvidos neste trabalho possuem como principal linguagem de programação o PERL. Todos eles foram testados em ambiente UNIX que possuem o interpretador PERL nas versões 5.10 e 5.16. Outras linguagens e recursos computacionais utilizados no desenvolvimento do programa foram o banco de dados MySQL, e as linguagens para a programação na web como o HTML/CSS e o Javascript. O pacote do TaxOnTree encontra-se disponível no repositório SourceForge na página sourceforge.net/projects/taxontree/. Já os programas HyperTriplets e ELDOgraph encontram-se na página http://biodados.icb.ufmg.br/taxonphylotools.

3.1.1. Bibliotecas

As seguintes bibliotecas em PERL foram utilizadas para auxiliar no desenvolvimento dos programas:

- BioPerl (Bio::TreeIO, Bio::Tree::TreeFunctionsI, Bio::Tree::TreeI e Bio::Tree::NodeI) - para a leitura e confecção das árvores filogenéticas;
- HTTP::Tiny e URI::Escape – realiza requisições HTTP;
- XML::Simple – para a leitura dos arquivos em XML recuperados via requisições HTTP;
- File::Which - verifica se os softwares utilizados no pipeline estão acessíveis pelo programa;
- Net::Wire10 – conecta a um banco de dados MySQL local;
- Parallel::ForkManager – auxilia na execução de operações em paralelo.

Os pacotes de cada programa acompanham uma pasta "lib*"* que reúne todas as bibliotecas requisitadas pelo software, prevenindo do usuário de instalar estas bibliotecas antes da execução. Além disso, todas as bibliotecas reunidas escolhidas para fazer parte dos programas são classificadas como "Pure Perl", ou seja, são bibliotecas que não utilizam extensões em C e, portanto, não requerem um compilador de C para a

sua execução. Este fato foi levado em conta para facilitar a distribuição do software, já que os compiladores de C pode não conseguir executar programas que foram compilados por outras versões de compiladores.

3.2. Pipeline de inferência filogenética

No programa TaxOnTree foi implementado um "pipeline" que reconstrói uma árvore filogenética a partir de um acesso de proteína, de uma sequência de aminoácido, de uma lista de acessos. O "pipeline" reúne vários softwares de terceiros que são comumente utilizados para cada etapa da reconstrução filogenética. Nesta seção, serão descritos cada uma destas etapas, apresentando algumas particularidades de cada etapa e os software utilizados. É pertinente relatar que as linhas de comando executadas por cada um dos software incluídos no "pipeline" encontram-se reunidos no arquivo CONFIG.xml que acompanha o pacote da TaxOnTree. Caso seja do interesse do usuário alterar algum comando desses software, bastaria alterar o comando listado neste arquivo.

3.2.1. Busca por ortólogos putativos

O pipeline do TaxOnTree utiliza o programa "blastp" do pacote StandAlone BLAST+ (ALTSCHUL *et al.*, 1990; CAMACHO *et al.*, 2009) para buscar proteínas que sejam ortólogos putativos de uma sequência de aminoácidos submetida pelo usuário. Portanto, para o programa realizar esta etapa, ele necessita de um banco de dados de sequência proteica formatado pelo programa "makeblastdb", que se encontra no mesmo pacote.

O pacote StandAlone BLAST+ pode ser obtido pelo endereço eletrônico ftp.ncbi.nlm.nih.gov/blast/executables/LATEST.

3.2.1.1. Banco de dados de sequências

Três bancos de dados de sequências proteicas foram montados e utilizados neste trabalho. Os bancos já pré-formatadas encontram-se disponíveis para qualquer usuário na página do SourceForge da TaxOnTree (sourceforge.net/projects/taxontree/). As descrições de cada banco e de como ele foi montado estão descritos a seguir:

- RefSeq: um banco de dados contendo apenas sequências de RefSeq do banco de dados do NCBI. Este banco também se encontra disponível no site do NCBI no endereço: ftp.ncbi.nlm.nih.gov/blast/db/.
- RefSeq without fragments: um banco de dados contendo apenas proteínas RefSeq que possuam toda a sua extensão sequenciada. Para a sua construção, primeiramente foram listados todos os acessos de proteína presentes no banco de dados de sequência RefSeq citado anteriormente. Foram também listados todos os acessos das proteínas do banco RefSeq que estejam anotados como parciais. Para a obtenção dessa segunda lista, foi utilizado o seguinte parâmetro de busca no site do NCBI Protein (www.ncbi.nlm.nih.gov/protein): refseq[filter] and "partial"[Properties]. As duas listas foram comparadas e apenas aquele acesso presente na primeira lista foi selecionado e incorporado a uma nova lista de acessos que farão parte desse banco. A nova lista foi então submetida ao programa "blastdb_aliastool" juntamente com o banco de dados de sequência RefSeq. Este programa, que se encontra disponível no pacote do StandAlone BLAST+, cria arquivos que permitem que a busca seja realizada no banco de dados original mas considerando apenas os acessos presentes na lista;
- Uniprot Reference proteomes: um banco de dados contendo apenas as proteínas que constituem o proteoma dos organismos de referência presente no banco do Uniprot. As sequências que constituem esse banco foram concatenadas em um único arquivo e submetidas ao programa "makeblastdb". As sequências deste banco podem ser baixadas acessando o endereço ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/reference_proteomes/.

A TaxOnTree funcionará normalmente com banco de dados de sequência montado pelo próprio usuário desde que ele seja gerado utilizando o programa "makeblastdb" a partir de sequências em formato FASTA provenientes do NCBI (The NCBI Handbook, 2013) ou Uniprot (CONSORTIUM, 2015). O usuário ainda pode optar em realizar esta busca de ortólogos nos servidores do NCBI caso ele não tenha

disponível em sua máquina local um banco de dados de sequências. Para este caso, a TaxOnTree envia uma requisição via HTTP aos servidores do NCBI com os dados de sequência do usuário. Esta requisição é executada quando o usuário não especifica um banco de dados local (utilizando o parâmetro -db) na linha de comando.

3.2.1.2. Filtro de hits do resultado do BLAST

Depois de obtidos os resultados do BLAST, a TaxOnTree filtrará os "hits" de forma a manter apenas aqueles que se enquadram nos parâmetros fornecidos pelo usuário. Os seguintes parâmetros podem ser utilizados para definir os "hits" do BLAST que devem ser filtrados:

- E-value (default = 1E-10) – valor estimado que reflete a probabilidade de encontrar o alinhamento observado de forma aleatória;
- Max target seq (default = 200) – número máximo de sequências que devem ser submetidas às análises posteriores.
- Threshold (default = 50) – Um valor de corte da porcentagem de identidade entre a proteína "query" e as proteínas "subjects". A porcentagem de identidade considerada neste filtro ("Tpident") não é a mesma apresentada nos resultados do BLAST pela coluna "pident". Nesta análise, considera-se a porcentagem de identidade como o número de sítios idênticos ou similares em relação ao comprimento total da proteína "query". A diferença entre esta e a porcentagem de identidade da coluna "pident" do BLAST é que o último compreende apenas o comprimento da porção alinhada entre as duas sequências. O "Tpident" pode ser calculado a partir dos valores das colunas fornecidas no resultado do BLAST pela seguinte fórmula:

$$Tpident = \frac{pident * length}{qlen}$$

Onde:

pident: porcentagem de sítios idênticos;

length: comprimento do alinhamento;

qlen: comprimento da proteína "query".

Em casos onde alguma proteína do banco de dados apresenta múltiplos HSPs (do inglês, High-scoring Segment Pair) com a proteína "query", ou seja, quando existem mais de um segmento que proporciona um bom alinhamento entre uma proteína "query" e uma proteína do banco de dados, o algoritmo verifica se as HSPs se sobrepõem uma com a outra. Caso haja alguma sobreposição, a HSP de menor score é descartada da análise. No final, uma média dos valores de identidade ("pident") das HSP ponderada pelos seus respectivos comprimento do alinhamento ("length"). Por fim, o valor da média ponderada é divido pelo tamanho da proteína ("qlen") e o seu resultado corresponderá ao "Tpident" da proteína.

3.2.2. Alinhamento de sequências

Os seguintes alinhadores de sequências foram selecionados para compor o pipeline:

- MUSCLE (EDGAR, 2004);
- PRANK (LÖYTYNOJA; GOLDMAN, 2010);
- ClustalOmega (SIEVERS; HIGGINS, 2014);
- Kalign (LASSMANN; FRINGS; SONNHAMMER, 2009).

Todos estes softwares são livres e estão inclusos no pacote da TaxOnTree acompanhado de suas respectivas licenças. Eles se encontram pré-compilados e reunidos na pasta "bin" que acompanha o pacote.

3.2.3. Refinamento do alinhamento

Para melhorar a inferência filogenética realizada pelo "pipeline", foi incluído nele uma etapa de refinamento do alinhamento. Esta etapa é realizada pelo programa trimAl (CAPELLA-GUTIÉRREZ; SILLA-MARTÍNEZ; GABALDÓN, 2009) que encontra-se também no pacote da TaxOnTree (na pasta "bin") acompanhado de sua licença.

3.2.4. Reconstrução da árvore filogenética

A reconstrução da árvore filogenética do pipeline é realizada pelo programa FastTree (PRICE; DEHAL; ARKIN, 2010). Este software realiza inferências filogenéticas que aproximam do método da máxima verossimilhança. Ele é conhecido por ser rápido e por conseguir lidar com uma grande quantidade de sequências. O software encontra-se também no pacote da TaxOnTree (na pasta "bin") acompanhado de sua licença.

3.3. Recuperação dos dados gênicos e taxonômicos das amostras

Os dados taxonômicos e gênicos de cada amostra que será utilizada na construção da árvore, ou que esteja presente na árvore, podem ser recuperados de duas maneiras pela TaxOnTree. A primeira forma é através de requisições via HTTP nos servidores do NCBI ou do Uniprot. A segunda é através de um banco de dados em MySQL que contém essas informações. A primeira abordagem pode demandar mais tempo na análise, mas previne do usuário de manter um banco de dados relativamente grande em sua máquina. Já a segunda abordagem permite uma rápida consulta à maior parte das informações requisitadas pelo programa, tornando as requisições via HTTP necessárias apenas àquelas informações faltantes no banco de dados local. O banco de dados em MySQL que contém esses dados encontra-se disponível no site do SourceForge da TaxOnTree (sourceforge.net/projects/taxontree/).

3.4. Adição das informações taxonômicas nas árvores filogenéticas

O formato comumente utilizado para representar uma árvore filogenética de forma programática é o formato Newick. A sua forma padrão utiliza parênteses e vírgulas para representar a topologia da árvore e permite a adição de valores do comprimento e do suporte dos ramos. Para a adição de mais informações além destas, outros formatos baseados no formato Newick são recomendados como o formato NHX e o formato Nexus. Em ambos os formatos, a adição de outras informações na árvore é realizada na forma de "tags" (etiquetas), que podem ser acrescentadas em cada nó da árvore. A TaxOnTree utiliza o formato Nexus para acrescentar os dados taxonômicos de cada amostra e de cada nó interno presente na árvore. A adição dessas etiquetas é realizada com o auxílio da biblioteca Bio::Tree::NodeI do BioPerl.

3.5. Determinação do LCA

O menor ancestral comum (LCA – Lowest Common Ancestor) entre dois ou mais organismos representa o ancestral mais recente que todos os organismos têm em comum. Este dado pode ser mais precisamente estimado utilizando os métodos filogenéticos, mas também é possível utilizar os dados da classificação hierárquica das espécies para ter uma noção do quão distante se encontra o LCA entre os organismos em análise.

A TaxOnTree utiliza a classificação hierárquica das espécies disponíveis no banco de dados Taxonomy do NCBI. Para todas as espécies em análise, a TaxOnTree recupera as suas linhagens taxonômicas e estas linhagens são comparadas entre elas para encontrar o último nível da linhagem onde todas as espécies compartilham o mesmo táxon. Para determinar o LCA de humano (*Homo sapiens*) e cachorro (*Canis familiaris*), o algoritmo primeiramente recupera as linhagens taxonômicas das duas espécies. Recuperadas as linhagens, o algoritmo percorre os níveis das duas linhagens taxonômicas, iniciando da raiz, até encontrar o nível em que ocorre a primeira divergência de táxons. O LCA, no caso, se encontrará no nível anterior ao nível em que ocorre a primeira divergência. A primeira divergência encontrada entre as duas linhagens ocorre no 22º nível, que é ocupado pelo táxon Euarchontoglires e Laurasiatheria nas linhagens de humano e de cachorro, respectivamente. Portanto, o LCA de humano e cachorro ocorre no 21º nível, que corresponde ao táxon Boreoeutheria (Figura 1). O mesmo procedimento pode ser utilizado para determinar o LCA de três ou mais espécies. É possível verificar, por exemplo, que o LCA entre humano, cachorro e sapo (*Xenopus tropicalis*) ocorre no 16º nível, que corresponde ao táxon Tetrapoda (Figura 1).

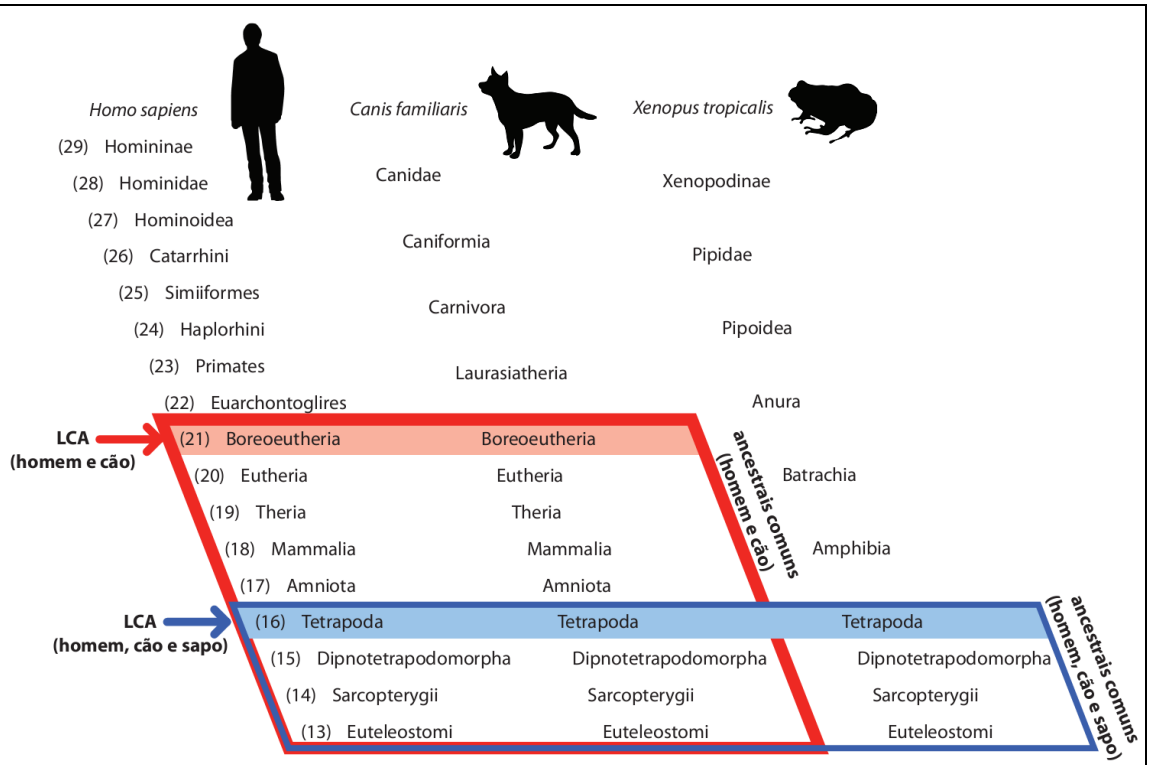


Figura 1: Determinando o LCA de homem e cachorro (vermelho) e de homem, cachorro e sapo (azul).

3.6. Simplificação das linhagens taxonômicas

Os níveis presentes nas linhagens taxonômicas recuperadas no NCBI Taxonomy podem ser classificados em categorias taxonômicas, como reino, filo, classe, ordem, etc. No entanto, duas observações devem ser consideradas ao associar a linhagem taxonômica do NCBI com as categorias taxonômicas:

a) Alguns táxons de uma linhagem taxonômica não se encontram enquadrados em nenhuma categoria taxonômica. Estes táxons são referidos no NCBI Taxonomy como “no rank”. Boreoeutheria, Eutheria e Theria são alguns exemplos de taxa na linhagem taxonômica de humanos (Figura 2A) que não estão enquadrados em uma categoria taxonômica;

b) Algumas categorias taxonômicas estão ausentes na linhagem taxonômica de uma espécie. Na linhagem taxonômica dos humanos (Figura 2A), não existe táxon para as categorias superclasse, subclasse, subgênero e subespécie.

Para lidar com estas situações, foi implementado na TaxOnTree um algoritmo que simplifica a linhagem taxonômica do NCBI (MELO; ORTEGA, 2014). A simplificação consiste em retirar os taxa que não possuem uma categoria taxonômica e em criar um táxon para cada categoria taxonômica que não está presente na linhagem. Para ilustrar como o algoritmo cria um novo táxon, tomemos como exemplo a categoria superclasse, que está ausente na linhagem humana. Para criar um táxon para esta categoria, o algoritmo verificará o primeiro táxon enquadrado em uma categoria taxonômica abaixo da categoria superclasse, que, neste caso, seria Mammalia (Classe). Então, o algoritmo extrairá o nome do táxon anterior, neste caso Amniota, e o usará para nomear a categoria taxonômica faltante. Amniota não é considerado superclasse de humano, mas o nome verdadeiro para esta superclasse poderia estar entre Amniota e Mammalia. Portanto, o algoritmo nomeará a superclasse de humano como uma superclasse e que é filha de Amniota (ou simplesmente de "spc of Amniota").

O fato de nem todas as categorias taxonômicas estarem presentes em uma linhagem taxonômica de uma espécie faz com que algumas categorias taxonômicas sejam mais frequentes do que outras. O algoritmo de simplificação da linhagem taxonômica considera na análise 17 categorias taxonômicas consideradas mais frequentes (Figura 2) (MELO; ORTEGA, 2014). Coincidentemente, quase sempre as categorias com o prefixo super e sub, além das principais, são as mais frequentes. Categorias taxonômicas pouco frequentes como infraordem, parvordem, tribo e entre outras, não são incluídas para a construção da linhagem taxonômica simplificada, mas podem ser utilizadas, assim como as "no rank", para nomear os táxons das categorias escolhidas, quando faltantes. Este procedimento torna as linhagens taxonômicas comparáveis entre elas, já que todas as linhagens taxonômicas submetidas ao algoritmo apresentarão as mesmas categorias taxonômicas.

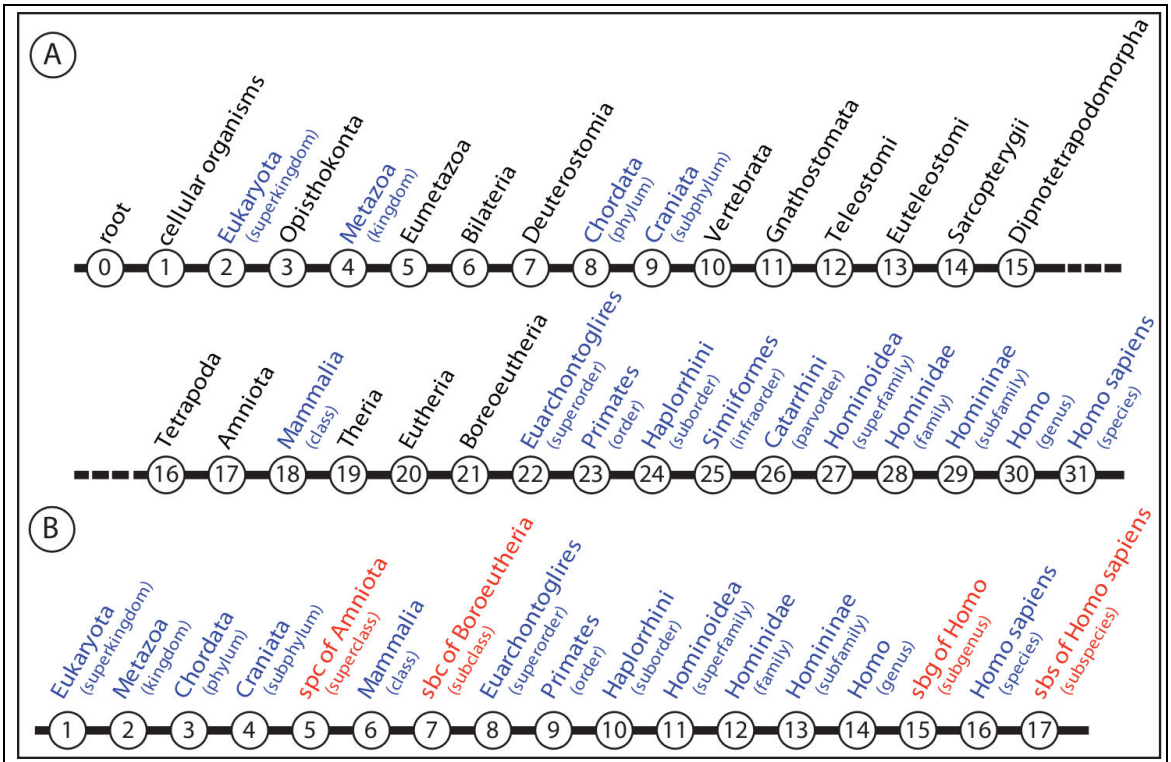


Figura 2: Linhagem taxonômica humana (A) e a sua simplificação (B).
Taxa em azul são taxa que estão enquadradas em uma categoria taxonômica. Taxa em vermelho são taxa criados pelo algoritmo.

3.7. Aplicação Web

Para permitir um acesso rápido ao programa, foi desenvolvida uma aplicação Web que permite ao usuário ter acesso à maior parte das funcionalidades que a TaxOnTree oferece (biodados.icb.ufmg.br/taxontree). O "front-end" da ferramenta foi construído utilizando a linguagem HTML/CSS e Javascript. As tarefas submetidas pelo usuário são enviadas para servidores do laboratório de Biodados (ICB/UFMG) por requisições via Ajax da biblioteca JQuery. No "back-end", as tarefas submetidas pelo usuário são recebidas por scripts escritos em PERL, utilizando a biblioteca CGI, e submetidas ao sistema de fila oferecido pela biblioteca PERL TheSchwartz. O sistema de fila da biblioteca TheSchwartz é constituído por um banco de dados MySQL, para armazenar as tarefas a serem executadas, e um script escrito em PERL denominado "worker", que recupera uma tarefa presente no banco de dados e a executa. Neste caso, o script "worker" foi elaborado de forma que ele execute a TaxOnTree no servidor com os parâmetros fornecidos pelo usuário. Os resultados são armazenados em uma pasta do

servidor e o seu conteúdo é acessado pelo "front-end" também através de requisições via Ajax.

O usuário ainda possui a opção de cadastrar uma conta que permite uma melhor organização das suas tarefas e que facilita o acesso aos seus resultados. Para isso dados de cadastro do usuário são armazenados em um banco de dados MySQL. Os dados de cadastro consistem apenas no nome do usuário e a senha. As senhas são armazenadas na forma criptografada. A criptografia das senhas é feita com o auxílio das bibliotecas: Data::Entropy::Algorithms, que gera um conjunto de caracteres aleatórios ("sal") a ser adicionada à senha, e Crypt::Eksblowfish::Bcrypt, que realiza a criptografia da senha. Este procedimento não permite que o administrador do site tenha o conhecimento das senhas do usuário e dificulta a decodificação das senhas criptografadas por invasores.

3.8. Visualização dos dados

Nesta seção, serão citados os recursos gráficos utilizados por cada arquivo de saída que são gerados pelos programas desenvolvidos neste trabalho.

3.8.1. Árvores filogenéticas

O arquivo de saída da TaxOnTree é uma árvore filogenética no formato Nexus que foi projetado para ser aberto no programa FigTree (RAMBAUT, 2009). FigTree é um programa de visualização de árvores filogenéticas escrito em Java. O programa é disponibilizado gratuitamente pelos autores e comumente utilizado no meio científico por fornecer vários recursos de edição da árvore e por gerar imagens apropriadas para publicações. O arquivo Nexus gerado pela TaxOnTree possui, além das informações taxonômicas incluídas na forma de etiquetas nos nós da árvore, dados de formatação que são específicos para o programa FigTree.

O programa HyperTriplets possui como saída uma superárvore no formato Newick. Portanto, a superárvore gerada pode ser visualizada pela maioria dos programas de visualização de árvores filogenéticas. Neste trabalho, a superárvore gerada foi submetida ao programa TaxOnTree e a árvore resultante foi visualizada no programa FigTree. O HyperTriplets também gera um arquivo que pode ser carregado em uma árvore presente no site iToL (LETUNIC; BORK, 2011). Este arquivo faz com

que os valores de suporte para a topologia atual e as topologias alternativas sejam representados na forma de um gráfico de pizza em cada nó interno da árvore em análise.

3.8.2. Grafos

Uma das saídas do programa ELDOgraph é um arquivo em HTML que apresenta um grafo. O grafo é gerado utilizando o "Force Layout" da biblioteca Javascript D3.js (BOSTOCK; OGIEVETSKY; HEER, 2011). A página pode ser aberta nos principais navegadores.

3.9. Dados reais de árvores filogenéticas

Para testar o algoritmo e verificar o seu desempenho, utilizamos dados reais de árvores filogenéticas disponíveis em dois bancos de dados públicos. O primeiro banco, OrthoMaM (DOUZERY *et al.*, 2014), possui em seu repositório árvores geradas a partir de sequências de mamífero. As árvores constituem-se de até 43 espécies de mamíferos e encontram-se enraizadas. Estas árvores foram construídas pelo método de verossimilhança utilizando o programa RAxML (STAMATAKIS, 2006) a partir de dados de alinhamento de nucleotídeos baseado em códon. O segundo banco, PhylomeDB (HUERTA-CEPAS *et al.*, 2008), possui várias coleções de árvores de diferentes clados taxonômicos. Deste banco de dados foram obtidas árvores reconstruídas a partir de sequências de fungo. As árvores utilizadas para o experimento foram recuperadas de uma coleção de árvores com o nome de acesso "Yeast Phylome (P60)". Esta coleção reúne árvores não enraizadas de até 60 espécies de fungo e de mais dois grupos externos (*Homo sapiens* e *Arabidopsis thaliana*). As árvores foram reconstruídas pelo método da máxima verossimilhança utilizando o programa PhyML (GUINDON *et al.*, 2010) a partir de dados de alinhamento de proteínas.

Como o programa HyperTriplets requer que as árvores estejam enraizadas, as árvores não enraizadas da PhylomeDB foram enraizadas utilizando o algoritmo de enraizamento baseado em dados taxonômicos implementado no TaxOnTree. O funcionamento desse algoritmo será abordado na seção de Resultados.

As árvores utilizadas neste trabalho encontram-se disponíveis no seguinte endereço eletrônico: http://biodados.icb.ufmg.br/taxonphylotools.

3.10. Execução dos programas

Esta seção abordará brevemente algumas instruções sobre a instalação e os comandos de linha que podem ser utilizados pelo usuário para executar os programas desenvolvidos neste trabalho. Todos os programas foram desenvolvidos em PERL e, por isso, a maioria dos computadores com ambiente UNIX pode ser utilizada. As particularidades de cada programa quanto às outras requisições serão abordadas a seguir. Todas as linhas de comando citadas nesta seção consideram as terminologias utilizadas em um terminal de ambiente UNIX.

3.10.1. TaxOnTree

Ao baixar o pacote do TaxOnTree na página do SourceForge (sourceforge.net/projects/taxontree/), descompacte o pacote e entre na pasta "taxontree" executando os comandos:

```
> tar -zxvf TaxOnTree_vXXX_XXX.tgz
> cd taxontree
```

Para verificar o funcionamento e os parâmetros aceitos pelo programa você pode executar os seguintes comandos:

```
> ./taxontree
```

Ou

```
> ./taxontree –man
```

O uso básico do programa TaxOnTree envolve um desses quatro comandos:

```
>./taxontree -singleID <sequence_ID>
```

(recebe um acesso de uma proteína do NCBI ou do Uniprot)

```
>./taxontree -seqFile <FASTA_file>
```

(recebe um arquivo contendo uma sequência no formato FASTA)

```
>./taxontree -listFile <list_file>
```

(recebe um arquivo contendo uma lista de acessos de proteína)

```
>./taxontree -treeFile <tree_file> -queryID <sequence_ID>
```

(recebe um arquivo contendo uma árvore filogenética no formato Newick. O parâmetro –queryID recebe o nome de uma amostra da sua árvore que deve ser considerada como uma proteína "query")

Todos os comandos citados acima requerem que o programa seja executado em um computador que tenha acesso a internet e não necessitam de banco de dados de sequência ou de dados taxonômicos instalados localmente.

3.10.2. HyperTriplets

O programa HyperTriplets é composto por dois scripts onde cada um deles realiza uma etapa específica da análise de ELDO.

O primeiro script que o usuário deve executar é o "hypertriplets_db", que é um script que construirá uma tabela contendo dados de topologia e de distância entre as amostras fornecidas pela árvore filogenética. Esse script pode ser executado utilizando o seguinte comando:

```
./hypertriplets_db –treedir <tree directory>
```

Onde –treedir é o parâmetro que recebe o endereço do diretório onde se encontram as árvores filogenéticas para a análise. Para executar este comando, o nome de cada amostra da árvore deve corresponder a um organismo específico. Em casos onde outras informações estejam inseridas no nome da amostra (como número de acesso, ou nome da proteína), se o nome do organismo estiver delimitado por caracteres específicos ("|", por exemplo), o usuário pode fornecer este caractere e a posição onde se encontra o nome do organismo para que o programa extraia apenas a informação necessária de cada amostra.

O arquivo de saída gerado pelo script “hypertriplets_db” será utilizado como entrada para o segundo script, “hypertriplets_analyse”, do programa HyperTriplets. A seguinte linha de comando pode ser utilizada para a execução desse script:

```
>./hypertriplets_analyse –table <table file> -out <out file>
```

Onde:

-table: recebe o nome do arquivo binário gerado pelo script “hypertriplets_db”;

-out: recebe um prefixo para nomear os arquivos de saída do programa.

Este comando inferirá uma superárvore com todos os organismos presentes nas árvores fornecidas pelo usuário. Caso o usuário deseje que alguns organismos sejam retirados da análise, basta criar um arquivo contendo uma lista de nomes dos organismos a serem incorporados na superárvore e fornecer o nome do arquivo utilizando o parâmetro “-restrict”.

3.10.3. ELDOgraph

O programa ELDOgraph é composto por dois scripts onde cada um deles realiza uma etapa específica da análise de ELDO.

O primeiro script que o usuário deve executar é o “eldograph_db.pl”, que é um script que construirá uma estrutura de dados contendo todos os dados de distância entre duas amostras fornecidas pela árvore filogenética. Esse script pode ser executado utilizando o seguinte comando:

```
>./eldograph_db –treedir <tree directory> -taxtable <taxonomy table>
```

Onde;

–treedir: recebe o endereço do diretório onde se encontram as árvores filogenéticas para a análise;

–taxtable: recebe o nome de um arquivo tabular onde estão listados os nomes dos organismos encontrados nas árvores filogenéticas na primeira coluna e o identificador taxonômico correspondente na segunda coluna.

Para executar este programa, é preferível que o nome de cada amostra nas árvores corresponda ao nome de um organismo específico. Em casos onde outras informações estejam inseridas no nome da amostra (como número de acesso, ou nome da proteína), se o nome do organismo estiver delimitado por caracteres específicos ("|", por exemplo), o usuário pode fornecer este caractere e a posição onde se encontra o nome do organismo para que o programa extraia apenas a informação necessária de cada amostra. Além disso, esta etapa do programa necessita que o computador em execução esteja conectado a internet para que os dados taxonômicos de cada amostra sejam recuperados. No final de sua execução, o script gerará um arquivo binário (eldo_db.hash) contendo informações sobre as distâncias filogenéticas entre pares de amostras e os dados taxonômicos de cada amostra.

O arquivo de saída gerado pelo script "eldograph_db.pl" será utilizado como entrada para o segundo script, "eldograph_analyse.pl", do programa ELDOgraph. A seguinte linha de comando pode ser utilizada para a execução desse script:

```
>./eldograph_analyse –hashfile <hash file> -rank <taxonomic rank> -threshold <threshold value> -out <out file>
```

Onde:

-hashfile: recebe o nome do arquivo binário gerado pelo script "eldograph_db.pl";

-rank: recebe um valor numérico de 1 a 17 que corresponde a uma categoria taxonômica para a análise de ELDO. As categorias taxonômicas que podem ser selecionadas para a análise correspondem às mesmas categorias utilizadas para construção da linhagem taxonômica simplificada dos organismos (Vide item 3.1.8) como listadas a seguir:

1 - "superkingdom";

2 - "kingdom";

3 - "phylum";

4 - "subphylum";

5 - "superclass";

6 - "class";

7 - "subclass";

8 - "superorder";

9 - "order";

10 - "suborder";

11 - "superfamily";

12 - "family";

13 - "subfamily";

14 - "genus";

15 - "subgenus";

16 - "species";

17 - "subspecies".

-threshold: uma margem de distância para considerar os casos de empates de ELDO (mais detalhes na seção dos Resultados).

-out: prefixo para o nome dos arquivos de saída.

4. RESULTADOS

Este trabalho compreendeu no desenvolvimento de ferramentas para análise filogenética e da distribuição taxonômica dos grupos de genes ortólogos. Serão apresentadas as ferramentas TaxOnTree, HyperTriplets e ELDOGraph.

4.1. TaxOnTree

TaxOnTree é uma ferramenta que confere ao pesquisador a capacidade de identificar visualmente todos os clados taxonômicos em uma árvore filogenética de acordo com os organismos que possuem as sequências representadas na árvore em análise. Os diferentes clados presentes em uma árvore filogenética são evidenciados atribuindo diferentes cores aos ramos da árvore. Esta sessão descreverá todo o funcionamento deste programa ("pipeline"), assim como as formas de visualização conferidas às árvores processadas por este programa e as possibilidades de análise que esta ferramenta possibilita.

4.1.1. Pipeline

A TaxOnTree foi primariamente desenvolvida para manipular uma árvore filogenética fornecida pelo usuário. No entanto, para fornecer ao usuário outras opções de entrada ao programa, foi implementado um pipeline de reconstrução de árvore filogenética que utiliza software comumente utilizados nas etapas da inferência da árvore filogenética, que vai desde a obtenção de sequências ortólogas putativas à reconstrução da árvore filogenética propriamente dita. Um esquema geral do pipeline do TaxOnTree está ilustrado na Figura 3.

Descrevendo o pipeline de forma resumida, o usuário pode fornecer como entrada um identificador de proteína ou uma sequência de aminoácidos ao programa. A proteína fornecida é então utilizada como entrada para a análise de similaridade realizada pelo algoritmo do BLAST (ALTSCHUL *et al.*, 1990), que recupera, de um banco de dados de sequência, sequências ortólogas putativas, em ordem de similaridade com a enviada. As bases de dados utilizadas por nós são a RefSeq completa ou processada para conter somente proteínas completas, ou a base UniProt Reference proteomes. Esta etapa pode ser saltada caso o usuário forneça ao programa uma lista de

identificadores de proteínas ortólogas como entrada, que pode ser obtida de uma vasta opção de bases de dados de homólogos. As sequências consideradas ortólogas são então alinhadas e, opcionalmente, o resultado do alinhamento é avaliado quanto a sua qualidade. Vários alinhadores múltiplos são oferecidos como opção: MUSCLE (EDGAR, 2004), PRANK (LÖYTYNOJA; GOLDMAN, 2010), Clustal Omega (SIEVERS; HIGGINS, 2014) e Kalign (LASSMANN; FRINGS; SONNHAMMER, 2009). A qualidade do alinhamento pode ser melhorada com o software trimAl (CAPELLA-GUTIÉRREZ; SILLA-MARTÍNEZ; GABALDÓN, 2009). No caso de um resultado de BLAST, pode ocorrer que um grupo de ortólogos se esgote e sejam adicionadas sequências de outro grupo, ou pode ocorrer que o limite de sequências escolhido pelo usuário (*default* = 200) seja alcançado. Uma adição útil é a possibilidade de retirar da lista as sequências com GeneID repetidos, mantendo somente a mais similar, para excluir isoformas que geralmente poluem a apresentação da árvore. Posteriormente, o alinhamento das sequências é submetido a um software de reconstrução da árvore filogenética que gera um arquivo de árvore no formato Newick. Devido à necessidade de velocidade de processamento, foi eleito o software FastTree. Alternativamente, uma árvore no formato Newick também serve como arquivo de entrada do programa, permitindo que o usuário inicie o pipeline a partir deste ponto, caso ele prefira utilizar uma árvore construída com software de sua preferência. Na próxima etapa, a TaxOnTree recupera as linhagens taxonômicas de todos os organismos representados na árvore filogenética e determina o menor ancestral comum (LCA – "Lowest Common Ancestor") entre a espécie que a proteína de entrada pertence e as outras espécies encontradas na árvore. O LCA representa o táxon mais recente que duas espécies compartilham ao longo das suas linhagens taxonômicas (para mais detalhes vide Materiais e Métodos). Por fim, a árvore é convertida para o formato Nexus e todas as informações taxonômicas são armazenadas na árvore na forma de "tags" em cada nó da árvore. O arquivo final da TaxOnTree é um arquivo de árvore no formato Nexus planejado para ser visualizado no programa FigTree (RAMBAUT, 2009).

A seguir serão descritos com mais detalhes cada uma das etapas que aplicamos no desenvolvimento do pipeline da TaxOnTree.

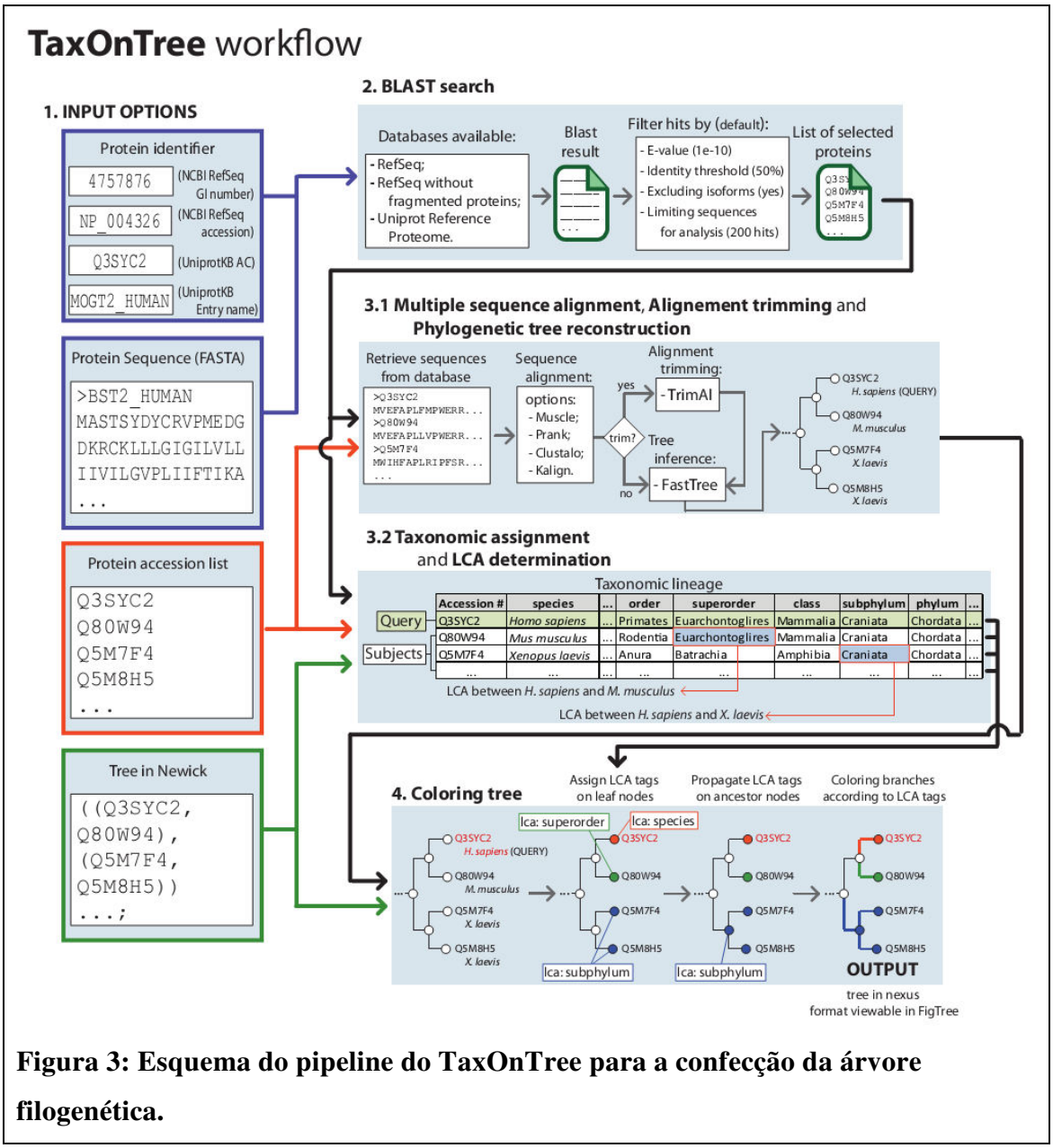


Figura 3: Esquema do pipeline do TaxOnTree para a confecção da árvore filogenética.

4.1.1.1. Inputs

Atualmente, a TaxOnTree oferece ao usuário quatro opções de entrada para o pipeline:

(1) um identificador de proteína – esta opção pode ser utilizada caso a proteína a qual o usuário deseja analisar possua um número de acesso no banco de dados do NCBI ou do Uniprot. A TaxOnTree aceita como identificadores de sequências proteicas: o número GI ou o número de acesso, do NCBI, ou número de acesso (Accession number) ou nome de entrada (Entry name), do Uniprot;

(2) uma sequência de aminoácidos – esta deve estar no formato FASTA. Um usuário pode utilizar esta opção de entrada para analisar sequências proteicas que não estão presentes nos bancos de dados de sequências do NCBI ou do Uniprot. Nesta opção, o usuário pode indicar o organismo que esta sequência pertence fornecendo o identificador taxonômico. Caso esta informação não seja fornecida, será atribuído à sequência de entrada o mesmo identificador taxonômico do melhor hit na análise do BLAST;

(3) uma lista de identificadores de proteínas – O usuário pode utilizar esta opção caso ele tenha em mãos uma lista de identificadores proteicos do NCBI ou do Uniprot que fazem parte do mesmo grupo de ortólogos. Os identificadores proteicos aceitos pelo TaxOnTree são os mesmos destacados anteriormente no primeiro tipo de entrada. Nesta opção, o usuário pode destacar qual das proteínas na lista deve ser considerada como "query" para a determinação do LCA. Caso esta não seja fornecida, o programa utilizará a primeira proteína da lista como "query" para realizar esta análise;

(4) uma árvore filogenética – a árvore deve estar no formato Newick que pode conter tanto valores de suporte nos nós, quanto os valores do comprimento dos ramos. As suas folhas devem conter pelo menos um identificador proteico válido para que o programa consiga mapear cada proteína presente na árvore no banco de dados de proteínas e assim identificar a qual organismo esta pertence. Caso os nomes das folhas não sejam constituídos apenas com o identificador proteico, o usuário pode fornecer ao programa o caractere que delimita o identificador proteico ("|", por exemplo) e a posição do identificador considerando este delimitador. Caso as folhas da árvore não contenham um identificador proteico, o usuário pode fornecer uma tabela delimitada por tabulação que contenha na primeira coluna o nome da folha e na segunda coluna o respectivo identificador taxonômico. Ao fornecer uma árvore filogenética como entrada, o usuário deve obrigatoriamente fornecer o nome da proteína que será considerada como "query" para a determinação do LCA. Esta opção visa atender usuários que criem árvores com software de sua preferência, principalmente aqueles de alta demanda computacional.

4.1.1.2. Busca por sequências ortólogas putativas

O primeiro passo do pipeline, quando fornecemos um identificador proteico ou uma sequência de proteína, é realizar a busca por proteínas ortólogas putativas a partir

de análise de similaridade entre a proteína fornecida pelo usuário e as proteínas encontradas nos bancos de dados. Para esta análise, foram disponibilizadas ao usuário três bases de dados de sequências proteicas:

- RefSeq – Um banco de dados de sequências do NCBI que constitui-se de sequências não redundantes e bem anotadas;

- RefSeq without fragments – O mesmo banco de dados RefSeq, no entanto, contendo apenas sequências com status "Complete". O descarte de proteínas consideradas fragmentadas oferece uma análise mais confiável nas etapas de reconstrução da árvore filogenética;

- Uniprot Reference Proteomes - Constitui-se de sequências proteicas do banco de dados do Uniprot de organismos que possuem o seu genoma completamente sequenciado. Esta opção, ao operar com menos genomas, permite que clados mais distantes da "query" sejam representados.

As sequências proteicas recuperadas a partir desta análise podem ser filtradas considerando um valor máximo de E-value, um valor mínimo de identidade com a proteína fornecida pelo usuário ou um valor máximo do número de sequências que devem ser utilizadas para as análises posteriores. A identidade entre a proteína fornecida pelo usuário e as proteínas recuperadas é calculada pela TaxOnTree considerando a cobertura do alinhamento em relação ao tamanho da proteína considerada "query" (vide Material e Métodos). O usuário pode também optar por retirar da lista os "hits" com identificador GeneID redundantes, com o intuito de retirar isoformas e diminuir a informação redundante.

Para usuários que preferem executar o programa localmente, os três bancos de dados de sequência disponíveis para o usuário da Web podem ser baixados na página do SourceForge do TaxOnTree no endereço sourceforge.net/projects/taxontree/. O usuário pode ainda formatar o seu próprio banco de dados contendo acessos de sequências do Uniprot ou do NCBI. Para isso, basta o usuário baixar as sequências que devem constituir o banco de dados e formata-lo utilizando o programa *makeblastdb* do pacote BLAST+ (CAMACHO *et al.*, 2009).

4.1.1.3.Alinhadores de sequência

Na etapa do alinhamento de sequências, as sequências proteicas que foram selecionadas da etapa da busca por sequências ortólogas, ou da lista de proteínas

fornecidas pelo usuário, são colocadas em formato Multi-FASTA e submetidas a um dos alinhadores múltiplos de sequência selecionado pelo usuário. Foram inclusos no pacote da TaxOnTree quatro softwares comumente utilizados no meio científico: (a) MUSCLE (EDGAR, 2004), (b) ClustalOmega (SIEVERS; HIGGINS, 2014), (c) Kalign2 (LASSMANN; FRINGS; SONNHAMMER, 2009) e (d) PRANK (LÖYTYNOJA; GOLDMAN, 2010). O resultado desta análise é também um arquivo Multi-FASTA contendo as sequências alinhadas.

4.1.1.4. Análise da qualidade do alinhamento de sequência

Após a etapa do alinhamento múltiplo de sequências, o alinhamento resultante pode ser, opcionalmente, submetido ao programa trimAl (CAPELLA-GUTIÉRREZ; SILLA-MARTÍNEZ; GABALDÓN, 2009). O trimAl é um software que analisa a qualidade do alinhamento múltiplo de sequências e remove sequências ou regiões pouco alinhadas. A remoção dessas sequências ou sítios tem o objetivo de melhorar a inferência da árvore filogenética do próximo passo. O seu resultado é também um arquivo Multi-FASTA, mas com as sequências e regiões pouco alinhadas filtradas pelo programa.

4.1.1.5. Inferência da árvore filogenética

A inferência filogenética da TaxOnTree é realizada utilizando o programa FastTree (PRICE; DEHAL; ARKIN, 2010). Este programa toma o arquivo de alinhamento em Multi-FASTA como entrada e realiza uma inferência filogenética utilizando um método que se aproxima do método de máxima verossimilhança. Este método permite que a inferência filogenética ocorra de forma rápida e com uma acurácia que supera outros métodos que também realizam uma inferência rápida, como o Neighbor-Joining, e que é comparável aos próprios métodos da máxima verossimilhança. O arquivo resultante desta análise é uma árvore filogenética no formato Newick.

4.1.1.6. Recuperação das informações taxonômicas das sequências em análise

A partir da árvore filogenética reconstruída pelo "pipeline", ou a partir da árvore fornecida pelo usuário como entrada do programa, a TaxOnTree lista todas as proteínas que constituem a árvore filogenética e recupera as informações taxonômicas dessas proteínas, que são necessárias tanto para a determinação do LCA quanto para atribuição dos clados taxonômicos para cada nó da árvore filogenética. As informações taxonômicas que a TaxOnTree necessita para estas análises são basicamente as linhagens taxonômicas de cada organismo representado na árvore filogenética em análise. Estas informações podem ser recuperadas a partir de uma consulta a uma tabela no banco de dados em MySQL, que pode ser instalado localmente (o banco pode ser obtido na página do SourceForge da TaxOnTree: sourceforge.net/projects/taxontree/), ou a partir de requisições via HTTP nos servidores do NCBI ou do Uniprot, realizadas pelo programa. No último caso, o computador ou o servidor que esteja executando a TaxOnTree deve estar conectado a internet, e é pertinente observar que esta abordagem eleva um pouco o tempo de execução do programa. Após a determinação do LCA entre a espécie considerada *query* e as outras espécies presentes na análise, a árvore, que se encontra no formato Newick, é convertida no formato NHX, que é um formato que permite a adição de *tags* nos nós da árvore que permite a inclusão de outras informações nos nós além daquelas comumente encontradas nas árvores no formato Newick (ex. valor de suporte). Serão nessas *tags* que as informações taxonômicas, como também o resultado da análise do LCA serão incluídas na árvore filogenética resultante. No final, a árvore em formato NHX é novamente convertida, desta vez para o formato Nexus.

4.1.2. Visualização das árvores filogenéticas

O arquivo de saída gerado pela TaxOnTree é um arquivo de árvore no formato Nexus que inclui etiquetas na árvore contendo as informações taxonômicas, além de parâmetros reconhecidos pelo programa FigTree (RAMBAUT, 2009). A fim de exemplificar o aspecto visual das árvores geradas pela TaxOnTree, o programa foi executado utilizando o número de acesso do Uniprot da proteína AGXT2 humana (Q9BYV1) como entrada. As várias marcações referem-se à Figura 4 e serão exploradas abaixo. Para esta análise, foi utilizado o banco de dados do "Uniprot Reference proteomes" com o corte de identidade de 70% para a busca dos ortólogos putativos e possíveis isoformas foram excluídas com o filtro. Além disso, as sequências

recuperadas foram alinhadas utilizando o MUSCLE e o alinhamento resultante submetido ao programa trimAl.

4.1.2.1. LCA

Ao abrir um arquivo Nexus gerado pela TaxOnTree no programa FigTree, o usuário já pode encontrar a árvore com os ramos coloridos de acordo com o critério LCA e com a proteína considerada como "query" evidenciada com a cor de fonte vermelha (Figura 4A). A gradação das cores apresentada nos ramos da árvore é relativa à distância taxonômica entre o organismo que a proteína "query" pertence e os outros organismos representados na árvore. Além das cores dos ramos, a distância taxonômica pode ser acessada pelo valor que se encontra entre parênteses e antecedido pela letra "n" nos nomes das folhas da árvore. Este valor representa o nível do LCA, que indica o número de nós taxonômicos presentes entre o nó raiz da linhagem ao nó do LCA na linhagem taxonômica do organismo ao qual a proteína "query" pertence. Quanto maior o valor do nível do LCA, mais recente é o ancestral comum entre os dois organismos em comparação. A legenda é gerada automaticamente pelo programa FigTree e ela indica respectivamente o nível do LCA, o nome do táxon e a categoria taxonômica (ex. classe, ordem, família, etc.). Nota-se que, na legenda, existem alguns itens que possuem um asterisco no final. Estes indicam que nenhum organismo representado na árvore filogenética possui estes táxons como LCA. É o caso do nível 30 (Homo/genus) e o nível 24 (Haplorrhini/suborder) na árvore exemplificada na Figura 4A. A partir desse recurso visual, o usuário pode notar facilmente a relação taxonômica entre o organismo ao qual a proteína "query" pertence com os outros organismos representados na árvore, neste caso o homem (*Homo sapiens*). A não ser que o usuário seja especialista em taxonomia de mamíferos, sem este recurso seria difícil, por exemplo, um usuário saber a princípio que gibão (*Nomascus leucogenys*) e o coelho (*Oryctolagus cuniculus*) compartilham, respectivamente, a mesma superfamília e a mesma superordem com o homem. A árvore também nos permite acessar facilmente os organismos na árvore que possuem a maior e a menor distância taxonômica com o homem. Nesta árvore, o organismo que possui uma relação mais longínqua com o homem seria o ornitorrinco (*Ornithorhynchus anatinus*), que compartilha a mesma classe (Mammalia), enquanto o chimpanzé (*Pan troglodytes*) e o gorila (*Gorilla gorilla gorilla*) seriam os organismos mais próximos do homem, pois compartilham a mesma subfamília.

4.1.2.2. Categoria taxonômica

Outra forma de visualização que o arquivo gerado pela TaxOnTree permite ao usuário é pelas categorias taxonômicas. Abrindo a árvore gerado pela TaxOnTree no programa FigTree e navegando no menu lateral do programa, em *Appearance* → *Colour by* podemos escolher uma das categorias taxonômicas e permitir que os ramos sejam coloridos de acordo os táxons presentes na árvore que se enquadram na categoria taxonômica selecionada. A legenda da árvore também pode e deve ser alterada. Basta navegar no menu lateral do programa em *Legend* → *Attribute* e escolher a mesma categoria taxonômica selecionada para a coloração dos ramos. Na Figura 4B, é ilustrada a mesma árvore gerada a partir da análise da proteína AGXT2 humana, mas tendo os seus ramos coloridos de acordo com as superordens presentes na árvore. Por meio desta visualização podemos rapidamente verificar que existem cinco táxons na categoria superordem na árvore exemplificada e que todos eles formam um grupo monofilético.

4.1.3. Outras aplicações

As árvores geradas pela TaxOnTree focam em valorizar o caráter visual das distribuições taxonômicas inerentes nas árvores filogenéticas. No entanto, o programa também pode ser utilizado para abordar outros aspectos de cunho evolutivo. Nesta seção serão exemplificadas algumas destas abordagens que os arquivos de saída da TaxOnTree permitem realizar.

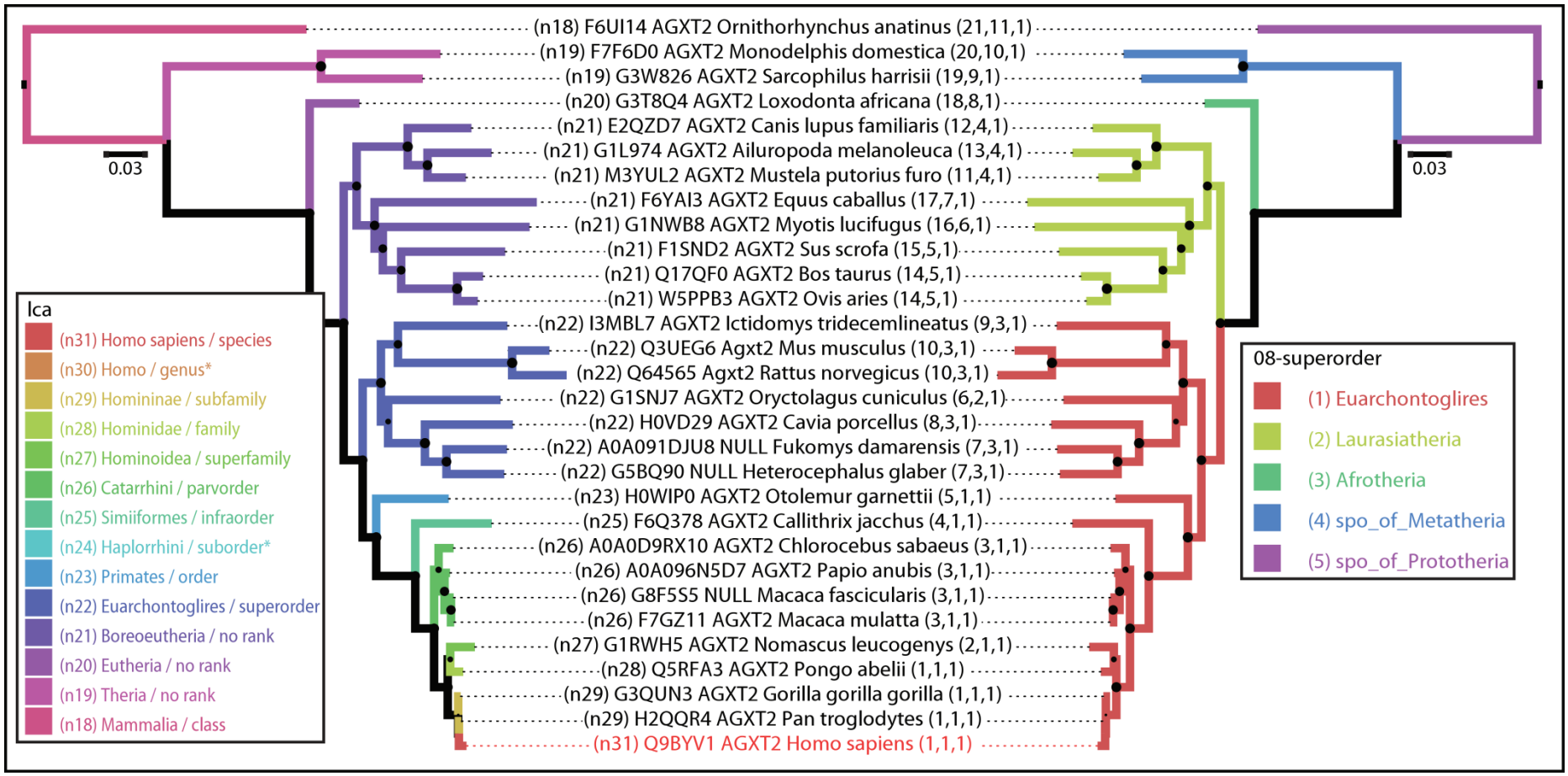


Figura 4: Exemplo de árvore gerada pela TaxOnTree utilizando o AGXT2 humana (Q9BYV1) como *query*.

Os ramos das árvores estão coloridos de acordo com o LCA (esquerda) e a categoria taxonômica superordem (direita).

4.1.3.1. Detecção de eventos de duplicação e deleção de genes

Uma vez que a TaxOnTree evidencia a distribuição taxonômica dos organismos encontrados em uma árvore filogenética, eventos de duplicação e deleção podem ser mais facilmente detectados ao longo de uma árvore filogenética gerada pela TaxOnTree. As árvores da Figura 5 foram geradas utilizando a proteína SLCO1B7 humana (*solute carrier organic anion transporter family member 1B7*; número de GI: 116812593). Os seus ortólogos foram recuperados do banco de dados RefSeq do NCBI com o corte de identidade de 50%. O alinhamento foi realizado com programa MUSCLE e o resultado do alinhamento foi analisado com o programa trimAl. Nesta análise foi verificado que, enquanto quase todos os organismos que não são da ordem primatas possuem apenas um membro desta subfamília (SLCO1B), as espécies da ordem primata, mais precisamente da subordem Haplorrhini, possuem até três membros proteicos desta família (SLCO1B1, SLCO1B3 e SLCO1B7) (Figura 5B). Na árvore, é possível identificar três nós que representam eventos de duplicação na linhagem dos primatas. Um deles deve ter ocorrido especificamente na linhagem do gênero *Tarsius*. Os outros dois eventos ocorreram no ancestral de todos os Simiiformes e deu origem aos outros dois membros da subfamília SLCO1B nas espécies desta infraordem. É possível verificar na árvore também que algumas espécies da infraordem Simiiformes não possuem um ou dois membros da subfamília SLCO1B citados anteriormente. Como todas as espécies de Simiiformes presentes na árvore possuem o genoma sequenciado (dados não mostrados), pode-se inferir que a ausência de alguns desses membros nestas espécies representam eventos de deleção. Assim como para os eventos de duplicação, podemos inferir em qual ancestral o evento de deleção pode ter ocorrido. No ramo onde estão representados os genes SLCO1B7, pode ser verificado que todas as espécies da subfamília Colobinae não possuem este gene, indicando que o evento de deleção deste gene nestas espécies aconteceu antes da diversificação da subfamília Colobinae.

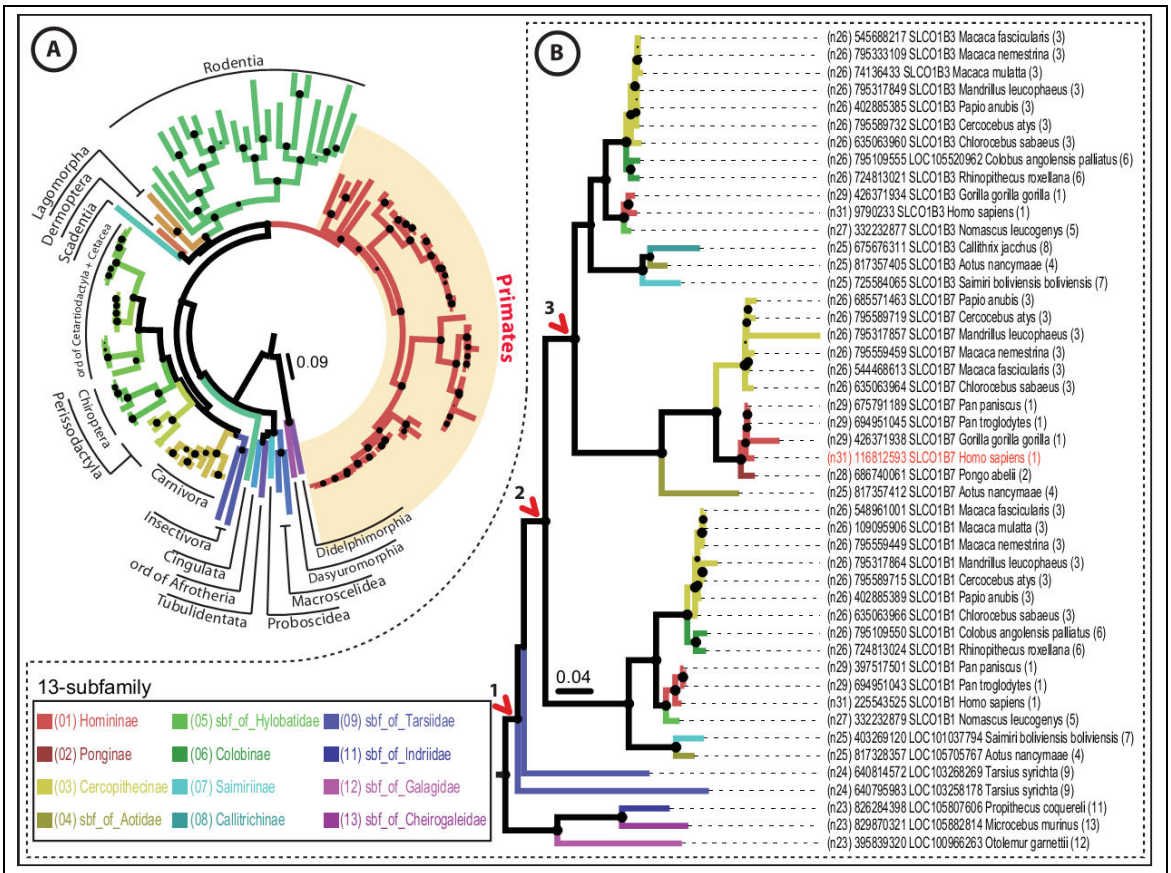


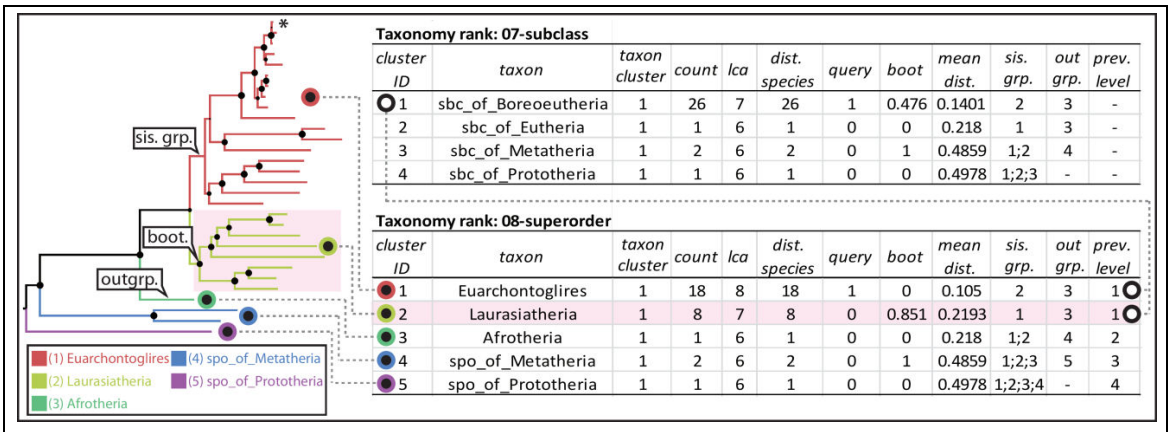
Figura 5: Árvore filogenética gerada pela TaxOnTree utilizando o gene SLCO1B7 humana (GI:116812593) como *query*.

Em (A), a árvore completa foi colorida de acordo com as ordens presentes na árvore. Em (B), a subárvore contendo apenas a ordem Primatas foi colorida de acordo com as subfamílias presentes na árvore. Setas vermelhas indicam os nós que correspondem a eventos de duplicação gênica.

4.1.3.2. Relatório taxonômico

O outro arquivo de saída que a TaxOnTree gera após a análise é um relatório contendo um sumário de todos os grupos taxonômicos que compõem a árvore filogenética (Figura 6). Neste relatório, o usuário pode verificar quais e quantos grupos taxonômicos encontram-se na árvore para cada categoria taxonômica, pois nem sempre os membros de uma superordem, por exemplo, estão agrupados em um ramo único. Além disso, para cada grupo taxonômico formado, outras informações úteis sobre o mesmo são apresentadas, como o número de sequencias dentro do grupo, o número de espécies distintas dentro do grupo, o valor de suporte de ramo e a distância média entre as amostras desse grupo com a amostra considerada como *query*.

O relatório também permite que o usuário tenha uma noção da topologia da árvore, informando, para cada grupo taxonômico, os grupos considerados irmãos e os grupos imediatamente externos ao grupo em análise. Além disso, o relatório apresenta uma coluna que relaciona um grupo taxonômico de uma categoria taxonômica com um grupo de uma categoria anterior. No relatório ilustrado na Figura 6, por exemplo, pode-se verificar que quando a análise chega ao nível "08-superorder" as superordens Euarchonthoglires e Laurasiatheria encontram-se no mesmo grupo "sbc_of_Boreoeutheria" na categoria taxonômica anterior ("07-subclass"). No nível "08-superorder" são detectados cinco grupos ("clusters"), sendo a "query" pertencente ao grupo 1 (número 1 na coluna "query"), compreendendo um ramo homogêneo agrupando 18 sequencias de 18 espécies distintas; o grupo 2 é um grupo irmão e o grupo 3, grupo externo; no nível 07-subclass as sequencias pertenciam ao grupo 1 correspondente.



Taxonomy rank: 07-subclass

cluster ID	taxon	taxon cluster	count	lca	dist. species	query	boot	mean dist.	sis. grp.	out grp.	prev. level
1	sbc_of_Boreoeutheria	1	26	7	26	1	0.476	0.1401	2	3	-
2	sbc_of_Eutheria	1	1	6	1	0	0	0.218	1	3	-
3	sbc_of_Metatheria	1	2	6	2	0	1	0.4859	1;2	4	-
4	sbc_of_Prototheria	1	1	6	1	0	0	0.4978	1;2;3	-	-

Taxonomy rank: 08-superorder

cluster ID	taxon	taxon cluster	count	lca	dist. species	query	boot	mean dist.	sis. grp.	out grp.	prev. level
1	Euarchontoglires	1	18	8	18	1	0	0.105	2	3	1
2	Laurasiatheria	1	8	7	8	0	0.851	0.2193	1	3	1
3	Afrotheria	1	1	6	1	0	0	0.218	1;2	4	2
4	spo_of_Metatheria	1	2	6	2	0	1	0.4859	1;2;3	5	3
5	spo_of_Prototheria	1	1	6	1	0	0	0.4978	1;2;3;4	-	4

Figura 6: Exemplo do relatório taxonômico gerado por TaxOnTree.
A árvore filogenética ilustrada na figura está com seus ramos coloridos de acordo com as superordens presentes na árvore. A amostra considerada *query* está evidenciada com asterisco na árvore. Legenda: *dist. species*: espécies distintas; *boot*: valor de suporte do nó; *mean dist.*: distância média entre a *query* e as amostras do cluster; *sis. grp.*: grupo irmão; *out grp.*: grupo imediatamente externo ao cluster; *prev. level*: identificador desse cluster no nível taxonômico anterior.

Usuários que possuem habilidades de programação podem aproveitar deste relatório para realizar uma análise programática das árvores geradas pela TaxOnTree em larga-escala. Para exemplificar uma das aplicações deste relatório, foram baixados 14.526 árvores filogenéticas geradas pelo método da máxima verossimilhança do banco de dados do OrthoMaM (DOUZERY *et al.*, 2014) e as árvores submetidas ao programa

TaxOnTree para gerar o relatório taxonômico de cada uma das árvores. O banco de dados de árvores filogenética do OrthoMaM possui uma coleção de árvore enraizadas constituídas por proteínas de mamíferos que possuem o seu genoma sequenciado. Nesta análise, foi desenvolvido um script em PERL que lê os relatórios taxonômicos e que classifica cada grupo taxonômico da categoria taxonômica superordem em uma das seguintes classes:

- *single*: o grupo taxonômico forma um grupo monofilético na árvore;

- *partial-in*: o grupo taxonômico se agrupa em um ramo juntamente com um pequeno número de sequencias de espécies de outros grupos taxonômicos inseridas;

- *partial-out*: a maioria das espécies de um grupo taxonômico encontra-se em um único ramo e algumas estão separadas na árvore;

- *multiple*: as espécies de um grupo taxonômico encontram-se espalhadas na árvore filogenética, não formando um grupo bem definido;

- *miss*: o grupo taxonômico em questão não se encontra na árvore em análise;

Analisando as quatro superordens de Eutheria (Euarchontoglires, Laurasiatheria, Afrotheria e Xenarthra), pode-se verificar que 5.310 árvores de genes (36,55%) possuem todas estas superordens formando um grupo monofilético cada qual. Se considerarmos as categorias *partial-in* e *partial-out* na contagem, o número de árvores de genes onde as quatro superordens são enquadradas em uma das três classes a contagem aumenta para 7.517 árvores de genes (51,74%). Todos os resultados apresentados nesta análise podem ser acessados no endereço eletrônico biodados.icb.ufmg.br/taxontree/orthomam. Exemplificamos assim como a TaxOnTree pode ser usada como um software ao invés de uma aplicação web, permitindo análises em larga escala.

4.1.3.3. Enraizamento da árvore filogenética baseado na informação taxonômica

A maioria dos software disponíveis para a reconstrução da árvore filogenética gera uma árvore não enraizada. Em casos onde a árvore não possui uma sequência considerada externa às demais sequências, o pesquisador pode recorrer ao método clássico do enraizamento pelo ponto médio ("Midpoint rooting"), que procura na árvore as duas folhas com a maior distância e cria um nó raiz no ponto em que divide esta distância em duas metades. Uma alternativa é procurar o melhor ponto de

enraizamento considerando as informações taxonômicas embutidas em uma árvore filogenética. Como esta última é comumente realizada de forma manual, foi implementado na TaxOnTree um algoritmo que automatiza esse procedimento utilizando as informações taxonômicas recuperadas no decorrer da análise.

Para ilustrar o funcionamento do algoritmo de enraizamento baseado nas informações taxonômicas, considere uma árvore não enraizada apresentada na Figura 7. Esta árvore apresenta três nós internos cada um ligados a três ramos. Considerando um desses nós, verifica-se que cada ramo ligado a ele conduz a um conjunto distinto de folhas da árvore e que se juntássemos esses diferentes conjuntos de folhas teríamos o conjunto todo das folhas presentes na árvore. Considerando estas propriedades, o algoritmo primeiramente obtém o conjunto total de espécies na árvore e visita cada um dos nós internos. Para cada nó interno visitado, o algoritmo escolhe um ramo ligado a este nó e elimina do conjunto total de espécies na árvore as espécies que compõe o ramo escolhido. As espécies que se mantiveram no conjunto total são submetidas para o cálculo do nível de LCA. Este procedimento é repetido para cada ramo ligado ao nó interno. Na árvore em questão, como todos os nós internos possuem três ramos ligados a ele, têm-se três níveis de LCA calculados para cada um deles. Em cada nó, o algoritmo interpreta que o ramo que deveria ser considerado como o ramo ancestral seria aquele que, quando o seu conjunto de folhas é retirado da análise do LCA, apresenta o maior nível de LCA. A estes ramos denominamos de ramos ancestrais putativos. Após o cálculo de todos os níveis de LCA, o algoritmo percorre novamente a árvore, desta vez visitando os ramos ancestrais putativos. Para cada ramo visitado, o algoritmo simula a criação de um nó raiz neste ramo e verifica quantos ramos ancestrais putativos de fato encontram-se como um ramo ancestral após o enraizamento. O algoritmo escolherá o ponto de enraizamento que apresenta o maior número de ramos ancestrais putativos em concordância. O benefício de se utilizar este algoritmo se encontra ilustrado na Figura 8, quando comparamos uma árvore que não passou por um método de enraizamento, com a mesma árvore que foi enraizada pelo ponto médio ou baseando-se nas informações taxonômicas. Verifica-se que nesta árvore, o método de enraizamento baseado nas informações taxonômicas apresenta uma árvore que melhor concorda com a evolução das espécies que constituem a árvore. Note que um grupo marcado por (n13) reúne sequencias que compartilham com o morcego o clado Euteleostomi, ficou posicionado próximo do ponto de enraizamento.

As duas abordagens de enraizamento (“Midpoint” e “Taxonomy-based rooting”) encontram-se implementados no TaxOnTree.

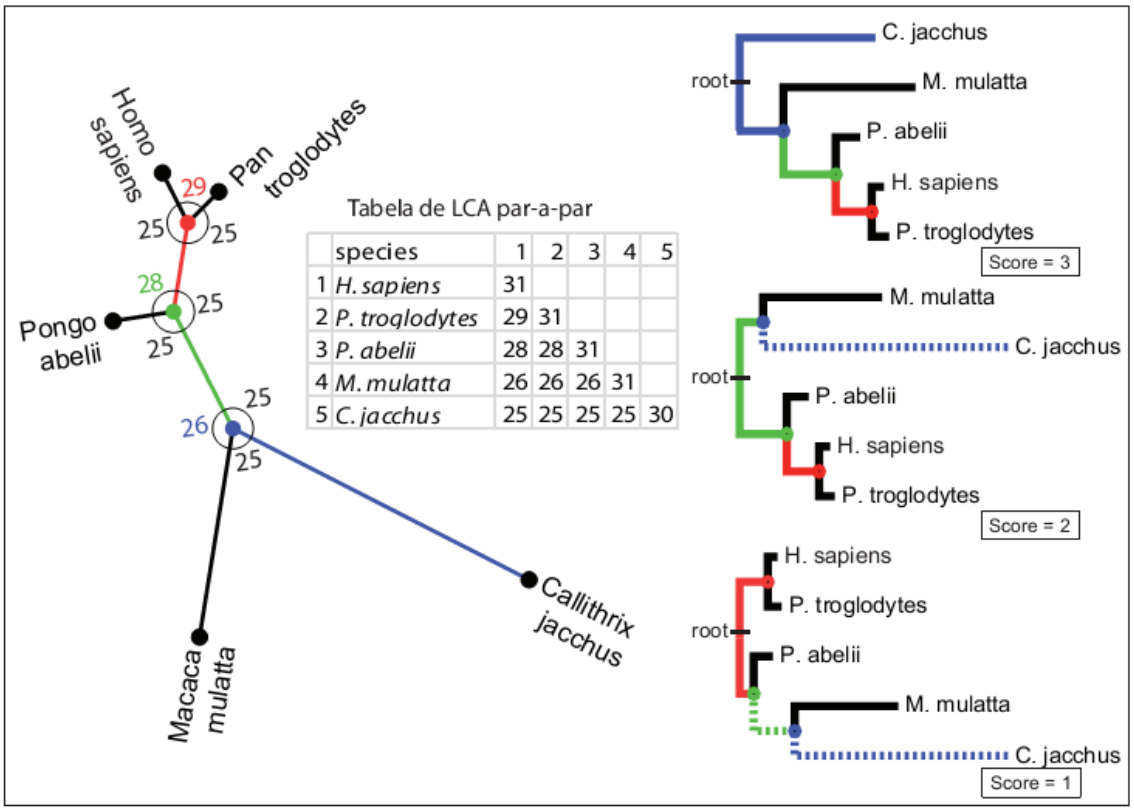


Figura 7: Enraizamento da árvore baseado em dados taxonômicos.

À esquerda, árvore não enraizada constituída por alguns primatas. Os números nos nós dessa árvore indicam o LCA entre os organismos que constituem dois ramos (ligados pelo arco em cada nó) e determina os ramos que deve ser considerado ancestral (ramos coloridos). O nível de LCA é determinado através de uma tabela que contém os níveis de LCA entre pares de organismo (meio da figura). Depois de determinado os LCAs em cada nó interno da árvore, o algoritmo simula o enraizamento da árvore (à direita) e para cada local de enraizamento, ele determina uma pontuação que leva em conta se os ramos considerados ancestrais encontram-se em acordo.

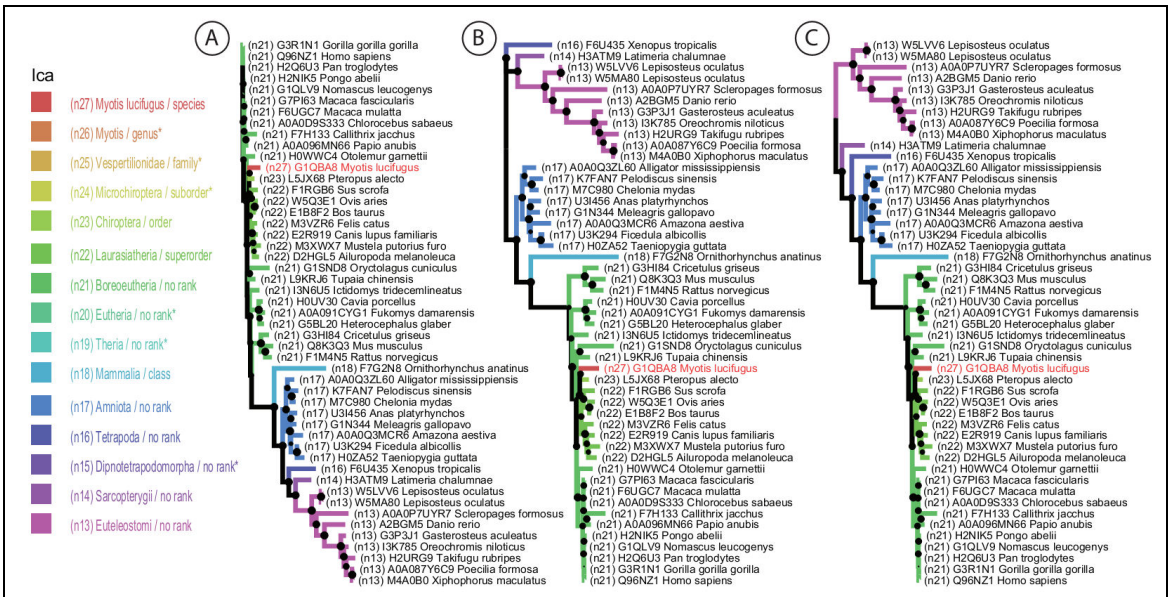


Figura 8: Árvores filogenéticas geradas pela TaxOnTree utilizando a proteína de morcego com o número de acesso G1QBA8.

Nesta figura estão representadas árvores: que não foi enraizada (A), que foi enraizada pelo método do ponto médio (B) e que foi enraizada pelo método baseado em informações taxonômicas (C).

4.1.4. Outras funcionalidades implementadas na TaxOnTree

Nesta seção encontram-se listadas outras funcionalidades da TaxOnTree que podem ser úteis ao usuário durante a análise que podem melhorar a apresentação da árvore final.

4.1.4.1. Descarte das isoformas de proteínas

Isoformas de proteínas são diferentes formas de proteínas que o mesmo gene pode codificar através do "splicing" alternativo. As isoformas podem ser encontradas nos bancos de dados de proteínas do RefSeq (NCBI), mas muitas vezes elas não são o foco da análise e acabam atrapalhando a visualização da árvore já que a sua presença pode aumentar dramaticamente o número de proteínas em uma árvore. Para esta situação, foi implementado na TaxOnTree um filtro que leva em consideração o identificador de gene (GeneID) atribuído a cada proteína. Proteínas que possuem o mesmo GeneID são proteínas codificadas pelo mesmo gene e, portanto, são isoformas. Caso o usuário opte em descartar as isoformas na análise, a TaxOnTree selecionará uma

isoforma representativa para cada gene representado na árvore (Figura 9). A isoforma selecionada é aquela que apresentou maior identidade com a proteína considerada "query".

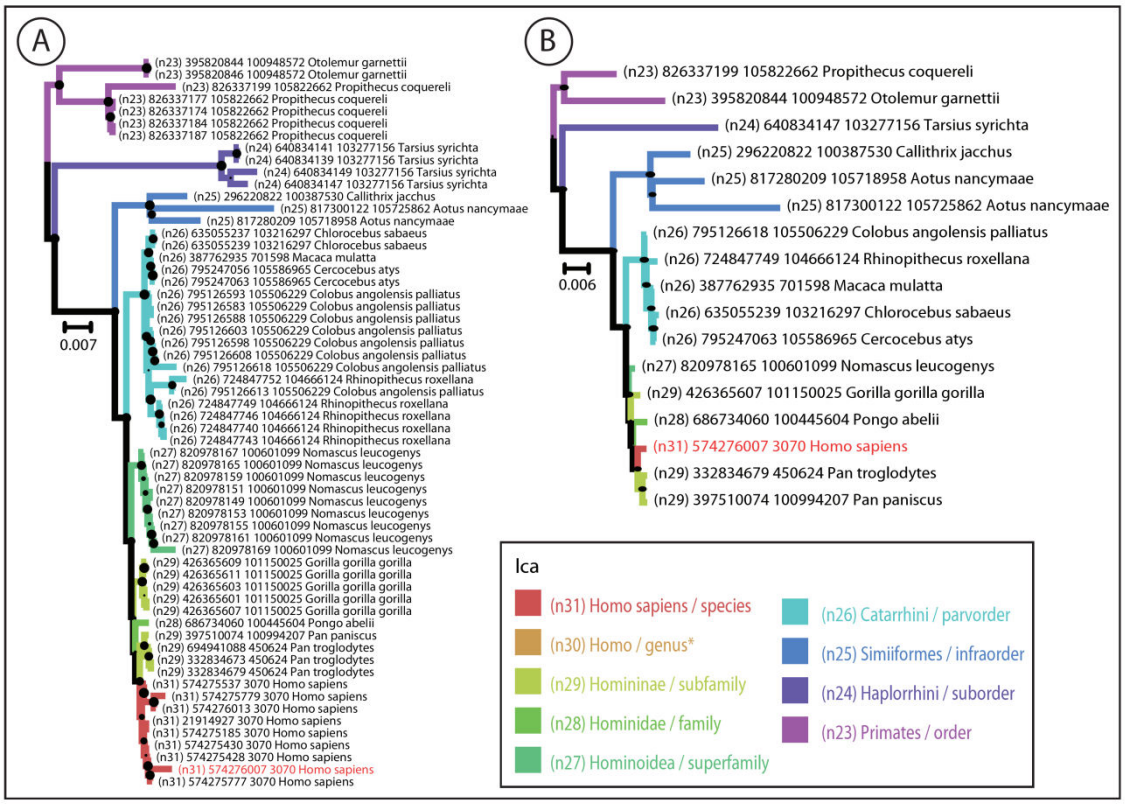


Figura 9: Árvores filogenéticas gerada pela TaxOnTree utilizando uma proteína humana com o número GI 574276007 e com os ramos coloridos de acordo com o LCA.

A árvore da esquerda foi gerada sem o filtro de isoforma, enquanto a árvore da direita foi gerada com o filtro de isoforma.

4.1.4.2. Formatação dos nomes das folhas

Os nomes presentes nas folhas são as partes da árvore que mais chamam a atenção dos usuários, uma vez que é através delas que acessamos as informações primárias sobre alguma amostra específica. Se as folhas contiverem informações apropriadas, elas podem facilitar o usuário durante a interpretação da árvore reduzindo o tempo que ele poderia gastar caso necessite buscar essas informações por outros meios. No entanto, as árvores geradas a partir dos programas de construção de árvore

filogenética apresentam nos nomes das folhas, em geral, o cabeçalho das sequências em FASTA submetidos ao programa, que em muitos casos contém apenas o número de acesso da proteína. Usuários podem ainda encontrar problemas mais específicas para determinados programas, como o limite de caracteres permitidos ou restrições quanto ao uso de alguns caracteres nos nomes das folhas. Dentro desta abordagem, a TaxOnTree foi implementada de forma a permitir ao usuário customizar o nome das folhas da árvore final, permitindo a ele incluir ou retirar algumas informações sobre a proteína que são recuperadas ao longo da análise. Atualmente, as informações que a TaxOnTree pode incluir nas folhas das árvores estão listadas abaixo:

- *Primary ID* – Número de GI (para acessos do RefSeq) ou número de acesso (para acessos do Uniprot);
- *Alternative ID* – Número de acesso (para acessos do RefSeq) ou nome de entrada (*Entry Name*; para acessos do Uniprot);
- *Fasta header* – Nome original extraído do arquivo Multi-FASTA;
- *LCA level* – Nível de LCA. Corresponde ao número de nós taxonômicos encontrado entre o nó raiz e o nó que representa o menor ancestral comum entre dois organismos, sendo um dos organismos é aquele associado à proteína *query*.
- *Gene name* – Abreviação oficial do gene fornecido pela base Gene do NCBI;
- *Species name* – Nome científico da espécie ou da linhagem;
- *Taxonomy rank code* – Cada grupo taxonômico presente em uma categoria taxonômica é identificado por um código numérico que representa a ordem de proximidade dos grupos com o grupo taxonômico que contém o organismo associado à proteína *query*. Em outras palavras, o grupo taxonômico que contém o organismo associado à proteína *query* é identificado como grupo 1, e os demais grupos taxonômicos são identificados pelos números sucessivos de acordo com a sua proximidade (distância filogenética) com o grupo 1. Estes números são os mesmos apresentados nas legendas da árvore para uma dada categoria taxonômica e podem ser acrescentadas aos nomes das folhas;
- *Taxonomy rank name* – Mostra o nome do grupo taxonômico que a amostra pertence de uma dada categoria taxonômica selecionada pelo usuário.

O único requisito deste recurso é que o usuário selecione pelo menos o item *Primary ID* ou *Alternative ID* para garantir que a árvore final não contenha folhas com nomes duplicados. Isso é necessário uma vez que o programa FigTree não permite abrir

árvores que contenham folhas ambíguas. Esse recurso possibilita a publicação de árvores que permitem inspeção visual muito clara.

4.1.4.3. Arquivo de saída em SVG

Dentro do pacote do programa TaxOnTree, foi disponibilizado um script escrito em PERL que executa o programa FigTree em linha de comando para gerar automaticamente a árvore filogenética em formato SVG. O script ("nexus2svg.pl") tem como entrada o arquivo Nexus gerado pelo programa TaxOnTree. O usuário pode escolher o modo de coloração dos ramos, se é pelo critério do LCA ou por uma das categorias taxonômicas. Este mesmo script foi utilizado para gerar as 14.526 árvores do banco de dados do OrthoMaM no formato SVG apresentados no endereço eletrônico: biodados.icb.ufmg.br/taxontree/orthomam.

4.1.5. Webservice da TaxOnTree

Para atender aqueles usuários que não possuem treinamento para executar programas em linha de comando no ambiente UNIX, ou que possuem uma pequena demanda do programa, foi gerado um Webservice que reúne a maior parte das funcionalidades do programa TaxOnTree em um único endereço eletrônico: biodados.icb.ufmg.br/taxontree (Figura 10).

Para executar uma tarefa no Webservice da TaxOnTree basta navegar na primeira aba da página "Run TaxOnTree", escolher o tipo de entrada que você quer fornecer ao programa e colocar na caixa de texto a sua entrada ou carregar um arquivo que contém a sua entrada. Antes de submeter a sua tarefa, você pode definir os parâmetros na aba central "Options". Nesta aba é possível escolher, por exemplo, um banco de dados de sequência para a busca de proteínas ortólogas putativas, escolher o alinhador, ativar/desativar o filtro de isoforma e escolher o método de enraizamento. Após a escolha de todos os parâmetros, você pode submeter a tarefa clicando no botão "Submit". Ao submeter a tarefa ao programa, um identificador único da sua tarefa ("jobID") surgirá na página. Este identificador deve ser guardado para que você consiga recuperar os resultados da sua tarefa. Os resultados da sua tarefa podem ser acessados utilizando a última aba da direita "Get your result". Nesta aba, você deve digitar o identificador da sua tarefa e selecionar o tipo de arquivo que deseje baixar. Os usuários

do Webservice têm a opção de baixar o arquivo de árvore no formato Nexus (arquivo principal da ferramenta, para ser aberto com FigTree) ou no formato SVG (imagem pronta), o relatório taxonômico, o alinhamento de sequência que foi submetido para a construção da árvore e o resultado do BLAST. Caso tenha ocorrido algum erro durante a análise do TaxOnTree, o usuário pode ainda baixar o arquivo log da sua tarefa e verificar o erro que o programa encontrou na sua tarefa.

É pertinente enfatizar que o usuário não necessita abrir uma conta para acessar a todas as funcionalidades deste Webservice. Qualquer usuário pode submeter as suas tarefas como convidado ("guest") e acessar os resultados na própria página. Mas qualquer usuário possui a opção de criar uma conta e ter as suas tarefas organizadas em tabela na própria página do TaxOnTree. A abertura da conta necessita apenas que o usuário defina um nome e uma senha de acesso. Após a criação e o acesso da conta, a aba inferior "Job tables" ficará disponível para o usuário (Figura 11). Nesta tabela, o usuário pode ter acesso a todas as tarefas submetidas enquanto logado nesta conta. Cada linha da tabela corresponde a uma tarefa submetida pelo usuário e contém algumas informações pertinentes da tarefa como o identificador da tarefa, o "status" da tarefa, tipo de tarefa, alguns valores dos parâmetros utilizados, dentre outros. Além disso, cada linha contém uma coluna denominada "Action" que contém ícones que, ao serem clicados, baixam o arquivo de saída correspondente. Por fim, caso seja do interesse do usuário, ele também pode remover as informações de alguma tarefa que consta na tabela, clicando no ícone da lixeira.

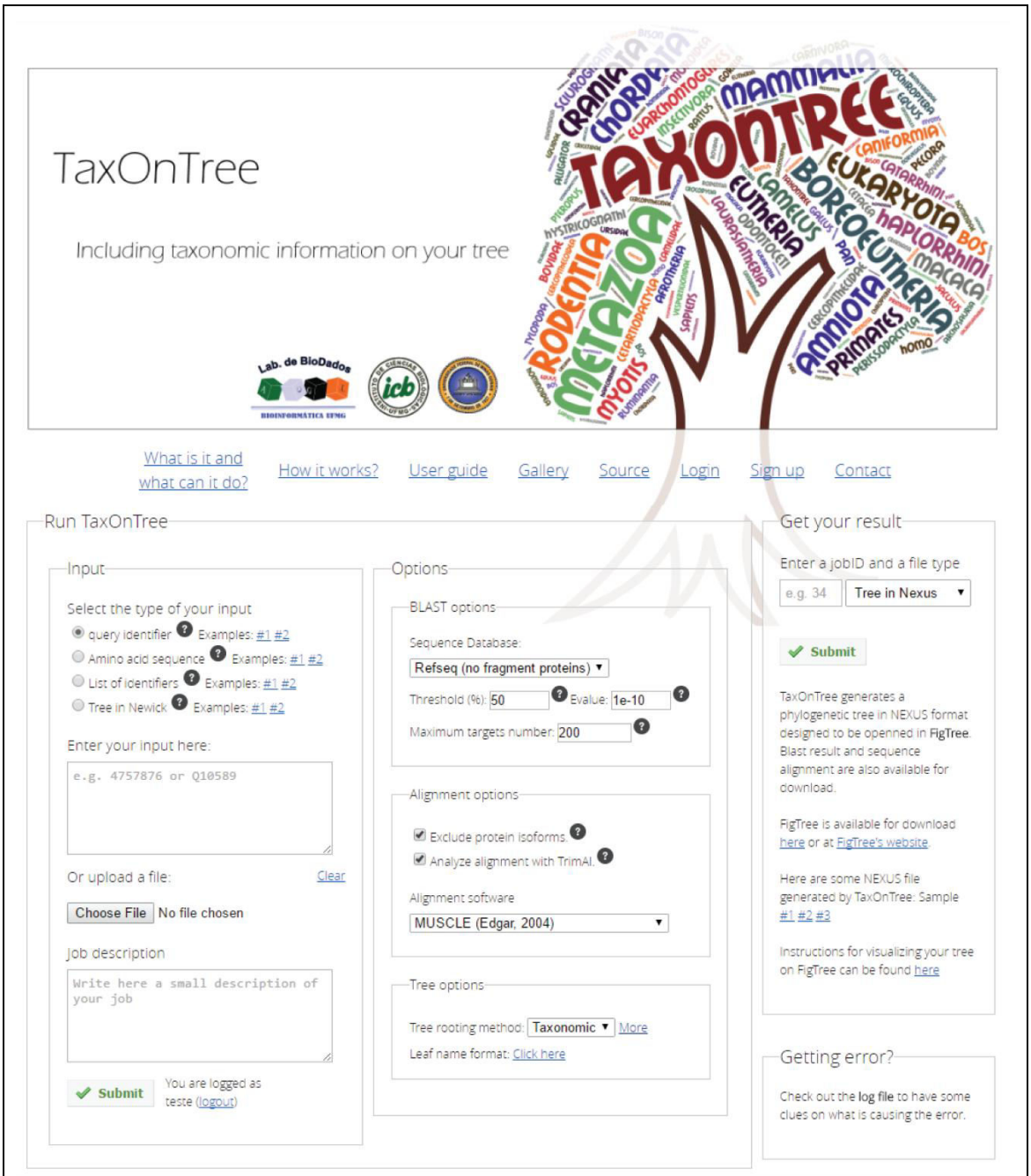


Figura 10: Página de entrada do Webservice da TaxOnTree.

As tarefas do usuário podem ser submetidos utilizando a aba “Run TaxOnTree” à esquerda. Os parâmetros podem ser alterados utilizando a aba “Options” no meio. Os resultados de uma tarefa podem ser acessadas utilizando a aba “Get results” à direita.

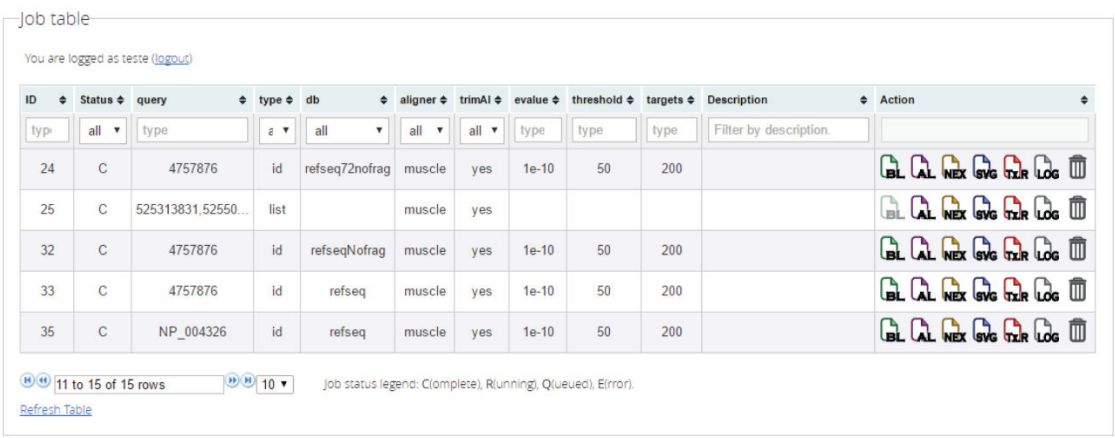


Figura 11: Tabela das tarefas acessível a aqueles usuários que cadastraram uma conta no Webservice da TaxOnTree.

Os resultados de cada tarefa podem ser acessados clicando em um dos ícones na coluna *Action*.

4.2. HyperTriplet

Com o desenvolvimento do programa TaxOnTree e juntamente com o desenvolvimento da capacidade de manipular árvores filogenéticas programaticamente ao longo do projeto, outras ferramentas foram elaboradas para facilitar e melhorar a análise de grande número de árvores filogenéticas. Um dos tópicos que foi possível abordar neste trabalho se refere à reconstrução da árvore de espécie a partir de dados moleculares (árvores gênicas). Na filogenia molecular existem duas principais abordagens utilizadas para a reconstrução da árvore de espécie. Uma delas é referida como superárvore que se baseia em várias árvores de genes para a reconstrução de uma árvore de espécies. Pode-se dizer que uma superárvore representa uma árvore consenso das árvores de genes em análise. Esta difere da segunda abordagem que é da supermatriz, que consiste em reconstruir uma árvore de espécie a partir de um único arquivo de alinhamento de sequências resultante da concatenação de alinhamentos de sequências de diferentes genes ou proteínas. Esta seção abordará a ferramenta denominada HyperTriplets, um programa que gera uma superárvore utilizando o conceito dos tripletos nas árvores filogenéticas.

4.2.1. Algoritmo do HyperTriplets

O HyperTriplets recebe como arquivos de entrada, múltiplas árvores no formato Newick. De forma resumida, a partir destas árvores, o programa constrói uma tabela contendo informações topológicas (tripletos) e de distância filogenética. Estes dados são posteriormente utilizados para a busca da melhor topologia da árvore. Determinada a melhor topologia, o algoritmo então incluirá os dados de distância nos ramos desta árvore. O programa ainda conta com um algoritmo que lida com árvores de genes que contém parálogos. Os itens a seguir descreverão com mais detalhes cada passo realizado pelo programa HyperTriplets para a confecção da superárvore.

4.2.1.1. Busca pela melhor topologia

Os tripletos são porções da árvore filogenética constituídas por três amostras (ou folhas). A partir de uma árvore que contém quatro amostras (Figura 12), por exemplo, podem-se formar quatro tripletos distintos, onde cada tripleto é resultante da remoção de uma folha da árvore. Cada tripleto extraído é classificado em uma das três formas possíveis de topologia que um conjunto de três amostras pode formar. Para exemplificar estas três possíveis topologias considere A, B e C como amostras que fazem parte da árvore. As possíveis topologias que estas amostras podem apresentar na árvore são: (1) A e B são irmãos e C é externa (C|A,B), (2) A e C são irmãos e B é externa (B|A,C) ou (3) B e C são irmãos e A é externa (A|B,C). Neste contexto, o algoritmo determina para cada tripleto o par de amostras que são irmãos entre eles e, consequentemente, a amostra que é externa às outras duas. Cada tipo de topologia de tripletos encontrado é contabilizado em uma tabela e esta é utilizada posteriormente para auxiliar na montagem da superárvore.

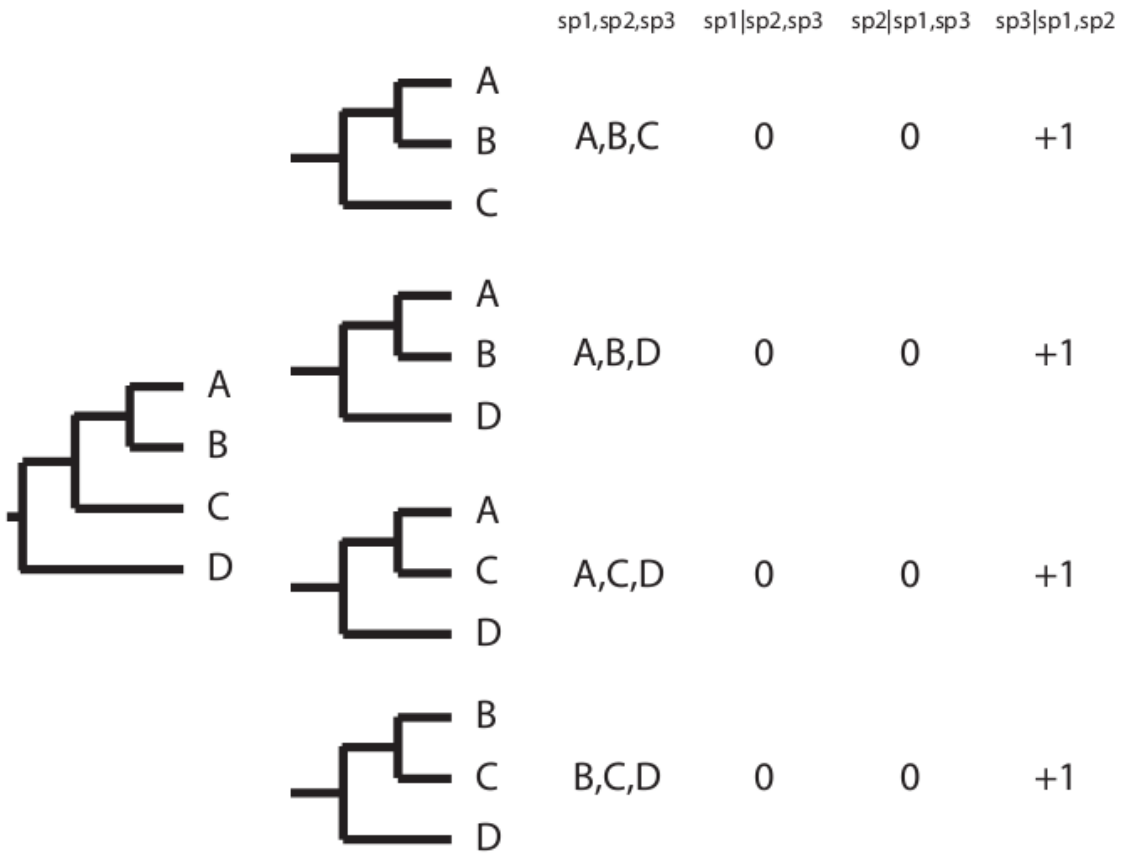


Figura 12: Decomposição da árvore filogenética em tripletos.

A árvore da esquerda pode ser decomposta em quatro tripletos. Os tripletos gerados são contabilizados em uma tabela (direita) de acordo com a sua topologia.

Após determinar e contar todos os tripletos presentes nas árvores fornecidas pelo usuário, o algoritmo do HyperTriplets cria uma árvore inicial aleatória, enraizada e dicotômica constituída pelos organismos presentes nas árvores de genes (Figura 13). Desta árvore inicial, o algoritmo percorrerá os seus nós internos e verificará se a topologia encontrada ao analisar este nó é a mais frequente na tabela de contagens da topologia dos tripletos. Para explicar esta etapa considere antes que, em uma árvore dicotômica e enraizada, de cada nó interno partem três ramos: um ramo que o liga a um nó ancestral (ramo ancestral) e dois ramos que o ligam a nós descendentes (ramos descendentes). Os dois ramos descendentes constituem-se de amostras que são irmãs entre elas e serão designados como ramo 1 e ramo 2. Considere também que do nó ancestral partem dois ramos descendentes. Um deles é o próprio ramo que o liga ao nó interno em análise, já o segundo é um ramo constituído de amostras imediatamente externo às amostras do nó em análise, e este será designado como ramo 3. A partir de

um nó interno, então, o algoritmo extrai três conjuntos de amostras da árvore que constituem cada um dos ramos designados anteriormente (ramos 1, 2 e 3). As amostras do ramo 1 são irmãs das amostras do ramo 2, enquanto as amostras do ramo 3 são externas às amostras dos ramos 1 e 2. Após a determinação desses três conjuntos de amostras, o algoritmo pegará uma amostra de cada conjunto e consultará a tabela de contagem da topologia dos tripletos e determinar se a topologia atual da árvore é a mais frequente na tabela. Se um ou mais conjuntos apresentarem mais de uma amostra, todas as combinações possíveis de amostras partindo dos três conjuntos serão analisadas. No final, uma média aritmética será calculada para determinar qual das três topologias é a mais frequente na tabela. Caso a topologia atual seja a mais frequente, esta topologia se mantém e o algoritmo se direciona para o próximo nó interno. Caso contrário, o algoritmo mudará a topologia da árvore para a topologia mais frequente na tabela. O valor de suporte inserido neste nó corresponde à frequência (ou a frequência média) da(s) topologia(s) selecionada(s) encontrada(s) na tabela e pode variar de 0,33 a 1. O algoritmo continuará a análise até que todos os nós internos apresente a topologia mais frequente determinada pela tabela dos tripletos.

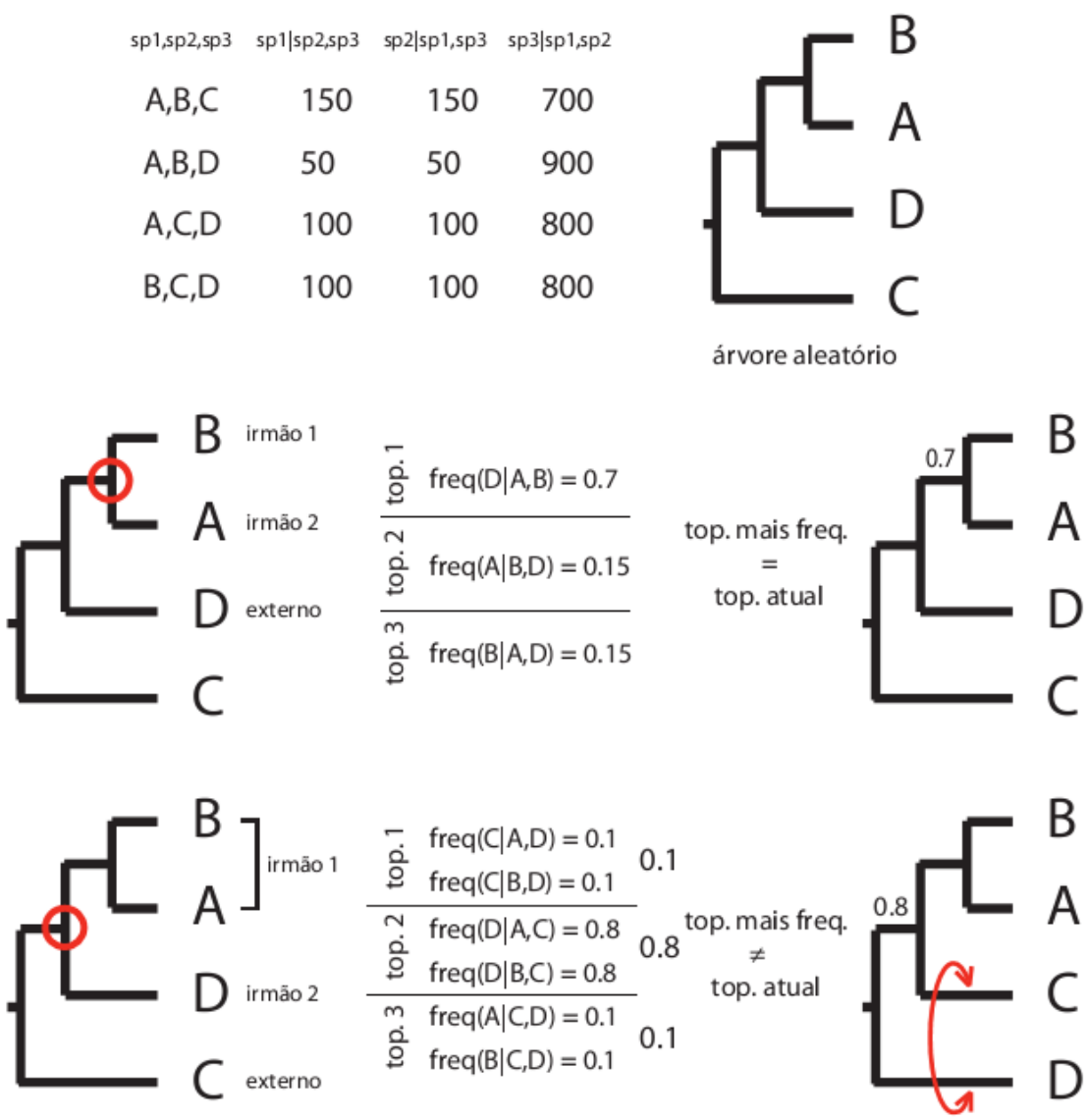


Figura 13: Construção da superárvore a partir da tabela da frequência dos tripletos.

Uma árvore aleatória contendo os organismos em análise é gerada. Posteriormente, a topologia desta árvore é mantida ou trocada baseado nas frequências de tripletos em cada topologia.

4.2.1.2. Inserindo dados de distância na árvore

O programa HyperTriplets também guarda dados de distância entre as folhas de cada árvore filogenética fornecido pelo usuário. Após o processamento de todas as árvores, o programa gera uma segunda tabela contendo a distância média encontrada entre duas amostras (Figura 14). Esta segunda tabela funciona como uma matriz de

distância e é utilizada posteriormente para atribuir o comprimento dos ramos à árvore que apresentou a topologia mais adequada em relação à tabela da contagem dos tripletos. Para a inclusão dos dados de distância, o algoritmo percorrerá novamente os nós internos (n) da árvore e extrairá em cada visita os mesmos três conjuntos de amostras citados na etapa anterior: dois conjuntos de amostras dos dois ramos descendentes (ramo 1 e 2) e um conjunto de amostras do ramo ancestral (ramo 3). Consultando a tabela de distância entre as folhas, podem-se obter três valores de distância: (1) a distância entre a amostra do ramo 1 e a amostra do ramo 2 (A_n), (2) a distância entre a amostra do ramo 1 e a amostra do ramo 3 (B_n) e (3) a distância entre a amostra do ramo 2 e a amostra do ramo 3 (C_n). Se um determinado conjunto contiver mais de uma amostra, o algoritmo utilizará a maior distância encontrada para os cálculos posteriores. A partir desses dados de distância, pode-se formar um sistema de três equações com três incógnitas, sendo cada incógnita (x_n, y_n e z_n) a distância do nó até a amostra de um dos conjuntos. Estes valores serão, posteriormente, utilizados pelo algoritmo para determinar o comprimento de cada ramo presente na árvore como demonstrado na Figura 14.

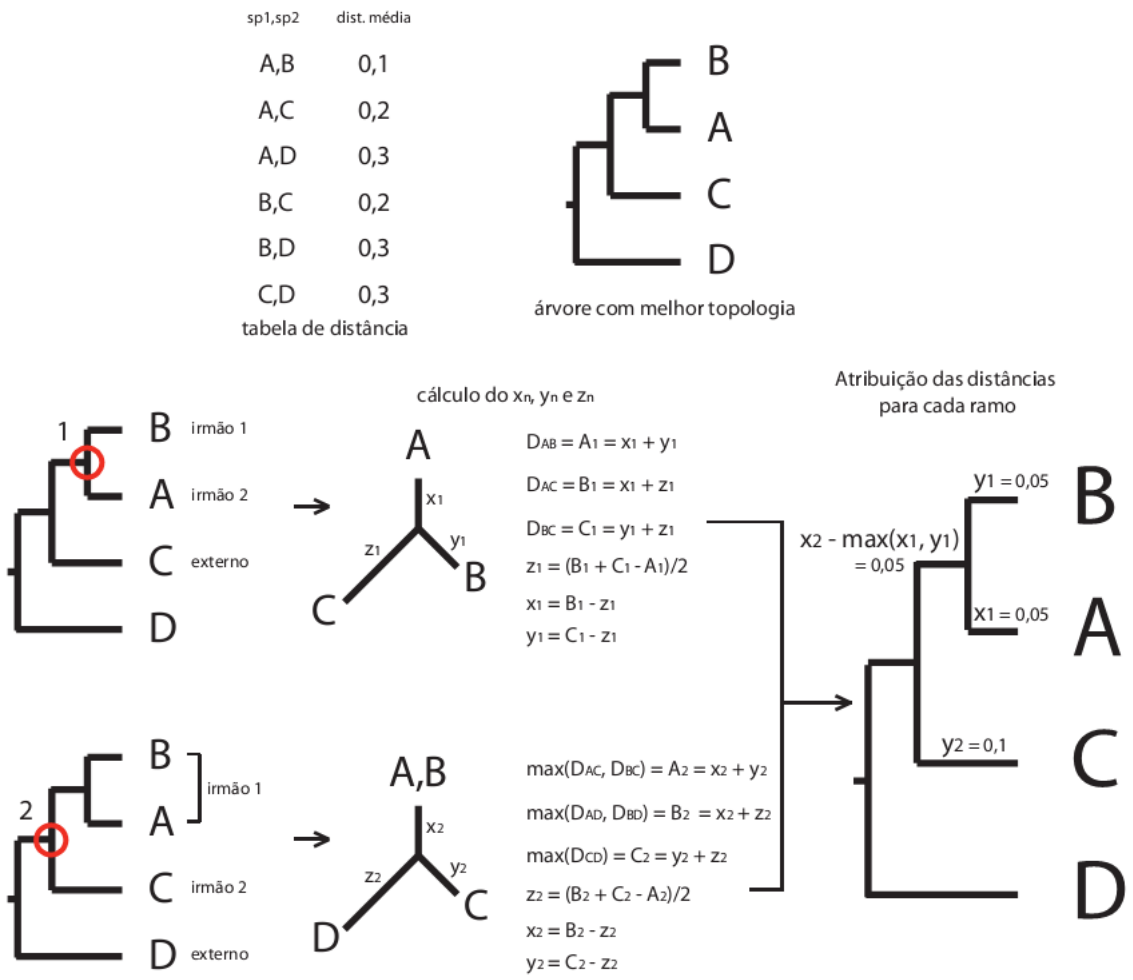


Figura 14: Incorporação dos dados de distância dos ramos da superárvore baseado na tabela da distância média entre as amostras.

A distância de um ramo é determinada por um sistema de três equações e três incógnitas.

4.2.1.3. Lidando com árvores de genes que contém parálogos

Árvores que contém parálogos impõem mais um nível de dificuldade para a elaboração de um algoritmo de reconstrução da superárvore, já que para a elaboração das relações de ancestralidade entre as espécies presentes na árvore devem ser realizadas considerando os eventos de duplicação. Para possibilitar a inclusão de árvores de genes com parálogos na análise, foi implementado no programa HyperTriplets um algoritmo que identifica os nós de duplicação e divide a árvore nesses nós de forma a gerar duas ou mais árvores sem os nós de duplicação. O número de árvores geradas a partir desse processo depende da quantidade de nós de duplicação presente na árvore.

Cada uma das árvores formadas a partir de uma árvore de gene pode ser analisada individualmente quanto à contagem dos tripletos e, posteriormente, calcular a média desses resultados e contabilizá-los na tabela definitiva. Existe uma situação em que o algoritmo não divide a árvore quando encontra um nó de duplicação. Isto acontece quando o nó de duplicação representa uma expansão gênica específica para uma linhagem ou espécie. Neste caso, o algoritmo fundirá as amostras de genes que sofreram expansão em uma única amostra e recalculará a distância desta amostra até o nó ancestral considerando a média das distâncias entre este nó ancestral e as amostras expandidas. A Figura 15 ilustra algumas situações resolvidas por esse algoritmo.

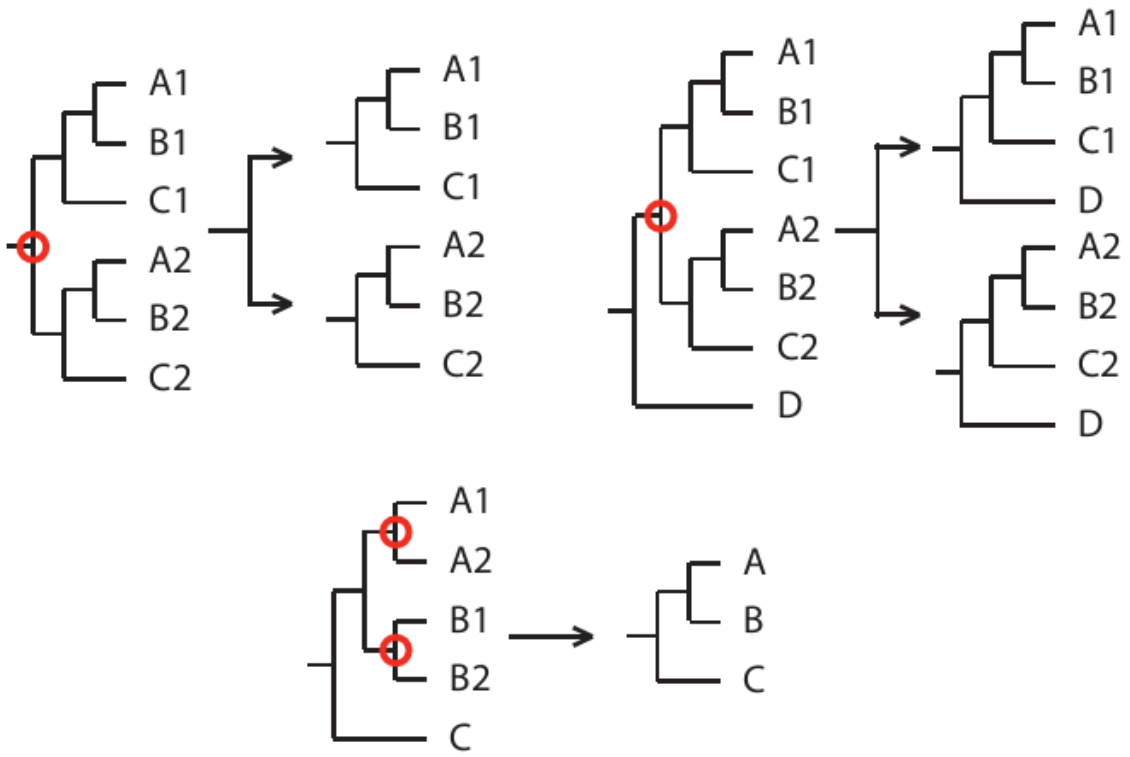


Figura 15: Esquemas de como o algoritmo de partição da árvore lida com as árvores que apresentam nós de duplicação gênica.

4.2.2. Aplicação em dados reais

Nesta seção abordaremos alguns resultados obtidos pelo programa HyperTriplets utilizando dados reais de árvores filogenéticas de organismos de diferentes clados taxonômicos.

4.2.2.1. Dados de mamífero

O programa HyperTriplets foi utilizado para elaborar uma árvore de espécie contendo mamíferos que possuem o seu genoma sequenciado e anotado. Para isso, árvores filogenéticas foram recuperadas do banco de dados OrthoMaM (DOUZERY *et al.*, 2014), que reúne 14.526 árvores de genes contendo 43 espécies de mamíferos. As árvores armazenadas neste banco encontram-se enraizadas e foram geradas a partir de dados de alinhamento de nucleotídeos baseado em códon e de método da máxima verossimilhança utilizando o programa RAxML (STAMATAKIS, 2006). A única observação que deve ser feita sobre estas árvores é que elas constituem-se apenas de genes considerados ortólogos do tipo 1:1, ou seja, presente em cópia única no genoma.

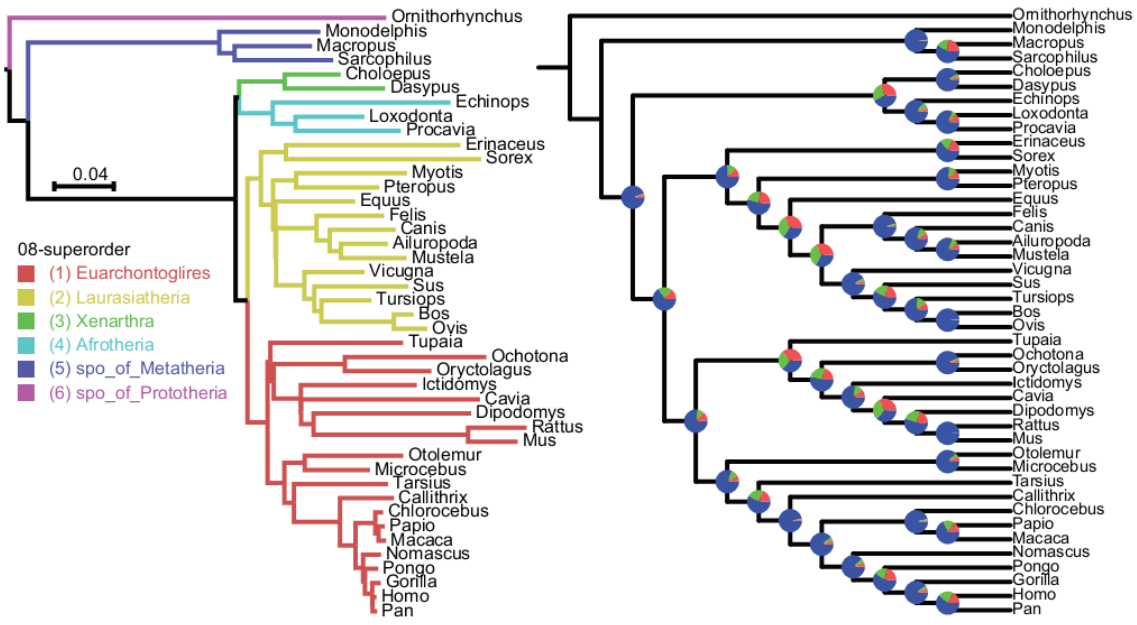


Figura 16: Superárvore gerada pelo HyperTriplet utilizando dados de mamífero. À direita, os ramos da árvore foram coloridas de acordo com as superordens presentes na árvore. À esquerda, a árvore foi colocada em cladograma, e em cada nó interno tem representado os valores de suporte em um gráfico de pizza. A porção em azul representa o suporte para a topologia atual. As outras duas porções (verde e vermelha) representam o suporte para topologias alternativas.

A superárvore gerada pelo HyperTriplets utilizando as 14.526 árvores como entrada é apresentada na Figura 16. Evidenciando os grupos taxonômicos da categoria superordem na árvore, utilizando o programa TaxOnTree, observa-se que a árvore

resultante organiza todos os seis grupos taxonômicos em um ramo monofilético. Dos 40 nós internos que possuem um valor de suporte, 19 (47,5%) deles apresentaram um valor acima de 0,70 e 5 (12,5%) apresentaram um valor abaixo de 0,50. Os nós que apresentaram um baixo valor de suporte representam as partes da evolução dos mamíferos que ainda gera dúvidas e discussões. Os cinco nós que apresentaram baixo suporte refletem aos quatro questionamentos sobre a história evolutiva dos mamíferos: (1) posicionamento do gênero *Tupaia*; (2) posicionamento do gênero *Cavia* e *Dipodomys*; (3) posicionamento das quatro ordens de Laurasiatheria (Perissodactyla, Carnivora, Cetartiodactyla e Chiroptera) e o posicionamento das superordens Xenarthra e Afrotheria. É possível observar também que os nós que apresentaram baixo suporte são antecedidos por ramos que possuem pequeno comprimento. Relacionando os valores de suporte dos nós com o comprimento dos ramos, é possível verificar que ramos mais longos tendem a apresentar valores de suporte mais alto e que nós com valores de suporte mais baixo possuem ramos curtos (Figura 17). Além disso, observa-se que ramos que possuem um comprimento acima de 0,02 substituições de nucleotídeo/sítio possuem valores de suporte acima de 0,8.

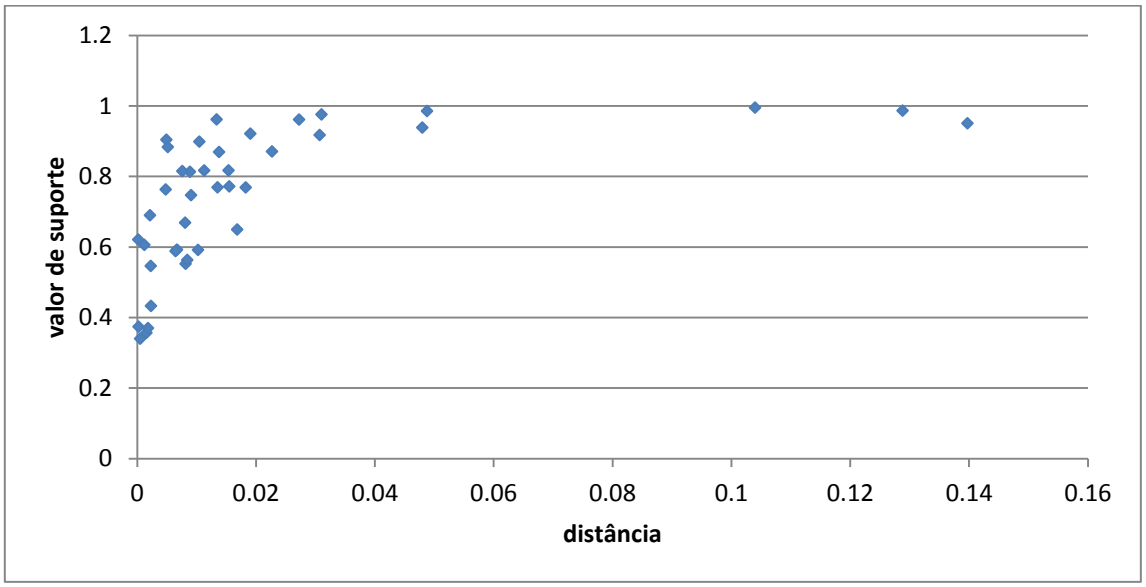


Figura 17: Relação entre a distância e o suporte do ramo da superárvore da Figura 16.

Dados de alinhamento de proteína disponível também no banco de dados do OrthoMaM também foram recuperados e utilizados para a geração da superárvore. Cada arquivo de alinhamento foi submetido ao programa RAxML e as árvores resultantes

foram submetidas ao programa HyperTriplets. A superárvore gerada a partir desses dados (Figura 18) apresentou uma topologia semelhante àquela apresentada pelos dados de alinhamento por códon. Dentre as diferenças topológicas encontradas foram o posicionamento do gênero *Tupaia* dentro do clado do Euarchontoglires e o posicionamento do gênero *Equus* (cavalo) dentro do clado do Laurasiatheria. Além disso, o suporte dos nós apresentou valores menores na superárvore gerada a partir de dados de aminoácido (valor médio = 0,6291) em relação à superárvore gerada a partir dos dados de nucleotídeos (valor médio = 0.7408). Dos 40 nós internos que possuem um valor de suporte, apenas cinco apresentaram um valor acima de 0,80 (20%), no entanto, os nós que apresentaram um valor de suporte menor que 0,50 foram os mesmos cinco nós (12.5%) apresentado na superárvore baseado em dados de nucleotídeos.

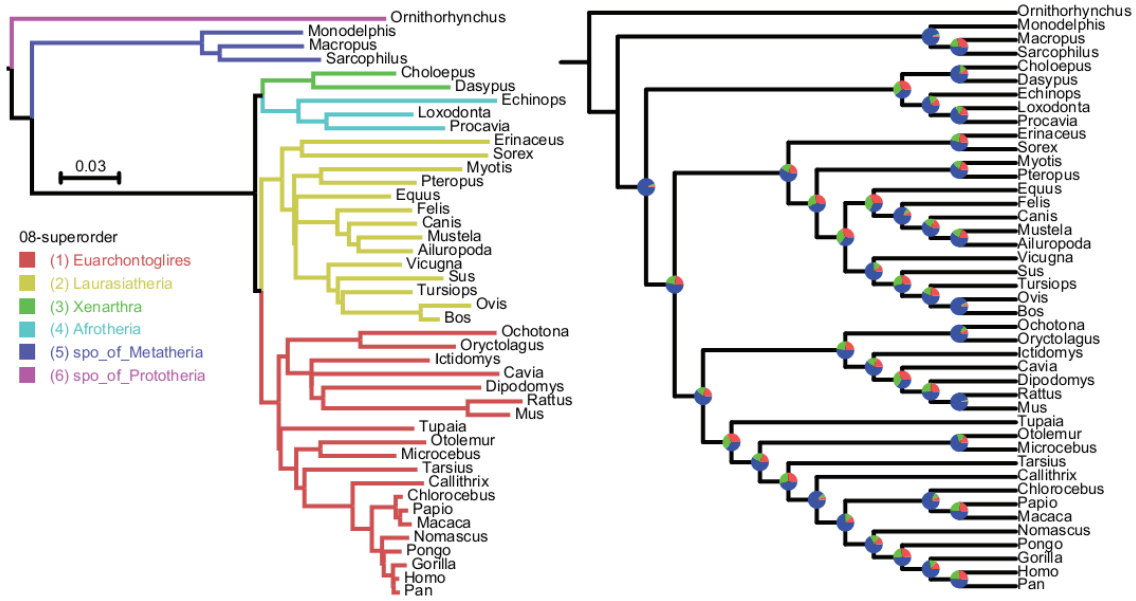


Figura 18: Superárvore gerada pelo HyperTriplet utilizando árvores de mamíferos geradas a partir de alinhamento de proteínas.

À direita, os ramos da árvore foram coloridas de acordo com as superordens presentes na árvore. À esquerda, a árvore foi colocada em cladograma, e em cada nó interno tem representado os valores de suporte em um gráfico de pizza. A porção em azul representa o suporte para a topologia atual. As outras duas porções (verde e vermelha) representam o suporte para a topologia alternativa.

Quanto à distribuição dos valores de suporte dos nós em relação ao tamanho dos ramos, verifica-se um padrão semelhante apresentado pela superárvore gerado a

partir dos dados de nucleotídeos, onde ramos mais curtos apresentam valores menores de suporte em relação àqueles ramos mais longos (Figura 19).

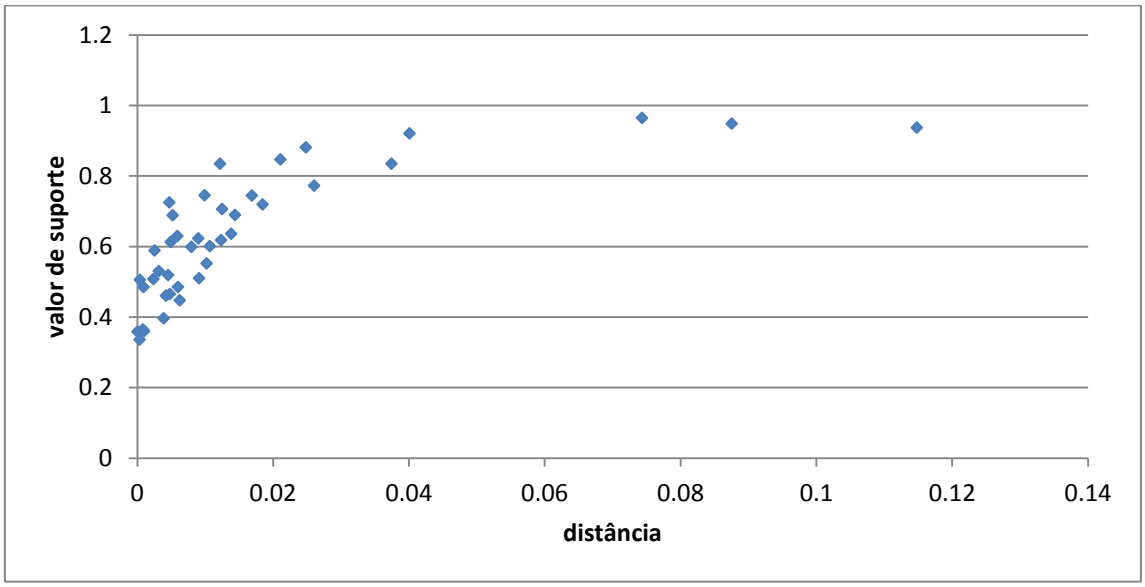


Figura 19: Relação entre a distância e o suporte do ramo da superárvore da Figura 18.

4.2.2.2. Dados de Fungo

Árvores filogenéticas construídas a partir de sequências de proteína de fungo foram obtidas no banco de dados PhylomeDB (HUERTA-CEPAS *et al.*, 2008) com o nome de acesso "Phylome Yeast (P60)". Estas árvores foram reconstruídas utilizando o software PhyML (GUINDON *et al.*, 2010) a partir do alinhamento de sequências proteicas gerado pelo programa MUSCLE (EDGAR, 2004). As árvores são constituídas de até 62 organismos, sendo 60 de fungos. Os outros dois organismos são organismos externos aos fungos e são *Homo sapiens* e *Arabidopsis thaliana*. Diferente das árvores armazenadas no banco de dados do OrthoMaM, as árvores do PhylomeDB não se encontram enraizadas, portanto, antes de submetê-las ao programa HyperTriplets, todas as árvores foram processadas pelo algoritmo de enraizamento baseado nas informações taxonômicas implementado no TaxOnTree. Além disso, algumas das árvores depositadas no banco de dados PhylomeDB possuem parálogos, o que possibilitou verificar o comportamento do programa HyperTriplets em lidar com árvores contendo duplicações de genes.

A superárvore gerada pelo programa HyperTriplets a partir das árvores de fungo enraizada com o auxílio de dados taxonômicos é apresentado na Figura 20 e

Figura 21 Evidenciando os grupos taxonômicos de categoria subfilo, utilizando o programa TaxOnTree, é possível verificar que a topologia gerada pelo programa HyperTriplets agrupa cada amostra presente na árvore em seus respectivos grupos taxonômicos (Figura 20). Quanto aos valores de suporte de cada nó interno (Figura 21), a média dos valores obtidos nesta árvore foi de 0,77, tendo 28 dos 60 nós internos com valores acima de 0,80 e oito nós com valores inferiores a 0,50. Ao relacionarmos a distância dos ramos com o seu valor de suporte, observa-se também nesta árvore um padrão semelhante ao que ocorre com a superávore gerada com os dados de mamíferos (Figura 22). Dentre os nós com baixo valor de suporte podem ser observado em relação ao posicionamento da espécie *Ustilago maydis* no clado do filo Basidiomycota, à história evolutiva da família Debaryomycetaceae e Saccharomycetaceae dentro do subfilo Saccharomycotina, e à história evolutiva dos membros da família Aspergillaceae dentro do subfilo Pezizomycotina. Outro fato notável que pode ser observado nesta árvore é a parafilia do grupo Fungi. O homem (*Homo sapiens*) encontra-se posicionado no clado que contém todas as amostras de fungo, enquanto a espécie de fungo *Encephalitozoon cuniculi* se posiciona como a espécie mais externa ao restante dos fungos. Apesar do valor de suporte apresentada por esta topologia não ser alta (0,5827), a topologia alternativa, que reposiciona o homem e o *Encephalitozoon cuniculi* de forma que grupo Fungi forme um clado monofilético é suportado apenas por 29,31% das árvores de genes.

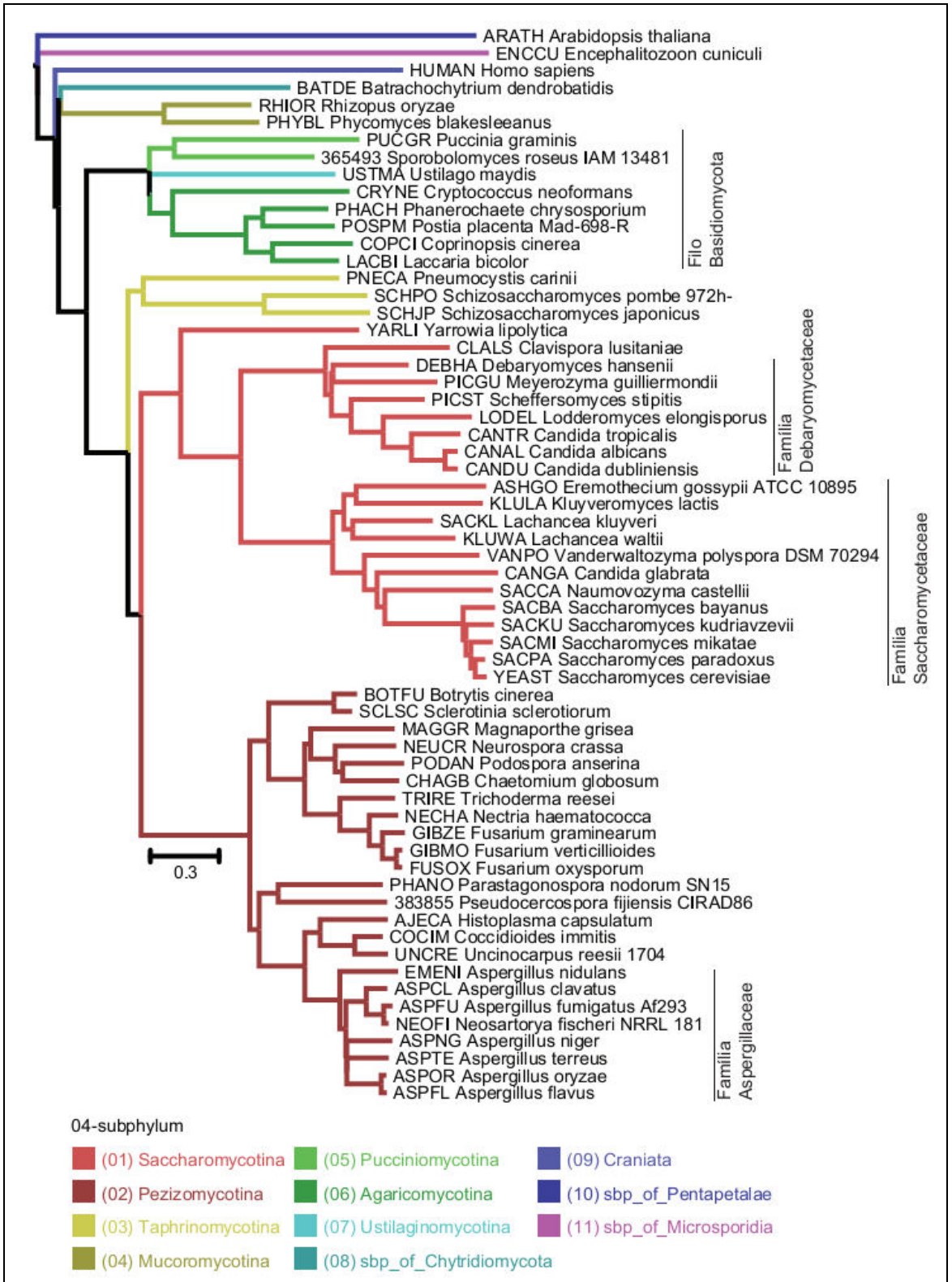


Figura 20: Superárvore gerada pelo programa HyperTriplet utilizando dados de fungo.

Os ramos encontram-se coloridos de acordo com os subfilos presentes na árvore. A coloração dos ramos da árvore foi realizada automaticamente pelo programa

TaxOnTree.

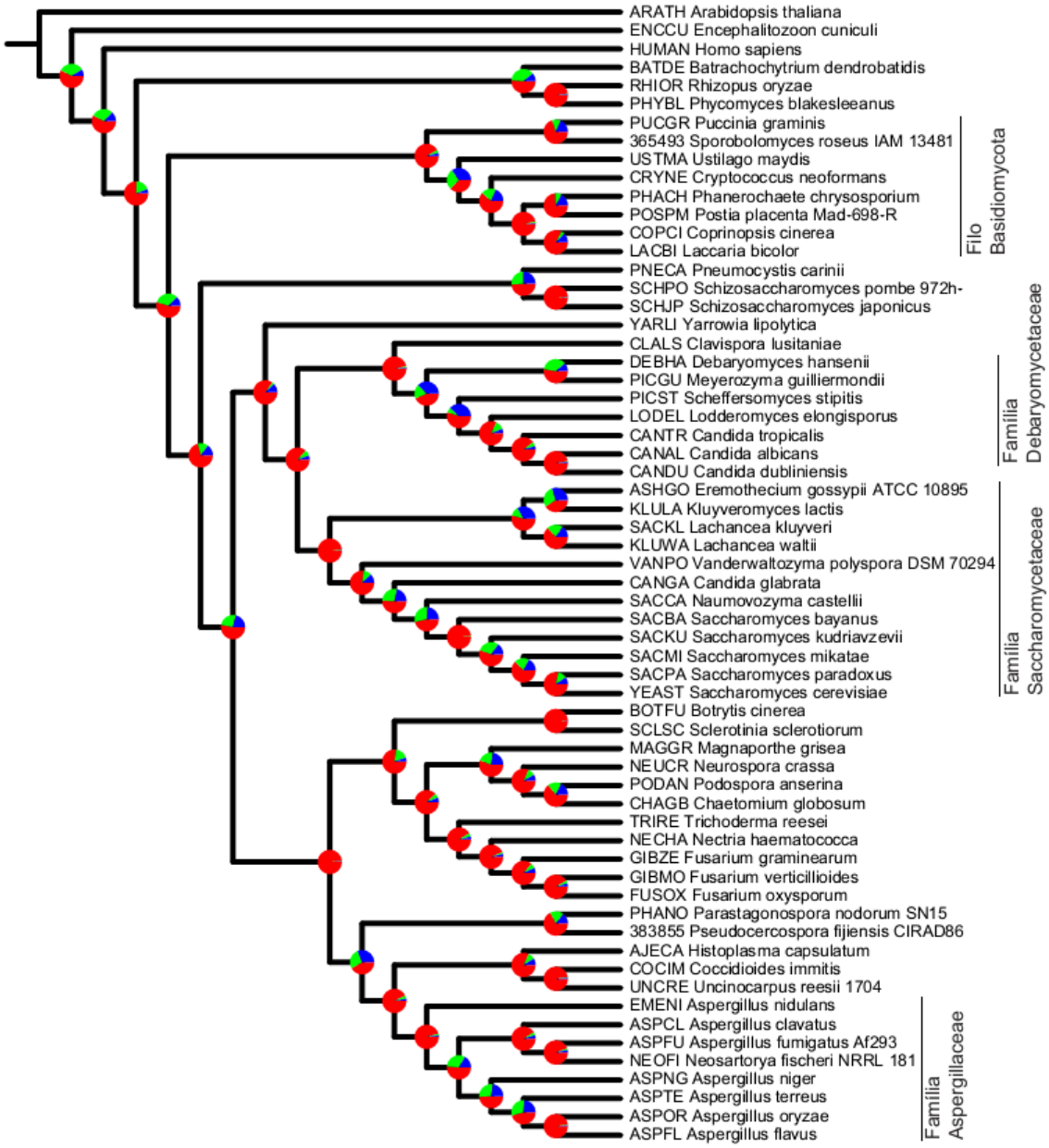


Figura 21: a árvore da Figura 20 representada em cladograma e com os valores de suporte em cada nó interno na forma de um gráfico de pizza.

A porção em vermelho representa o suporte da topologia atual. As outras duas porções (verde e azul) representam o suporte para as topologias alternativas.

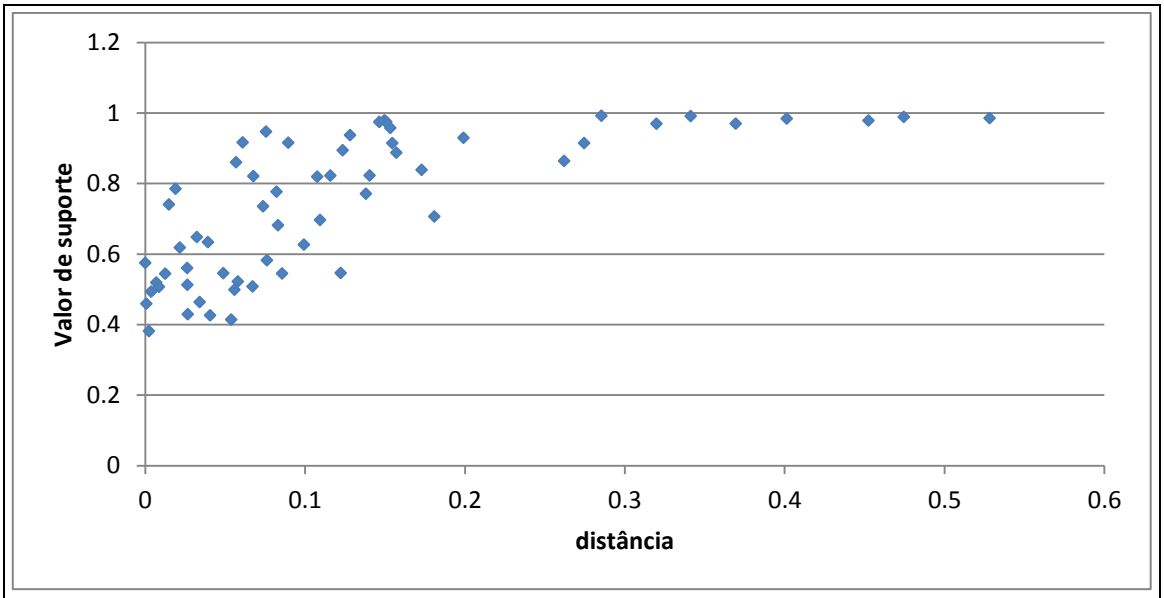


Figura 22: Relação entre a distância e o suporte do ramo da superárvore da Figura 20.

4.3. ELDOgraph

A principal forma de visualização da evolução e da relação entre as espécies é por meio de árvores filogenéticas. Quando esta se encontra na forma enraizada e dicotômica é possível inferir a ordem dos eventos evolutivos que ocorreram ao longo da história evolutiva dos organismos em análise por meio da sua topologia. As árvores filogenéticas em geral também apresentam informações que refletem a taxa de mutação ou a similaridade das sequências amostradas, que são obtidos extraindo-se os valores do comprimento dos ramos. Apesar das informações sobre a taxa de mutação ou sobre a similaridade de sequências serem de grande valia para os estudos evolutivos, um maior destaque é dado para as interpretações topológicas das árvores principalmente quando o estudo lida com uma grande quantidade de árvores (ex. na construção de superárvores). Neste contexto, foi desenvolvida neste trabalho a ferramenta ELDOgraph, que apresenta uma forma alternativa de ilustrar as relações entre as amostras de uma árvore filogenética através dos valores de distâncias entre as amostras. Destaca-se também a introdução de um novo conceito que foi denominado neste trabalho de ELDO (External Least Divergent Ortholog).

4.3.1. Conceito de ELDO

Considerando uma árvore filogenética, onde a distância do ramo reflete a taxa de substituição ou a similaridade entre as amostras, se considerarmos uma amostra A na árvore é possível ordenar as amostras restantes de acordo com a distância que estas amostras possuem em relação à amostra A. Se a distância entre a amostra A e a amostra B apresenta a menor distância em relação às demais, podemos destacar que a amostra B é o ortólogo menos divergente (LDO – Least Divergent Ortholog) da amostra A. No entanto, a amostra B pertence ao mesmo gênero que a amostra A, e pode ser do interesse do pesquisador não destacar o LDO que seja do mesmo gênero. Então podemos recorrer à lista ordenada das distâncias entre a amostra A e o restante das amostras e verificar quais das amostras não é do mesmo gênero e que apresenta a menor distância com a amostra A. Se a amostra C atende a esses requisitos, então esta amostra é considerada o ortólogo externo menos divergente (ELDO – External Least Divergent Ortholog) da amostra A em relação à categoria taxonômica "gênero". A princípio, na implementação da TaxOnTree, quando destacamos alguma categoria taxonômica na árvore, os grupos taxonômicos são ordenados de acordo com a distância filogenética da árvore. Neste contexto, o primeiro grupo destacado na legenda sempre será o grupo que contém o organismo associado à proteína "query". Já o segundo grupo taxonômico pode ser considerado como o grupo que contém o organismo associado ao ELDO da proteína "query".

Em uma árvore que ilustra a evolução de gene com taxa de substituição constante, o ELDO de uma amostra será sempre amostras que se encontram no grupo irmão. No entanto, nos casos reais de árvores filogenéticas, os diferentes ramos frequentemente apresentam taxas de substituição que variam entre eles, o que permite que o ELDO de uma determinada amostra não seja necessariamente aquela presente no grupo irmão. Um exemplo de uma árvore gerada pela TaxOnTree utilizando o acesso 239735506 (Tetherin, *Macaca mulatta*) do RefSeq ilustra esta situação (Figura 23). Quando destacamos os gêneros presentes na árvore, observa-se que a amostra que possui a menor distância com a proteína "query" e que não é do gênero *Macaca* é uma amostra do gênero *Papio*, que se encontra destacado como o segundo grupo da legenda. Este primeiro exemplo ilustra a situação onde o ELDO não ocorre nas amostras do grupo irmão, já que o grupo irmão do gênero *Macaca* nesta árvore é constituído pelas amostras do gênero *Cercocebus* e *Mandrillus*. A análise do ELDO da mesma proteína e

na mesma árvore filogenética, mas na categoria taxonômica "família" nos remete a outra característica de ELDO. O ELDO da proteína "query" neste caso é uma amostra de um organismo pertencente à família Hylobatidae, que nesta árvore encontra-se no grupo irmão do ramo que reúne as amostras da família Cercopithecidae. No entanto, outra família de primatas encontra-se no mesmo ramo que a amostra da família Hylobatidae, que é a família Hominidae. Apesar desta família também fazer parte do grupo irmão de Cercopithecidae, ela não se destaca como ELDO da proteína "query" devido a sua maior distância.

Na maioria dos casos, para cada amostra existe apenas uma amostra que seja a mais próxima filogeneticamente. No entanto, não se descarta a possibilidade de ocorrência de empates durante a procura da amostra com o menor valor de distância. Caso para uma proteína exista várias amostras com o valor da menor distância, o ELDO dessa proteína será todas estas amostras que apresentam a menor distância. Adicionalmente, existem casos em que a diferença entre o valor da menor distância e da segunda menor distância é pequena, e que seria biologicamente plausível considerar estes também como casos de empate. Para lidar com esta situação, foi criado um parâmetro ("threshold") que o usuário define uma margem de distância para que o algoritmo decida os casos de empates Esta margem corresponde a uma fração, determinada pelo usuário, da distância de um ELDO. Por exemplo, se a fração foi de 0,05 (5%), o algoritmo considerará também como ELDO as amostras de um grupo taxonômico distinto que tiver uma distância com a amostra em análise menor que a distância do ELDO somado com a fração de 0,05 dessa distância.

O conceito de ELDO foi derivado da inspeção manual de várias árvores produzidas pela TaxOnTree, já que ela denota, por exemplo quando a legenda é ajustada para gênero, qual o gênero distinto que possui sequencia menos distante da "query". A análise de ambiguidades na resolução de alguns ramos das superárvores produzidas pelo HyperTriplets sugere que uma fração dos genes de um organismo podem apresentar seus ELDO em um organismo, e outras frações, em outros organismos. Isso poderia ser representado na forma de uma rede, na estrutura de grafos.

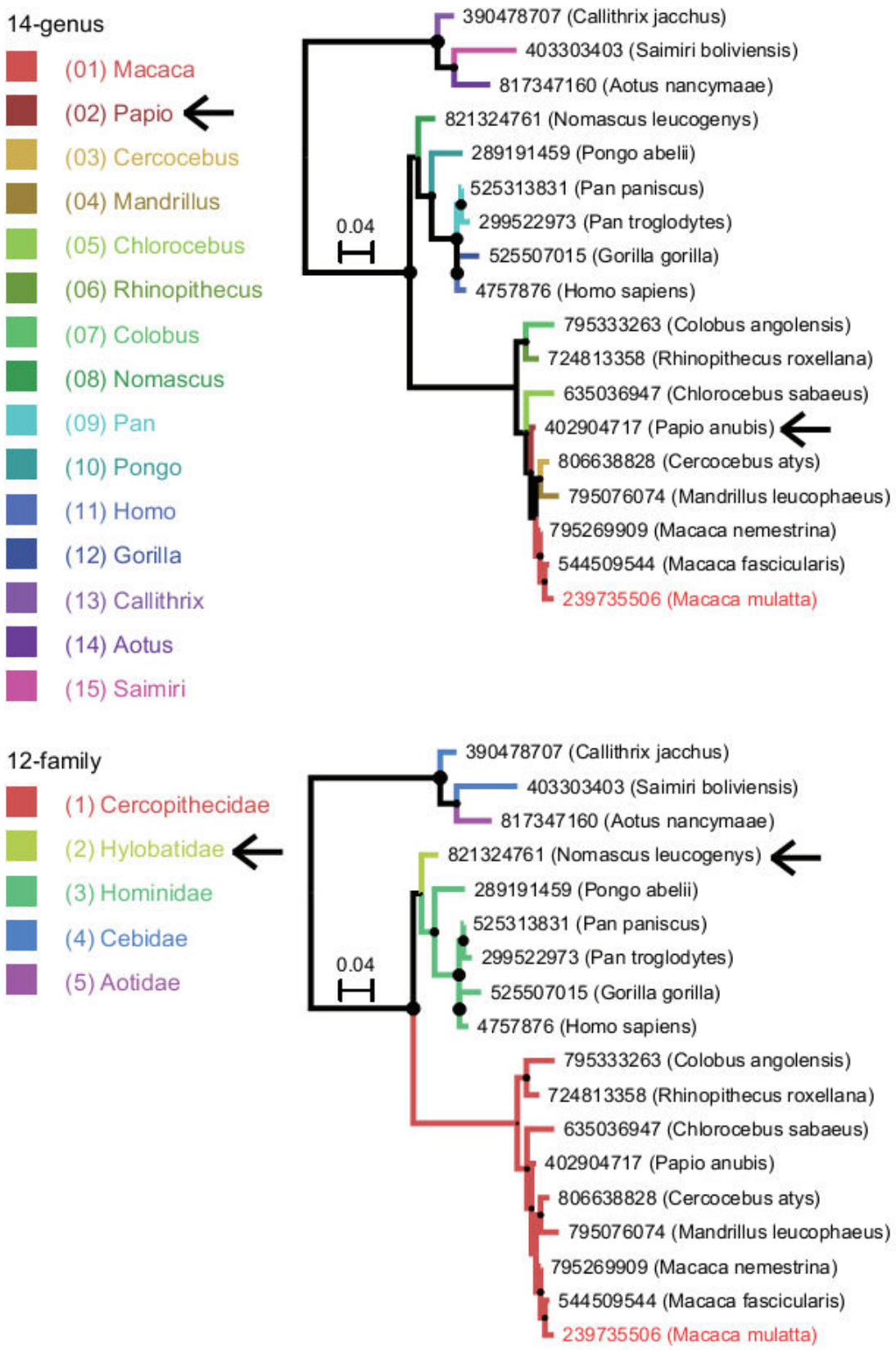


Figura 23: O ELDO pode ser determinado utilizando a TaxOnTree. O ELDO da proteína *query*, em vermelho, é o segundo táxon na legenda da TaxOnTree (setas). As árvores evidenciam as diferentes os diferentes gêneros (superior) e família (inferior).

4.3.2. Análise de ELDO em múltiplas árvores filogenéticas e o ELDOgraph

A análise de ELDO pode ser estendida para várias árvores filogenéticas (todos os genes do organismo, por exemplo) e dessas análises é possível verificar quantas amostras de um determinado grupo taxonômico possui como ELDO amostras de um segundo grupo. A Tabela 1 reúne o resultado da análise de ELDO em múltiplas árvores de genes de mamíferos considerando a categoria taxonômica "superordem". A tabela, por sua vez, pode ser utilizada para gerar redes ou grafos que facilitem a sua visualização como mostrado na Figura 24. Estas funcionalidades foram reunidas em um único programa que foi denominado como ELDOgraph.

Tabela 1: Frequência de ELDO entre os taxa de mamíferos na categoria taxonômica "superordem"

query/ELDO	Afrotheria	Euarchontoglires	Laurasiatheria	Xenarthra	spo_of_Metatheria	spo_of_Prototheria
Afrotheria	-	0.507	0.680	0.431	0.001	0.000
Euarchontoglires	0.109	-	0.951	0.149	0.000	0.000
Laurasiatheria	0.156	0.924	-	0.195	0.001	0.000
Xenarthra	0.361	0.488	0.675	-	0.001	0.000
spo_of_Metatheria	0.366	0.593	0.627	0.407	-	0.147
spo_of_Prototheria	0.246	0.389	0.417	0.259	0.663	-

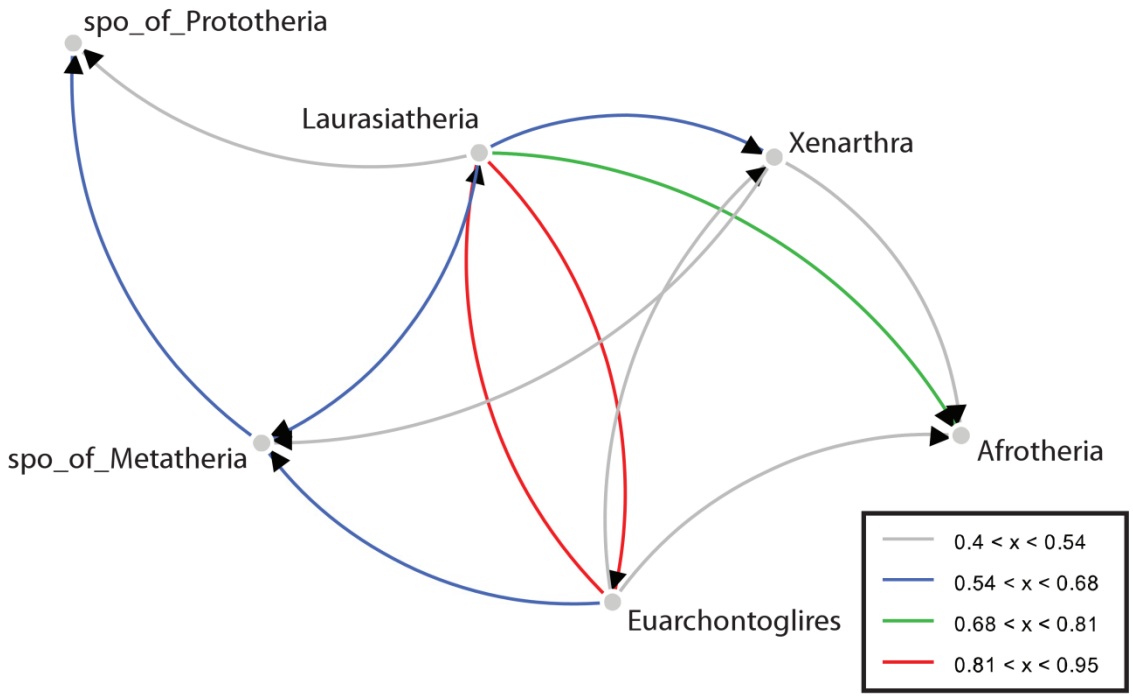


Figura 24: Representação em grafos dos dados da Tabela 1.

O ELDOgraph realiza dois principais procedimentos para realizar a análise global de ELDOs. No primeiro passo, o ELDOgraph construirá uma estrutura de dados onde estarão armazenados os dados de distância entre as amostras e as informações taxonômicas das amostras presentes nas árvores. Para isto, o ELDOgraph recebe como entrada múltiplas árvores filogenéticas e uma tabela contendo o identificador taxonômico para cada amostra presente na árvore. Esta tabela é constituída por duas colunas separadas por tabulação que possui o nome da amostra presente nas árvores e o identificador taxonômico do NCBI Taxonomy respectivamente na primeira e na segunda coluna. Após a construção dessa estrutura de dados, o ELDOgraph realizará a montagem da tabela de frequência de ELDOs considerando os parâmetros recebidos pelo usuário, como a categoria taxonômica a ser analisada e a margem de corte para os empates entre ELDOs. Por fim, o ELDOgraph montará uma página em HTML que utiliza recursos da biblioteca D3.js (BOSTOCK; OGIEVETSKY; HEER, 2011, p. 3) para a montagem dos grafos a partir da tabela de frequência de ELDOs gerada (Figura 25). Neste grafo, cada nó representa um grupo taxonômico presente nas árvores. Cada nó pode receber uma aresta de outros nós, e isto ocorre quando estes representam grupos taxonômicos que possuem ELDOs do grupo em análise. A frequência de ELDOs em um grupo taxonômico é refletida pelas cores das arestas, onde o vermelho denota

uma alta frequência, seguido pelo verde, azul e pelo cinza, que denota uma baixa frequência. A página em HTML gerado pelo ELDOgraph oferece ainda ao usuário alguns recursos listados abaixo (Figura 26) que podem ser utilizados para melhorar a visualização do grafo:

- *Threshold* - determina a frequência mínima para que uma aresta seja representada no grafo;
- *Link distance* - determina o comprimento das arestas;
- *Charge* - determina a força com que os nós são repelidos entre eles.

O usuário ainda conta com recursos que permitem que a visualização seja focada em um nó, que pode ser útil ao analisar grafos que possuam um grande número de arestas. Para isto, basta clicar com botão esquerdo do mouse sobre um nó de interesse e segurar o clique (Figura 27). Por fim, a página permite ao usuário salvar o grafo apresentado no formato SVG.

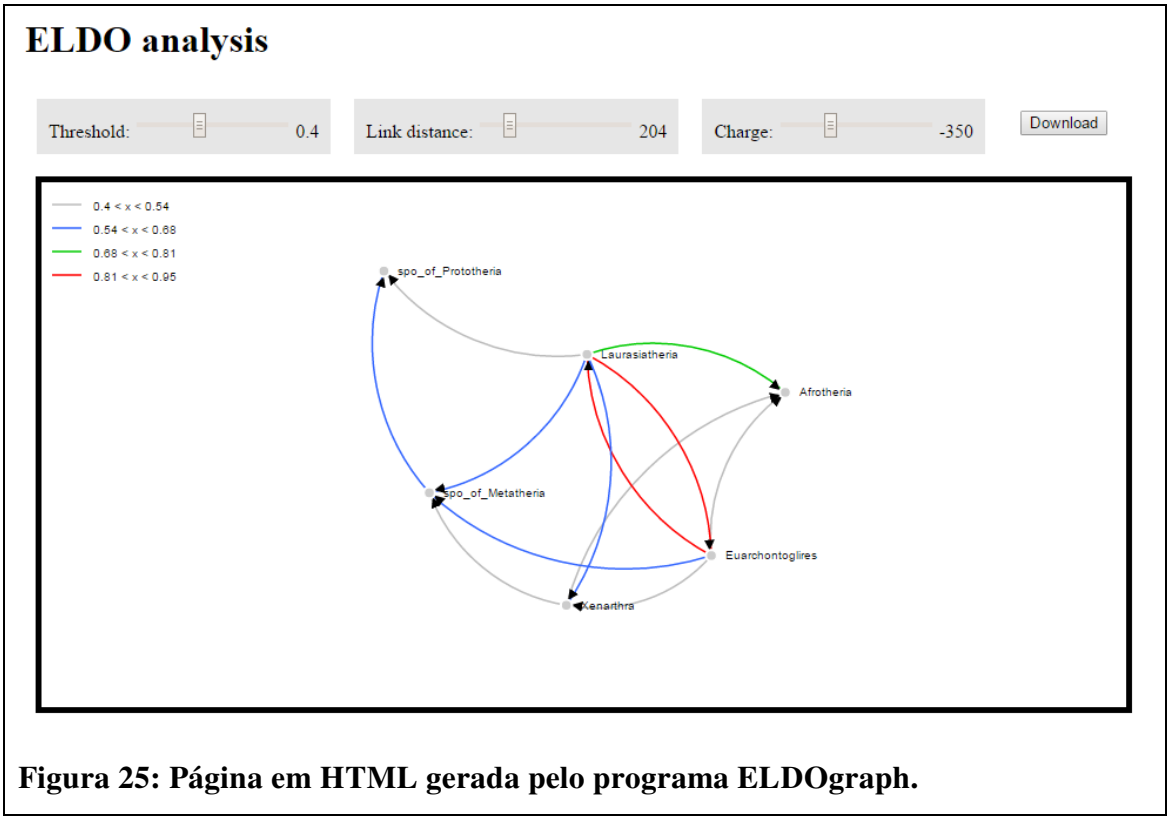


Figura 25: Página em HTML gerada pelo programa ELDOgraph.

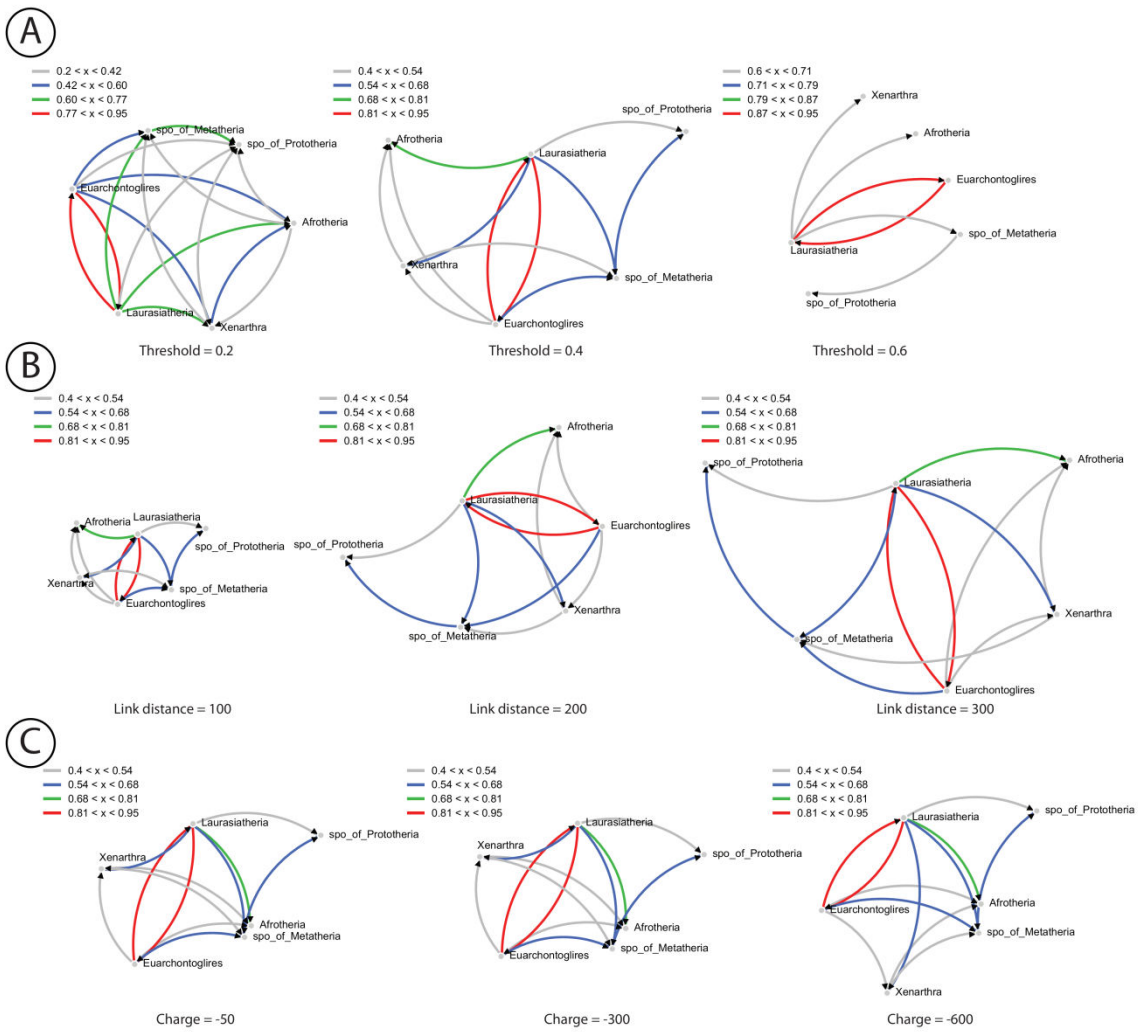


Figura 26: Efeito de cada parâmetro do ELDOgraph no desenho do grafo.

A figura ilustra o desenho dos grafos ao variar o *threshold* (A), *link distance* (B) e *charge* (C). A variação do desenho do grafo alterando o parâmetro *charge* pode ser percebida ao verificar a distância entre os nós Afrotheria e spo_of_Metatheria.

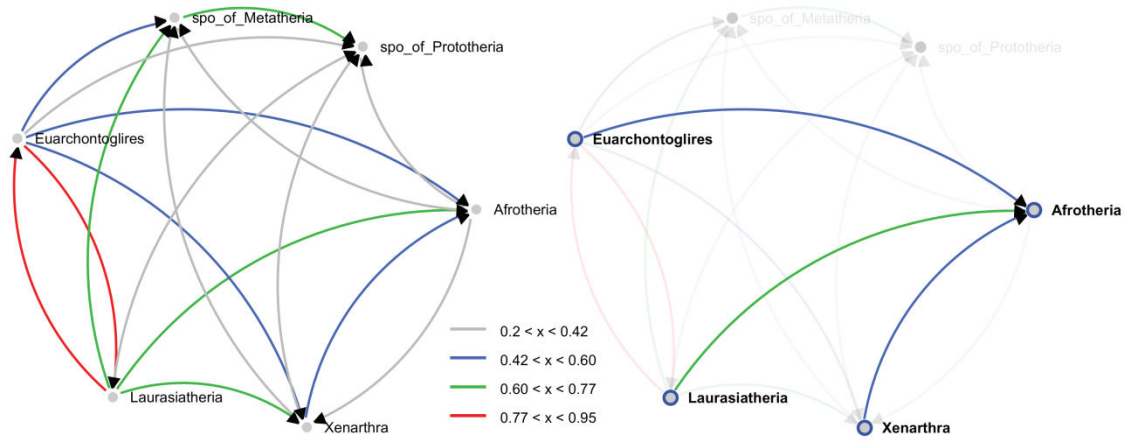


Figura 27: Demonstração do recurso foco do ELDOgraph.
Ao clicar e segurar um nó específico, o programa evidencia os nós e as arestas que apontam para ele.

4.3.3. Estudo de caso

O programa ELDOgraph foi aplicado em um conjunto de árvores disponíveis em bases de dados público como o OrthoMaM (DOUZERY *et al.*, 2014) e o PhylomeDB (HUERTA-CEPAS *et al.*, 2008). A seguir, serão apresentados alguns resultados obtidos a partir da análise de ELDO nestas árvores.

4.3.3.1. Dados de mamíferos

Árvores filogenéticas constituídas por gêneros de mamíferos que possuem o seu genoma sequenciado foram obtidas na base de dados do OrthoMaM (DOUZERY *et al.*, 2014) e submetidas ao programa ELDOgraph. O grafo desta análise, ilustrado na Figura 28, foi gerado considerando a categoria taxonômica "gênero" e utilizando uma fração de 0,05 da distância mínima para os eventuais empates de ELDO. Considerando também uma árvore de espécie constituída pelos organismos presentes na análise, é possível observar que os maiores valores de frequências de ELDO ocorrem entre dois organismos que se posicionam como taxa irmãos. Exemplo disso são os gêneros *Mus* e *Rattus*, que apresentam uma frequência de ELDO entre eles acima de 0,9. Outros pares de gêneros considerados irmãos e que apresentam uma alta frequência de ELDO entre eles são *Bos*/*Ovis*, *Homo*/*Pan*, *Loxodonta*/*Procavia*, *Ochtona*/*Oryctolagus*, *Ailuropoda*/*Mustela* e *Macaca*/*Papio*.

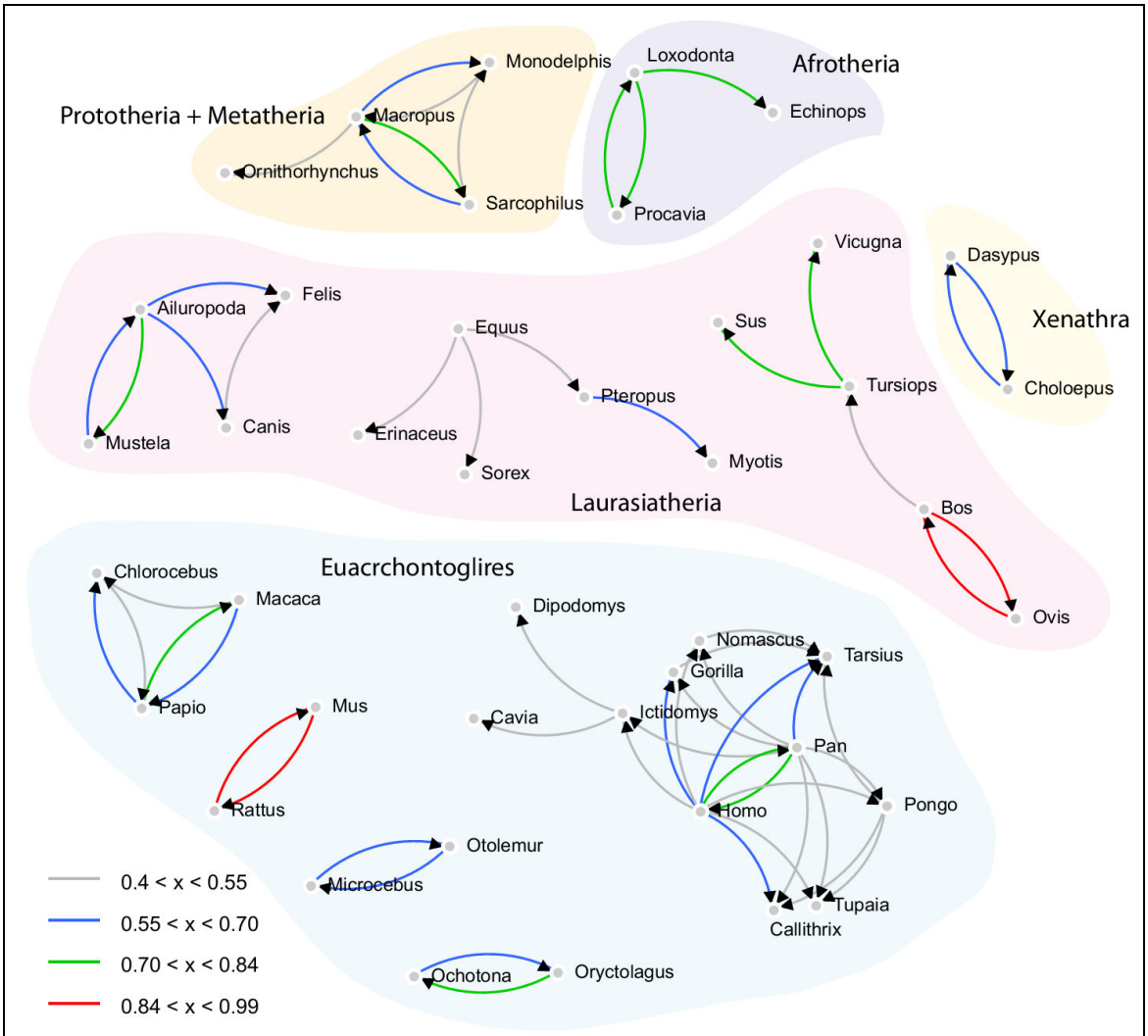


Figura 28: Grafo gerado pelo programa ELDOgraph utilizando dados de mamífero na categoria taxonômica "gênero".

As superordens evidenciadas no grafo foram posteriormente adicionadas à figura.

Algumas exceções a essa abordagem podem ser verificadas neste grafo. O gênero *Myotis* (ordem: Chiroptera) possui cerca de 60% de seus genes que tem os genes do gênero *Pteropus* (ordem: Chiroptera) como ELDO (Figura 29A e Tabela 2). Seria de esperar que a análise recíproca (ELDO dos genes do gênero *Pteropus*) apresentasse uma frequência semelhante, no entanto, o grafo demostra que menos de 40% dos genes de *Pteropus* possui como ELDO os genes de *Myotis*. Ao invés disso, uma boa parte dos ELDOs de *Pteropus* pertence ao gênero *Equus* (43,6%), que pertence à ordem Perissodactyla. Uma situação ainda mais drástica acontece nos dois gêneros

classificados na ordem Insectivora. A maior parte dos genes dos gêneros *Erinaceus* (50,6%) e *Sorex* (48,5%) possuem os genes de *Equus* como ELDOs e menos de 10% dos genes dos dois gêneros possuem ELDOs entre eles (Figura 29A e Tabela 2).

A subrede formada pelos gêneros da ordem Cetartiodactyla (Figura 29B) apresenta quatro arestas indicando alta frequência de ELDOs. Duas delas, como já foram citadas, são ELDOs entre os gêneros *Bos* e *Ovis*, que apresentam alta frequências pelo fato da divergência desses dois organismos ter ocorrido recentemente na evolução. Já as outras duas arestas indicam que mais de 70% dos ELDOs nos gêneros *Sus* (porco) e *Vicugna* (alpaca) pertencem ao gênero *Tursiops* (golfinho), um gênero que possui características fenotípicas bem distintas dentro da ordem Cetartiodactyla.

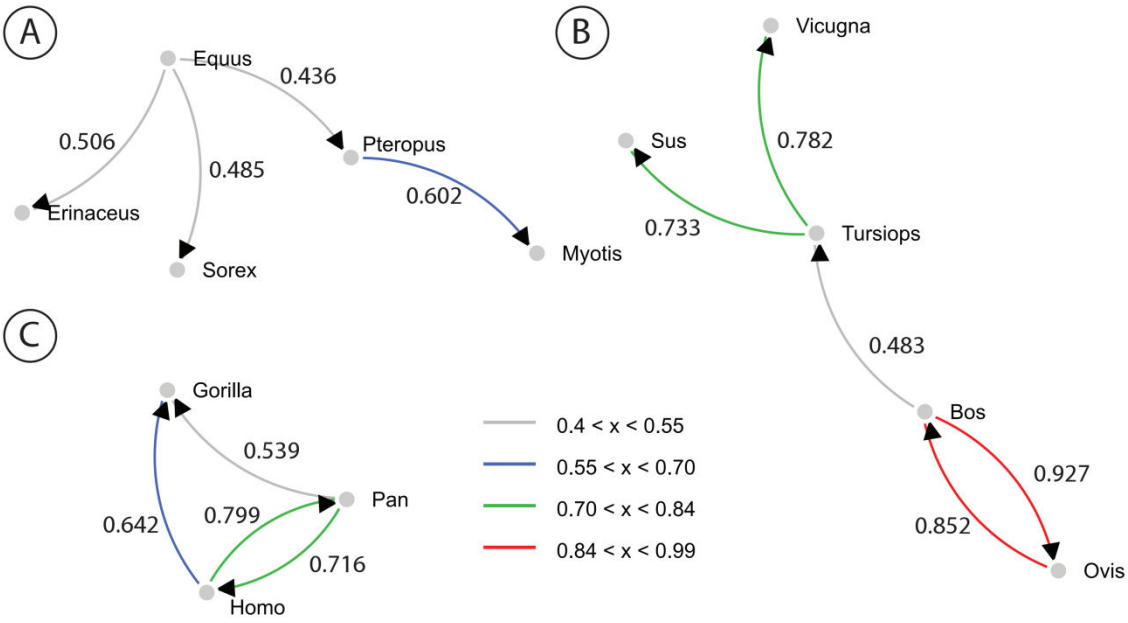


Figura 29: Subgrafos retirados do grafo da Figura 28.
Os números sobre a aresta representam o valor da frequência de ELDOs. Os subgrafos compreendem (A) o gênero Equus e os gêneros da ordem Insectivora e Chiroptera, (B) os gêneros da ordem Cetartiodactyla e (C) gêneros da subfamília Homininae.

Analisando a relação entre os gêneros da subfamília Homininae (*Homo*, *Pan* e *Gorilla*) (Figura 29C), observa-se que a maior parte dos genes dos gêneros *Homo* e *Pan* (mais de 70%) compartilham os ELDOs entre eles. Além disso, observa-se que, dentro da perspectiva do gênero *Gorilla*, a maior parte de seus genes possuem como ELDO genes dos gêneros de *Homo* e *Pan*. No entanto, os ELDOs pertencentes ao

gênero *Homo* prevalecem em quantidade, demonstrando que a maior parte dos genes de *Gorilla* é mais semelhante aos genes de *Homo* do que de *Pan*. Esta análise pode ser mais aprofundada verificando quais ELDOs são exclusivos de um organismo ou se existem casos de empates. Classificando os genes de acordo com os gêneros de seus ELDOs e organizando estes dados em um diagrama de Venn (Figura 30), observa-se que 89,9%, 89,1% e 76,1% dos genes de *Homo*, *Pan* e *Gorilla*, respectivamente, possuem ELDOs exclusivos para um gênero. A maior parte dos casos de empate ocorre entre gêneros da subfamília Homininae. Pelo diagrama de Venn é possível observar também que os outros quase 30% dos genes dos gêneros *Homo* e *Pan* possuem como ELDO os genes de *Gorilla*. Dentro da perspectiva do gênero *Gorilla*, a maior parte dos empates de ELDO é observado entre os gêneros *Homo* e *Pan*, que foi observado em 2976 genes.

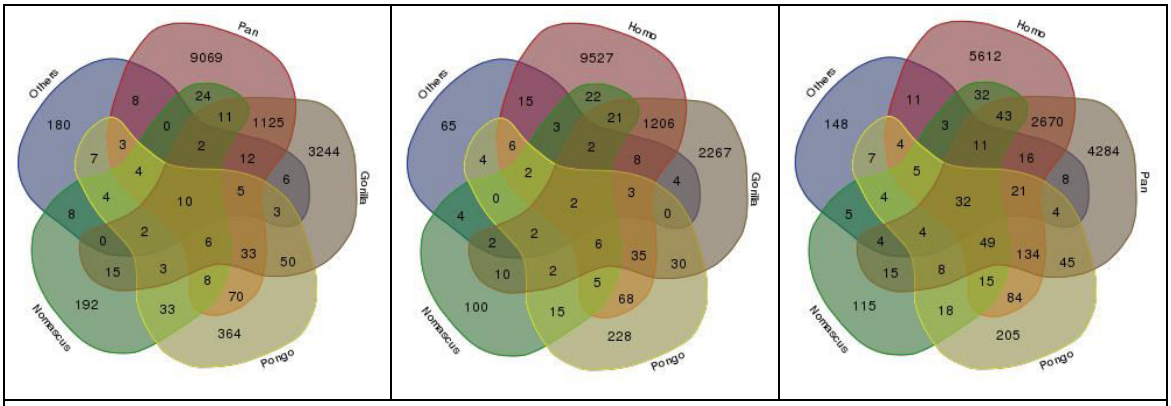


Figura 30: Diagrama de Venn que compara os organismos pertencentes aos ELDOs dos gêneros *Homo* (esquerda), *Pan* (meio) e *Gorilla* (direita).
As interseções no diagrama representam os casos de empate. A categoria "Others" engloba outros gêneros não representados no diagrama.

A Figura 31 ilustra a análise de ELDO utilizando as mesmas árvores filogenéticas de mamíferos e a mesma margem de corte, mas considerando a categoria taxonômica "ordem". Nesta análise, é possível analisar também que vários pares de ordens que são irmãos na árvore de espécie compartilham o ELDO entre eles, como pode ser observado nos pares Cetaceae/Cetartiodactyla, Pilosa/Cingulata, Diprotodontia/Dasyuromorphia e Proboscidea/Hyracoidea. No entanto, outros dados interessantes podem ser retirados desta análise. Analisando a ordem dos Primatas, observa-se que a maior parte de seus genes possuem como ELDO genes da ordem

Perissodactyla e não um gene de outras ordens que fazem parte da mesma superordem Euarchonthoglires (Rodentia, Scadentia e Lagomorpha). Verifica-se também que a maior parte dos genes da ordem Rodentia, Scadentia e Lagomorpha possui como ELDO genes de Primatas. Na superordem Laurasiatheria, verifica-se que os genes da ordem Perissodactyla são ELDOs de várias ordens dentro da superordem como Chiroptera, Carnivora, Insectivora. Os dados de frequência de ELDO entre pares de ordens estão reunidos na Tabela 3.

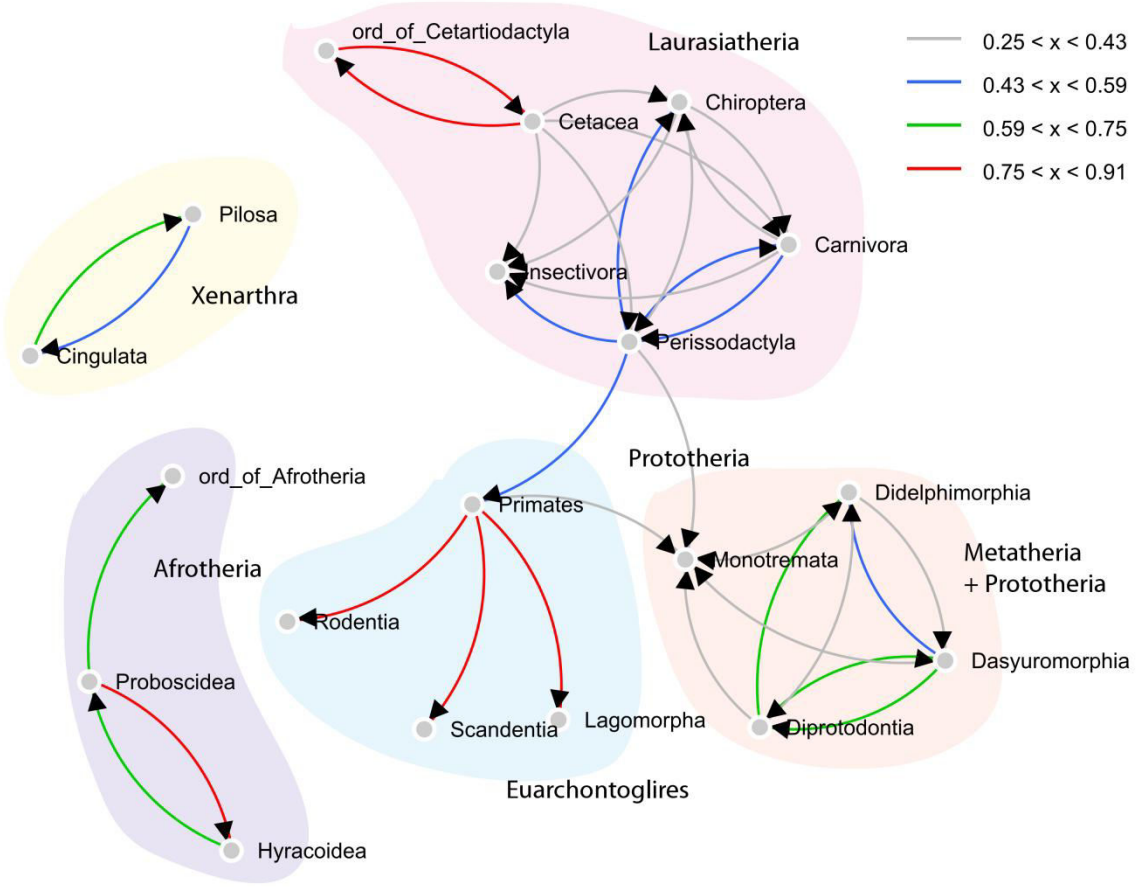


Figura 31: Grafo gerado pelo programa ELDOgraph utilizando dados de mamífero na categoria taxonômica "ordem".

As superordens evidenciadas no grafo foram posteriormente adicionadas à figura.

Tabela 2: Frequência de ELDO entre os taxa de mamíferos na categoria taxonômica "gênero".

query/ELDO		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43
Ailuropoda	1	-	0.0	0.0	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Bos	2	0.0	-	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0
Callithrix	3	0.0	0.0	-	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.6	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.0	0.5	0.3	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Canis	4	0.7	0.0	0.0	-	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Cavia	5	0.0	0.0	0.1	0.0	-	0.2	0.0	0.0	0.1	0.0	0.1	0.0	0.0	0.3	0.4	0.5	0.0	0.2	0.0	0.2	0.0	0.0	0.0	0.0	0.3	0.0	0.0	0.1	0.1	0.0	0.3	0.2	0.3	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.0
Chlorocebus	6	0.0	0.0	0.0	0.0	0.0	-	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Choloepus	7	0.0	0.0	0.0	0.0	0.0	0.0	-	0.7	0.0	0.0	0.1	0.0	0.0	0.1	0.1	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0
Dasypus	8	0.0	0.0	0.0	0.0	0.0	0.1	0.6	-	0.0	0.0	0.2	0.0	0.1	0.1	0.1	0.0	0.1	0.1	0.0	0.1	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.0
Dipodomys	9	0.0	0.0	0.1	0.0	0.1	0.2	0.0	0.0	-	0.0	0.1	0.0	0.0	0.3	0.4	0.5	0.0	0.2	0.0	0.2	0.0	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.0	0.3	0.2	0.3	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.0
Echinops	10	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	-	0.1	0.0	0.0	0.0	0.1	0.0	0.7	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Equus	11	0.2	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	-	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.1	0.0	0.0	0.4	0.1
Erinaceus	12	0.1	0.0	0.0	0.1	0.0	0.1	0.0	0.0	0.0	0.0	0.5	-	0.2	0.1	0.1	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.0	0.3	0.0	0.0	0.0	0.1	0.1	0.0	0.0	0.3	0.1
Felis	13	0.6	0.0	0.0	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Gorilla	14	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Homo	15	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	-	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Ictidomys	16	0.0	0.0	0.1	0.0	0.1	0.3	0.0	0.0	0.1	0.0	0.2	0.0	0.0	0.4	0.5	-	0.0	0.3	0.0	0.3	0.0	0.0	0.0	0.0	0.3	0.0	0.0	0.1	0.1	0.0	0.4	0.3	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.0
Loxodonta	17	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.0	0.1	0.0	0.0	0.1	0.1	0.0	-	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.1	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Macaca	18	0.0	0.0	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Macropus	19	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-	0.0	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.6	0.0	0.0	0.0	0.0	0.0	0.0
Microcebus	20	0.0	0.0	0.1	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.3	0.0	0.0	0.2	0.0	-	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.6	0.0	0.3	0.2	0.3	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0
Monodelphis	21	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7	0.0	-	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.0	0.0	0.0	0.0	0.0	0.0
Mus	22	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Mustela	23	0.8	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Myotis	24	0.1	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.6	0.0	0.0	0.0	0.0	0.0	0.1	0.0
Nomascus	25	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-	0.0	0.0	0.0	0.0	0.0	0.4	0.0	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Ochotona	26	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.0	0.1	0.0	0.1	0.0	0.0	0.0	0.0	0.1	-	0.0	0.8	0.0	0.0	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Ornithorhynchus	27	0.1	0.1	0.1	0.1	0.0	0.2	0.2	0.2	0.0	0.0	0.3	0.0	0.2	0.2	0.3	0.1	0.2	0.2	0.4	0.2	0.4	0.0	0.1	0.1	0.2	0.0	-	0.1	0.1	0.1	0.2	0.2	0.2	0.1	0.2	0.0	0.3	0.0	0.1	0.1	0.1	0.2	0.1
Oryctolagus	28	0.0	0.0	0.1	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.2	0.3	0.1	0.0	0.1	0.0	0.2	0.0	0.0	0.0	0.0	0.2	0.6	0.0	-	0.0	0.0	0.2	0.2	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Otolemur	29	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.2	0.0	0.0	0.1	0.0	0.7	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	-	0.0	0.2	0.1	0.2	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0
Ovis	30	0.0	0.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0
Pan	31	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Papio	32	0.0	0.0	0.0	0.0	0.0	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Pongo	33	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.4	0.0	-	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Procavia	34	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Pteropus	35	0.1	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-	0.0	0.0	0.0	0.0	0.0	0.2	0.1
Rattus	36	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-	0.0	0.0	0.0	0.0	0.0	0.0
Sarcophilus	37	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7	0.0	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-	0.0	0.0	0.0	0.0	0.0
Sorex	38	0.2	0.0	0.0	0.1	0.0	0.1	0.0	0.0	0.0	0.0	0.5	0.1	0.2	0.1	0.1	0.0	0.0	0.1	0.0	0.1	0.0	0.0	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.0	0.2	0.0	0.0	-	0.1	0.0	0.0	0.3	0.1
Sus	39	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-	0.0	0.0	0.7	0.2
Tarsius	40	0.0	0.0	0.1	0.0	0.0	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.6	0.0	0.0	0.3	0.0	0.3	0.0	0.0	0.0	0.0	0.4	0.0	0.0	0.0	0.1	0.0	0.6	0.3	0.5	0.0	0.0	0.0	0.0	0.0	0.0	-	0.0	0.0	0.0
Tupaia	41	0.0	0.0	0.1	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.4	0.5	0.1	0.0	0.3	0.0	0.4	0.0	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.1	0.0	0.5	0.3	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.1	-	0.1	0.0
Tursiops	42	0.0	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.0	0.0	-	0.2
Vicugna	43	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.8	-

Tabela 3: Frequência de ELDO entre os taxa de mamíferos na categoria taxonômica "ordem".

query/ELDO		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Carnivora	1	-	0.32	0.27	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.58	0.01	0.09	0.01	0.01	0.01	0.00	0.20
Cetacea	2	0.04	-	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.00	0.01	0.00	0.00	0.00	0.00	0.91
Chiroptera	3	0.35	0.32	-	0.01	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.56	0.01	0.08	0.01	0.01	0.01	0.00	0.18
Cingulata	4	0.12	0.09	0.08	-	0.00	0.00	0.00	0.02	0.01	0.01	0.00	0.18	0.55	0.21	0.14	0.02	0.02	0.01	0.06
Dasyuromorphia	5	0.01	0.01	0.01	0.01	-	0.33	0.72	0.00	0.00	0.00	0.00	0.01	0.00	0.02	0.01	0.01	0.00	0.00	0.01
Didelphimorphia	6	0.01	0.01	0.01	0.01	0.44	-	0.65	0.00	0.00	0.00	0.01	0.01	0.01	0.02	0.01	0.01	0.00	0.00	0.01
Diprotodontia	7	0.01	0.01	0.01	0.01	0.64	0.42	-	0.00	0.00	0.00	0.00	0.01	0.01	0.02	0.01	0.01	0.00	0.00	0.01
Hyracoidea	8	0.03	0.03	0.03	0.03	0.00	0.00	0.00	-	0.00	0.00	0.00	0.05	0.04	0.08	0.83	0.01	0.01	0.04	0.02
Insectivora	9	0.34	0.32	0.31	0.02	0.00	0.00	0.00	0.00	-	0.03	0.00	0.51	0.02	0.19	0.03	0.03	0.02	0.01	0.19
Lagomorpha	10	0.10	0.09	0.08	0.02	0.00	0.00	0.00	0.00	0.01	-	0.00	0.20	0.02	0.81	0.02	0.19	0.11	0.00	0.06
Monotremata	11	0.24	0.21	0.22	0.18	0.31	0.39	0.43	0.07	0.04	0.06	-	0.27	0.18	0.34	0.22	0.11	0.07	0.03	0.17
Perissodactyla	12	0.46	0.38	0.34	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	0.01	0.09	0.01	0.00	0.00	0.00	0.21
Pilosa	13	0.06	0.06	0.05	0.70	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.12	-	0.16	0.10	0.01	0.01	0.00	0.03
Primates	14	0.25	0.22	0.19	0.04	0.00	0.00	0.00	0.00	0.00	0.07	0.00	0.44	0.06	-	0.06	0.20	0.20	0.00	0.12
Proboscidea	15	0.07	0.05	0.04	0.07	0.00	0.00	0.00	0.70	0.00	0.00	0.00	0.11	0.08	0.15	-	0.01	0.01	0.05	0.03
Rodentia	16	0.11	0.09	0.08	0.02	0.00	0.00	0.00	0.00	0.01	0.09	0.00	0.21	0.02	0.87	0.02	-	0.10	0.00	0.06
Scandentia	17	0.08	0.07	0.06	0.01	0.00	0.00	0.00	0.00	0.01	0.04	0.00	0.17	0.02	0.85	0.02	0.08	-	0.00	0.04
ord_of_Afrotheria	18	0.05	0.05	0.04	0.05	0.00	0.00	0.00	0.18	0.01	0.01	0.00	0.08	0.06	0.13	0.71	0.01	0.01	-	0.03
ord_of_Cetartiodactyla	19	0.09	0.85	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.00	0.02	0.00	0.00	0.00	0.00	-

4.3.3.2. Dados de fungo

Uma análise de ELDO semelhante àquela realizada com dados de mamíferos também foi realizada utilizando dados de fungo. As árvores filogenéticas contendo genes de fungo foram obtidas do banco de dados do PhylomeDB (HUERTA-CEPAS *et al.*, 2008) e submetidas ao programa ELDOgraph. As árvores de máxima verossimilhança recuperadas desse banco de dados faz parte de uma coleção de árvores com o nome de acesso "Yeast phylome (P60)". As figuras Figura 32, Figura 33, Figura 34 e Figura 35 mostram grafos gerados pelo ELDOgraph, considerando a categoria taxonômica "espécie", "gênero", "família" e "ordem", respectivamente. Para todas as análises foram utilizadas um valor de margem de corte de 0,05 para considerar os casos de empates.

Todos os grafos demostram padrões semelhantes àqueles verificados nos grafos gerados utilizando os dados de mamíferos. A maior parte das arestas que representam maiores frequência de ELDO encontram-se entre taxa que estão posicionados como taxa irmãos na árvore de espécie.

Analisando os ELDOs da espécie *Saccharomyces cerevisae* ao longo das categorias taxonômicas, na categoria "espécie" (Figura 32), observa-se que a maior parte dos seus genes possuem como ELDO genes do seu táxon irmão *Saccharomyces paradoxus*. Os ELDOs das outras espécies de *Saccharomyces* na análise se distribuem de acordo com a ordem dos tempos de divergência entre estas espécies. A divergência mais recente provavelmente foi entre *S. cerevisae* e *S. paradoxus*, e este deve ter sido antecedido pela divergência de *S. mikatae*, já que a maior parte de seus ELDOs pertence àquelas duas primeiras espécies de *Saccharomyces*. A análise de ELDO de *S. mikatae* ainda demonstra que a frequência de seus ELDOs que pertencem à *S. paradoxus* (67.8%) é maior àquelas que pertencem à *S. cerevisae* (41.9%). Como a análise do ELDO está intimamente ligada à distância filogenética, essa diferença deve ser reflexo de uma taxa de mutação diferenciada encontrada entre essas duas espécies, que deve ser menor em *S. paradoxus*. Este mesmo fenômeno pode ser observado quando analisamos os ELDOs na espécie *S. kudriavzevii*, que, apesar de ter ELDO pertencente às espécies *S. cerevisae*, *S. paradoxus*, *S. mikatae* e *S. bayanus*, uma maior parte deles pertence à *S. paradoxus*.

Na categoria taxonômica "gênero" (Figura 33), a maior parte dos ELDOs do gênero *Saccharomyces* pertencem ao gênero *Naumovozyma*, seguido dos gêneros

Nakaseomyces e *Vanderwaltozyma*. É possível observar também que os ELDOs dos três gêneros citados anteriormente a maior parte deles pertencem a uma amostra do gênero *Saccharomyces*.

A família do gênero *Saccharomyces* é a Saccharomycetaceae e a maior parte de seus genes (mais de 80%) possui como ELDO genes que pertencem à família Debaryomycetaceae (Figura 34). A família Metschnikowiaceae, que é um grupo irmão da família Debaryomycetaceae, completa a maior parte dos 20% dos ELDOs restante de todos os Saccharomycetaceae em análise.

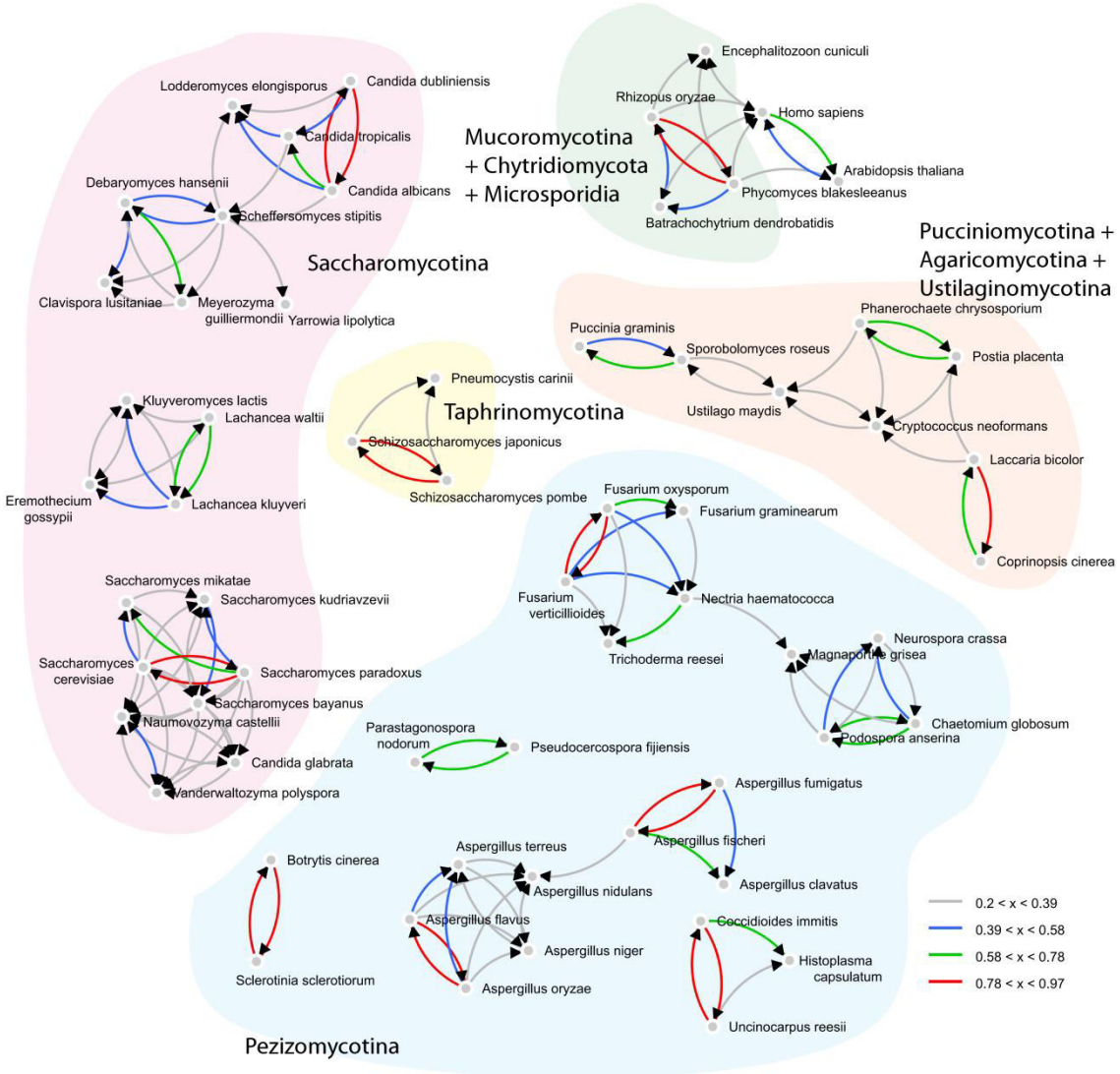


Figura 32: Grafo gerado pelo programa ELDOgraph utilizando dados de fungo na categoria taxonômica "espécie".

Os subfilos evidenciados no grafo foram posteriormente adicionados à figura.

Por fim, analisando os ELDOs em nível de ordem (Figura 35), observa-se que a única ordem dentro do subfilo Saccaromycotina é a ordem Saccaromycetales. Esta possui a maior parte dos ELDOs genes que pertencem a ordens do subfilo Pezizomycotina, como a Helotiales (*Botrytis* e *Sclerontinia*), Eurotiales (*Aspergillus*) e Onygenales (*Histoplasma*, *Coccidioides* e *Uncinocarpus*), e uma pequena porção dos ELDOs (aproximadamente 20%) que pertencem à ordem Schizosaccharomycetales, que pertence ao subfilo Taphrinomycotina.

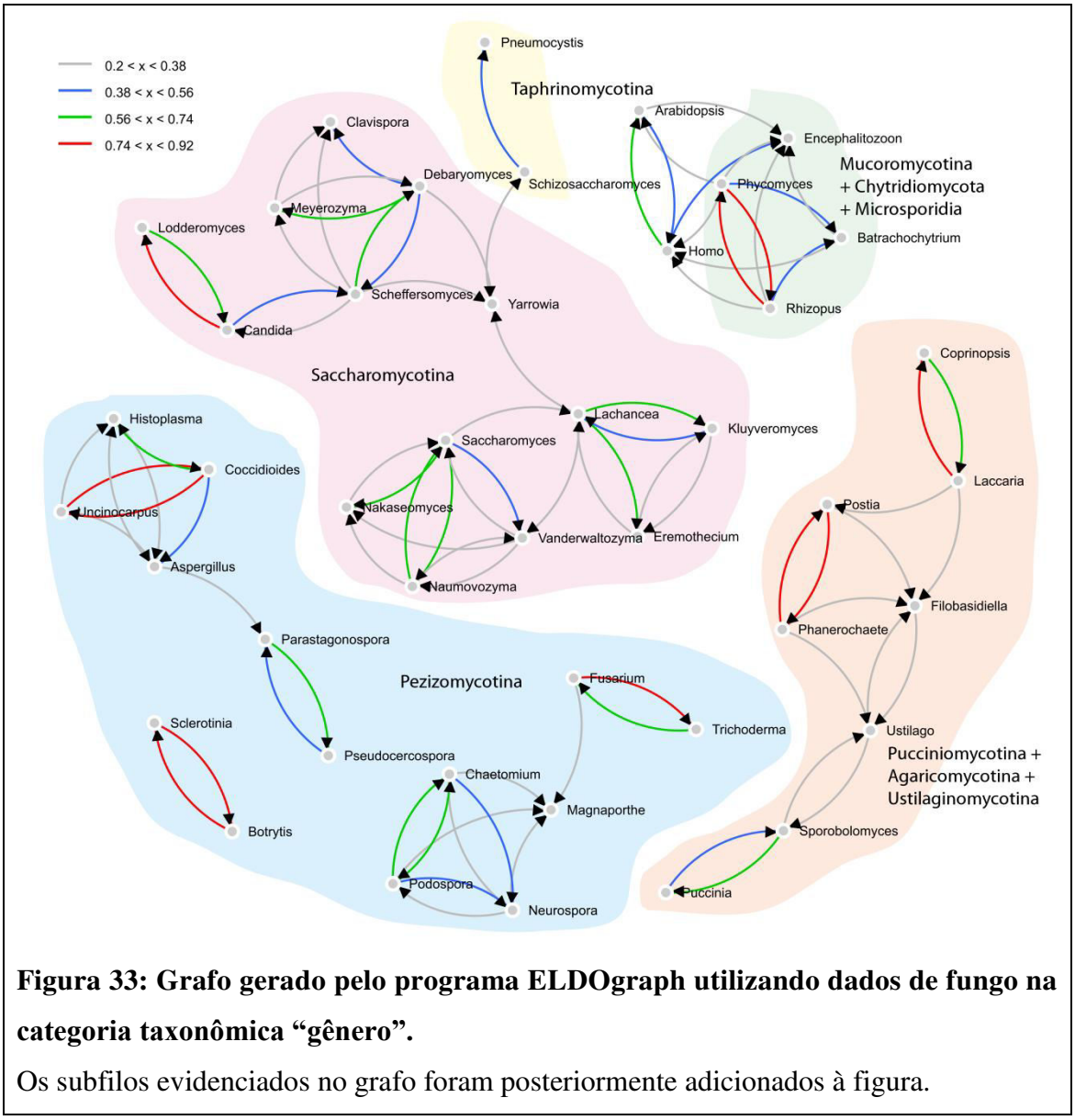


Figura 33: Grafo gerado pelo programa ELDOgraph utilizando dados de fungo na categoria taxonômica "gênero".

Os subfilos evidenciados no grafo foram posteriormente adicionados à figura.

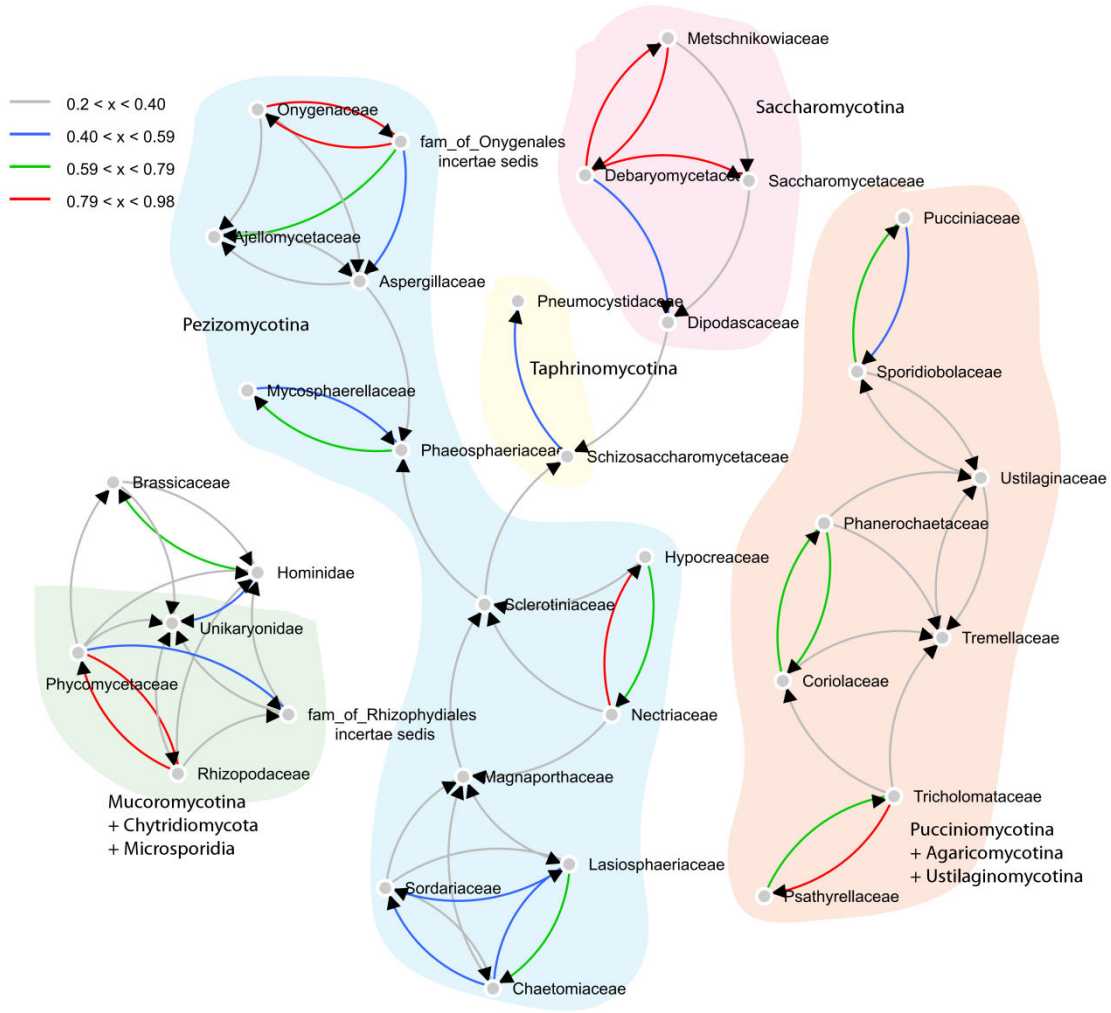


Figura 34: Grafo gerado pelo programa ELDOgraph utilizando dados de fungo na categoria taxonômica "família".

Os subfilos evidenciados no grafo foram posteriormente adicionados à figura.

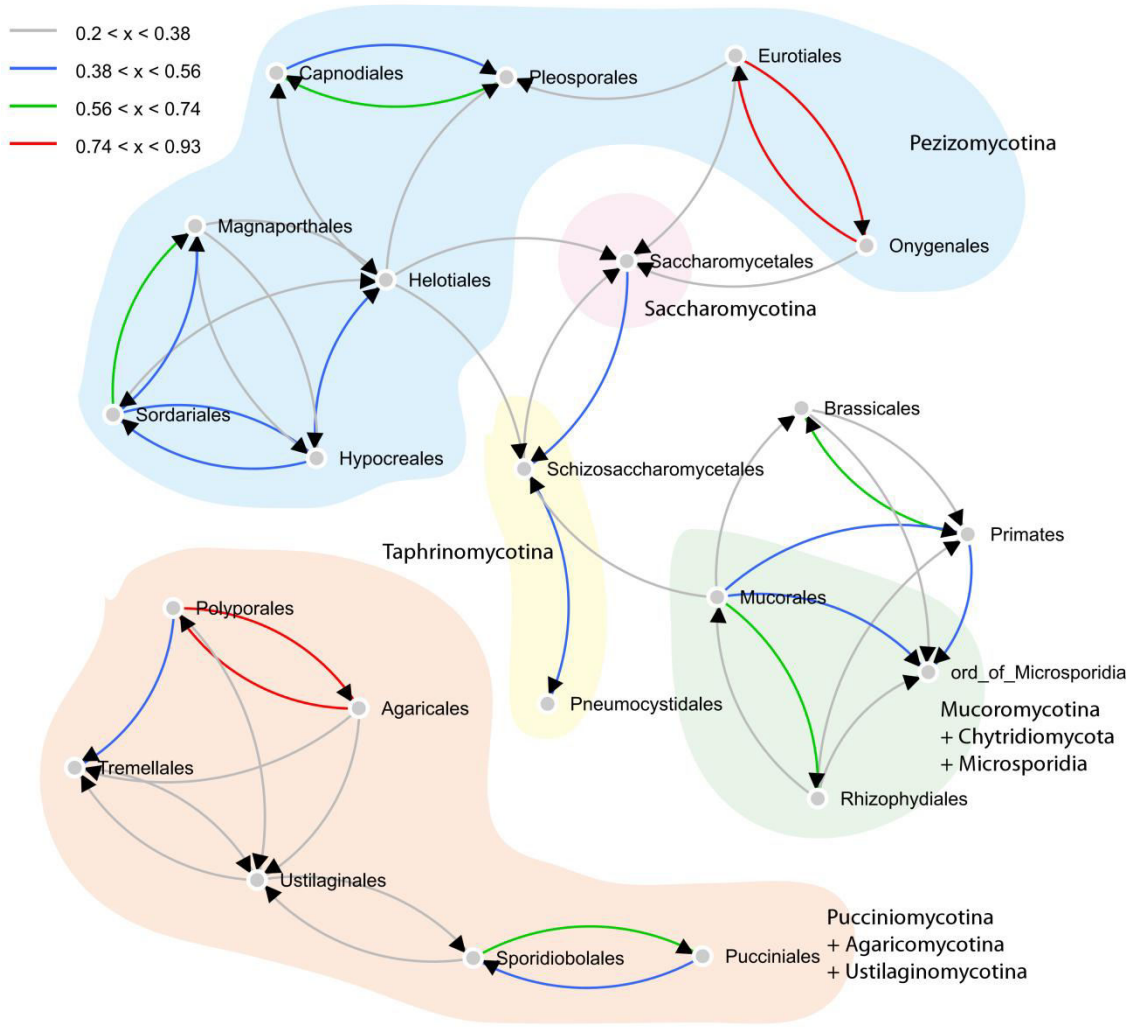


Figura 35: Grafo gerado pelo programa ELDOgraph utilizando dados de fungo na categoria taxonômica "ordem".

Os subfilos evidenciados no grafo foram posteriormente adicionados à figura.

A representação apresentada no ELDOgraph sugere que um organismo pode apresentar uma certa fração de seus genes mais próxima a genes de um organismo, enquanto outra fração é mais próxima a de outro organismo, e uma terceira fração ser igualmente próxima dos dois organismos preclaros. O ELDOgraph é uma solução computacional completa, pois a partir de um conjunto de árvores produz o "output" gráfico final, além de fornecer a tabela de distâncias que pode ser usada em pós-processamentos.

5. DISCUSSÃO

Neste trabalho foram exploradas as informações taxonômicas provenientes da base de dados Taxonomy do NCBI de forma programática, gerando três aplicações e novas conceituações, que serão discutidas em partes a seguir.

5.1. Anotação automática das informações taxonômicas na árvore filogenética

A anotação taxonômica na árvore filogenética pode ser realizada por alguns software disponíveis publicamente como o PhyloView (PALIDWOR; REYNAUD; ANDRADE-NAVARRO, 2006), iToL (LETUNIC; BORK, 2011) e ETE toolkit (HUERTA-CEPAS; DOPAZO; GABALDÓN, 2010). O programa PhyloView é uma ferramenta web que mais apresenta semelhanças com as funcionalidades da TaxOnTree. Assim como a TaxOnTree, o PhyloView possui como entrada uma árvore filogenética no formato Newick e recupera os dados de taxonomia das amostras presentes na árvore através de requisições aos servidores do NCBI e, posteriormente, colore os ramos de acordo com os grupos taxonômicos presentes na árvore. Apesar dessas semelhanças, a TaxOnTree provê algumas funcionalidades adicionais para os usuários, como por exemplo a edição automática dos nomes das amostras, enraizamento automático da árvore de acordo com as informações taxonômicas e também uma fácil visualização das relações evolutivas entre a proteína "query" e as outras amostras presentes na árvore. A abordagem centrada na "query" contempla o interesse que muitos pesquisadores têm pela filogenia induzido pelo interesse em uma dada sequência ou organismo. Além disso, pelo fato da árvore final da TaxOnTree estar acoplado às funções do programa FigTree, um programa de visualização de árvores comumente utilizado no meio científico, o programa provê ao usuário funções adicionais de edição da árvore e de exportação da árvore em formatos apropriados para a publicação, o que parece não ser possível pela forma de visualização da árvore implementada no PhyloView. iToL e ETE toolkit são duas aplicações bem documentadas que possuem também funções que permitem uma anotação automática da taxonomia na árvore. No entanto, a anotação taxonômica para ambos os programas requerem que o usuário proveja os identificadores taxonômicos de cada amostra. Como a recuperação dos identificadores taxonômicos não é uma tarefa trivial, o procedimento utilizado por estes dois programas pode requerer um esforço maior para conseguir resultados semelhantes oferecidas pela TaxOnTree. O

mesmo procedimento é requerido pela TaxOnTree apenas quando as amostras não apresentam em seus nomes um identificador de proteínas do NCBI ou do Uniprot. Uma característica que pode ser uma desvantagem da TaxOnTree em relação aos programas citados acima, é que a TaxOnTree não apresenta uma forma de manipular a árvore na web. Como uma alternativa, foi incluído no pacote um "script" que utiliza as funções de linha de comando do FigTree e gera uma figura da árvore filogenética no formato SVG. Este formato permite a visualização da árvore nos principais navegadores. Como exemplo de uso desta funcionalidade, nós criamos um site com árvores geradas automaticamente para explorar árvores de genes de morcegos.

De modo similar, software que leem as anotações presentes em uma árvore filogenética, como o ColorTree (CHEN, WEI-HUA; LERCHER, 2009) e o MixtureTree Annotator (CHEN, SHU-CHUAN; OGATA, 2015), também podem prover visualização que evidencie os grupos taxonômicos presentes na árvore. No entanto, como estas aplicações foram desenvolvidas para propósitos diversos, o usuário necessitará recuperar manualmente as informações taxonômicas de cada amostra da árvore e configurar o programa de forma que ele apresente a coloração dos ramos de forma apropriada. Como a TaxOnTree foi desenvolvida especificamente para lidar com as informações taxonômicas, estes procedimentos são realizados de forma automatizada.

A TaxOnTree oferece todo aparato para que as análises sejam realizadas em larga escala. A sua execução pode ser realizada a partir de linhas de comando em um terminal de um sistema operacional UNIX. Os bancos de dados de sequência e de dados taxonômicos utilizados pela TaxOnTree podem ser baixados e instalados localmente para que a TaxOnTree seja executada sem a necessidade de uma conexão na rede, e assim diminuir o tempo de execução. Assim, as funções implementadas na TaxOnTree podem ser aplicadas em um grande volume de árvores. A análise programática de árvores filogenética, que pode não ser uma tarefa trivial, pode ser evitada utilizando os relatórios taxonômicos gerados pela TaxOnTree. Estes relatórios são apresentados na forma de tabelas e podem auxiliar na análise em larga escala dos dados taxonômicos das árvores amostradas. O algoritmo de enraizamento baseado em dados taxonômicos pode também ser aplicado em larga escala para a construção de uma superárvore a partir de métodos que dependem de árvores enraizadas para a análise. A sua aplicação é demonstrada na análise realizada com o HyperTriplets com as árvores de fungo provenientes do banco de dados do PhylomeDB.

5.2. Inferindo árvores de espécie a partir de tripletos

A análise de múltiplos genes para a reconstrução da história evolutiva de um conjunto de espécies fornece um maior suporte nas inferências filogenéticas (DELSUC; BRINKMANN; PHILIPPE, 2005; PHILIPPE *et al.*, 2005). Se um grande número de genes infere uma história evolutiva semelhante, existe uma grande chance de que a história evolutiva das espécies tenha passado pelos mesmos caminhos. A abordagem da supermatriz, que consiste na concatenação dos dados de alinhamento de diferentes genes, tem sido o principal método utilizado para a inferência da árvore de genes. No entanto, algumas tendências sobre os dados utilizados para os estudos de filogenômica torna a prática desta abordagem limitada. Um dos requisitos para a reconstrução filogenética utilizando a supermatriz é que os genes incluídos na análise devem estar presentes como em cópia única em todas as amostras. Com o aumento do número de genomas sequenciados, o número de genes presentes exatamente em uma única cópia em todas as amostras diminui dramaticamente (SNEL; HUYNEN; DUTILH, 2005), o que restringe o seu uso apenas a um conjunto de amostras filogeneticamente próximas. Além disso, existem dúvidas de que as análises estatísticas do suporte dos ramos realizadas pelos programas comumente utilizados na reconstrução da árvore a partir de uma supermatriz (bootstrap e probabilidade posteriori) sejam adequadas para as análises filogenômicas (SALICHOS; ROKAS, 2013). Neste contexto, apesar de não estar imune a críticas (GATESY; SPRINGER, 2004; NOVACEK, 2001), o uso da abordagem da superárvore vem aumentando por melhor se adaptar ao contexto dos dados genômicos. A reconstrução de uma superárvore não requer, por exemplo, que os genes em análise estejam presentes em todas as espécies em análise, mas a maioria dos programas requer que os genes estejam em cópia única. Vários algoritmos e métodos foram elaborados para a construção da superárvore, mas o mais utilizado é o método da Representação de Matriz com Parcimônia (MRP, ou do inglês, "Matrix Representation with Parsimony") (BAUM, 1992; RAGAN, 1992).

O uso de tripletos para a reconstrução da superárvore oferece algumas vantagens em relação à manipulação dos dados e o tempo computacional necessário para a resolução do problema. Diferentes métodos foram sugeridos para reconciliar os tripletos presentes em um conjunto de árvores em uma única superárvore. A primeira proposta foi denominada T(I)LI (do inglês "Triplet (Inference and) Local Inconsistency") (MOSSES; PEDERSEN, 2005) que é uma adaptação do método Q(I)LI

(do inglês “Quartet (Inference and) Local Inconsistency”) (PIAGGIO-TALICE; BURLEIGH; EULENSTEIN, 2004), que por sua vez reconstrói superárvores a partir de quartetos. O método T(I)LI atribui pesos para os tripletos de acordo com a sua frequência encontrada nas árvores. Para a reconstrução da superárvore, o algoritmo escolhe um tripleto que possui o maior peso e as outras amostras são adicionadas iterativamente em uma posição da árvore que melhor se ajusta aos dados de frequência. Um segundo método utiliza na etapa da reconstrução da superárvore o algoritmo “Hill Climbing” para encontrar a árvore que melhor se ajusta aos dados (LIN; BURLEIGH; EULENSTEIN, 2009). Já o programa SuperTriplets (RANWEZ; CRISCUOLO; DOUZERY, 2010) inicia esta etapa construindo uma árvore estrelar semelhante àquele utilizado no algoritmo do Neighbor-Joining e realiza vários rearranjos da árvore baseados no “Nearest-Neighbor Interchange” (NNI) para encontrar o melhor arranjo das amostras. O algoritmo implementado no HyperTriplets para a determinação da árvore final assemelha-se àquele utilizado pelo SuperTriplets, diferindo na forma como o algoritmo inicia a primeira árvore e na forma como a melhor topologia é encontrada. Estas diferenças parecem não influir no desempenho do algoritmo e no resultado. Comparando os resultados com os dados do OrthoMaM, que também foi utilizado pelos autores do SuperTriplets, a topologia da superárvore resultante foi idêntica. No entanto, uma etapa que pode necessitar de uma revisão no algoritmo do HyperTriplets é a da decomposição das árvores em tripletos, uma vez que o SuperTriplets apresentou um desempenho bem superior nesta etapa (dados não mostrados). O baixo desempenho do HyperTriplets nesta etapa pode ser justificado pela extração dos dados de distância entre os pares de amostras na árvore e pela verificação da existência de nós de duplicações.

Apesar de, algoritmicamente, o HyperTriplets se assemelhar ao SuperTriplets, uma grande contribuição do HyperTriplets para os métodos das superárvores é que ele adiciona dados de distância aos ramos da superárvore e permite o uso de árvores com parálogos como entrada. Apesar da topologia da árvore ser do maior interesse para aqueles que inferem uma árvore de espécie, dados de distância podem fornecer informações interessantes, principalmente para aqueles pontos da árvore que apresentam baixo suporte. As distâncias dos ramos nas superárvores geradas pelo HyperTriplets são baseadas em uma média de distâncias entre as amostras encontradas nas árvores fornecidas pelo usuário. Através da sua inclusão, foi possível verificar neste trabalho que os nós de baixo suporte encontrados nas superárvores geradas a partir de dados de mamíferos e de fungos geralmente estão associados a eventos antigos de

rápida radiação, que são representados por ramos curtos na árvore (PHILIPPE *et al.*, 2011; WHITFIELD; KJER, 2008).

O HyperTriplets possui na sua implementação um algoritmo de particionamento das árvores que permite lidar com árvores que possuem parálogos. Uma árvore é particionada em duas ou mais árvores à medida que o algoritmo encontra um nó de duplicação, gerando, no final, árvores que não contêm nós de duplicação. Todas as árvores geradas por este algoritmo são utilizadas na contagem dos tripletos, mas as suas frequências são ponderadas, de forma que um "cluster" de genes que contenha muitos membros seja igualmente representado em relação aos "clusters" que possuam um membro de cada espécie. Com a inclusão de novas famílias gênicas na inferência da árvore de espécie, o presente procedimento representa mais uma iniciativa para a quebra da concepção de que as análises filogenômicas devem ser restritas a genes de cópia única (HELLMUTH *et al.*, 2015), já que as árvores que contém parálogos normalmente são descartadas tanto na análise de superárvores quanto de supermatrizes. Embora as aplicações TaxOnTree e HyperTriplets explorem árvores filogenéticas, estudos de caso feitos com elas chamaram a atenção para a possibilidade de sublimar-se a informação da história da filogenia para exploras as similaridades entre os proteomas de diversos organismos não necessariamente irmãos em uma árvore.

5.3. Um novo conceito para a bioinformática evolutiva: ELDO

O termo ELDO é derivado do termo LDO (do inglês "Least Diverged Ortholog") utilizado no banco de dados do PANTHER (MI; MURUGANUJAN; THOMAS, 2013). Em um "cluster" de ortólogos deste banco, um par de proteínas de espécies distintas é classificado como LDO se a distância filogenética entre elas representar a menor distância encontrada entre as proteínas dessas duas espécies. O uso desta terminologia se torna interessante na análise de grupos de ortólogos que possuem parálogos. Se um par de espécies possui uma relação "one-to-many" ou "many-to-many" entre suas proteínas, apenas aquele par de ortólogos que apresenta a menor distância filogenética é classificado como LDO. Casos em que o par de espécies possui ortólogos do tipo "one-to-one", o seu par de ortólogos sempre será classificado como LDO.

O conceito de ELDO utilizado neste trabalho difere daquele utilizado no banco PANTHER. Enquanto no PANTHER o LDO é utilizado como uma forma de classificar os pares de ortólogos de forma a encontrar os genes "equivalentes" entre

duas espécies (MI; MURUGANUJAN; THOMAS, 2013), o ELDO é uma propriedade que cada proteína amostrada possui e que representa o ortólogo mais próximo e externo àqueles ortólogos que sejam do mesmo táxon da proteína em análise. Assim, haverá o ELDO para o nível de gênero, família, ordem, etc. O conceito de ELDO, quando aplicado em grande número de genes, possibilita verificar alguns questionamentos da genômica comparativa como saber quais organismos possuem um conjunto de genes mais similar com os genes de um segundo organismo. Não é difícil perceber que estudos de caso feitos com a TaxOnTree inspiraram este conceito, dado que esta última aplicação enumera os demais grupos por ordem de proximidade com a "query".

Se considerarmos uma árvore cuja taxa de mutação é constante para todos os ramos ou uma árvore ultramétrica, onde a distância de qualquer folha até a raiz é a mesma, o ELDO dessas amostras será sempre representado por amostras presentes nos seus grupos irmãos. No entanto, esta mesma análise em árvores aditivas, onde o tamanho dos ramos pode refletir diferentes taxas de mutações, nem sempre a afirmativa anterior é verificada. Analisando a subfamília Homininae, foi possível verificar que a maior parte dos genes do homem (71,6%) possui uma menor distância filogenética com os genes do chimpanzé. Esta característica pode ser justificada pelo fato do homem e chimpanzé serem taxa irmãos nesta análise. No entanto, o fato de quase 30% de seus genes não terem o ELDO de chimpanzé remete à ocorrência de outros processos evolutivos que cabe ser analisada. Observando na perspectiva do gorila, apesar dele ser um táxon irmão ao clado formado pelo homem e pelo chimpanzé, boa parte dos seus genes (64,2%) possuem ELDOs de humanos (em comparação a 53,9% dos genes possuírem ELDOs de chimpanzé), implicando que mais genes dos gorilas são mais próximos, em termos de distância filogenética, com os genes do homem.

Os taxa *Myotis* e *Pteropus* (ordem: Chiroptera) apresentam um caso interessante de taxa irmãos onde um possui a maior parte dos ELDOs de um táxon irmão, mas o outro não. O *Pteropus* possui mais de 40% de seus genes com ELDOs de cavalo contra 30% de *Myotis*. Isto sugere que as taxas de mutações das linhagens de *Myotis* e de *Pteropus* se diferenciaram, tendo a linhagem de *Myotis* sido submetida a maiores taxas, em mais genes. Um caso mais extremo do não compartilhamento de ELDOs entre taxa irmãos ocorre entre o par *Erinaceus*/*Sorex*, onde menos de 10% de seus genes possuem ELDOs entre eles. A maior parte de seus ELDOs também é de genes do cavalo. Um caso intrigante também foi observado ao se verificar que a maior parte dos genes de *Sus* (porco) e *Vicugna* (alpaca) possuem como ELDOs genes de

Tursiops (golfinho). É de se pensar que a grande divergência nos aspectos morfológico, fisiológico e ecológico encontrada nos golfinhos entre os gêneros da ordem Cetartiodactyla tenha sido acompanhada também por uma grande divergência gênica. No entanto, resultados deste trabalho sugerem que apenas uma porção menor e especializada de genes de *Tursiops* acompanhou essa divergência.

A inclusão eletrônica dos dados taxonômicos e a manipulação programática da estrutura hierárquica da árvore taxonômica permitiram também que as comparações entre os conjuntos de genes dos organismos amostrados fossem realizadas considerando-se um determinado nível taxonômico. Assim, pôde-se verificar a similaridade filogenética entre classes, ordens ou famílias taxonômicas presentes na amostra. Em uma análise entre as ordens presentes nos dados de mamíferos, era de se esperar que a ordem dos primatas compartilhasse um maior número de ELDOs com alguma ordem que faz parte da mesma superordem que ele (Euarchontoglires), como a ordem Rodentia (ratos e camundongos), Lagomorpha (coelhos) ou Scadentia (escadêncios). No entanto, verificou-se que a maior parte dos genes de primatas possuem como ELDO genes de organismos pertencentes à ordem Perissodactyla (44%), que incluem os cavalos. Os gêneros da ordem de Rodentia amostrados são caracterizados por possuírem um tamanho de população efetiva grande (PHIFER-RIXEY *et al.*, 2012). Como a teoria da genética de população sugere que quanto maior o tamanho efetivo da população, maior também deve ser a taxa de evolução adaptativa (KIMURA, 1984), uma vez que maior número de indivíduos contribui para o surgimento de novas mutações, isto pode explicar a maior divergência gênica nesses grupos em relação a outros presentes na amostra. Os roedores são comumente utilizados como organismos modelo para o entendimento de diversos mecanismos e funções biológicas encontradas no homem (NGUYEN; XU, 2008; VANHOOREN; LIBERT, 2013). No entanto, a presença de poucos ELDOs de roedores nos genes de primatas pode ter implicações nas especulações realizadas a partir dos resultados obtidos desses organismos em humanos (SEOK *et al.*, 2013). Como perspectiva, podemos estudar as categorias funcionais enriquecidas nos ELDOs e entender, por exemplo, quais processos de humanos seriam melhor modelados em cavalos que em roedores.

6. CONCLUSÃO

Neste trabalho foram desenvolvidas aplicações no campo da filogenia molecular tendo em mente a necessidade deles serem capazes de lidar com um grande volume de dados moleculares. Entre as aplicações desenvolvidas incluem a TaxOnTree, que manipula programaticamente as árvores de genes e incluem múltiplos dados taxonômicos, o HyperTriplets, que propõe superárvores com suporte de arquitetura e distância de ramos, e o ELDOgraph, que realiza uma análise comparativa entre os genes utilizando os dados de distância presentes nas árvores. Todas as aplicações foram desenvolvidas no formato de programas, e uma delas, TaxOnTree, no formato de uma aplicação web. Os aplicativos criados neste trabalho confrontam com as dificuldades encontradas nos estudos filogenéticos como a manipulação e análise de um grande volume de árvores de genes, e a visualização e o acesso às informações (com foco em dados taxonômicos) das amostras de uma árvore filogenética. Desta forma, o presente trabalho contribui com algumas das principais demandas atuais observadas nos estudos filogenéticos.

7. REFERÊNCIAS BIBLIOGRÁFICAS

ALTENHOFF, Adrian M. *et al.* Inferring Hierarchical Orthologous Groups from Orthologous Gene Pairs. *PLOS ONE*, v. 8, n. 1, p. e53786, 14 jan. 2013.

ALTENHOFF, Adrian M. *et al.* The OMA orthology database in 2015: function predictions, better plant support, synteny view and other improvements. *Nucleic Acids Research*, PMID: 25399418PMCID: PMC4383958, v. 43, n. Database issue, p. D240–D249, 28 jan. 2015.

ALTSCHUL, S. F. *et al.* Basic local alignment search tool. *Journal of Molecular Biology*, PMID: 2231712, v. 215, n. 3, p. 403–410, 5 out. 1990.

BAUM, Bernard R. Combining Trees as a Way of Combining Data Sets for Phylogenetic Inference, and the Desirability of Combining Gene Trees. *Taxon*, v. 41, n. 1, p. 3–10, 1992.

BININDA-EMONDS, Olaf R. P. *et al.* The delayed rise of present-day mammals. *Nature*, v. 446, n. 7135, p. 507–512, 29 mar. 2007.

BOSTOCK, Michael; OGIEVETSKY, Vadim; HEER, Jeffrey. D3 data-driven documents. *IEEE transactions on visualization and computer graphics*, v. 17, n. 12, p. 2301–2309, 2011.

CAMACHO, Christiam *et al.* BLAST+: architecture and applications. *BMC Bioinformatics*, v. 10, p. 421, 2009.

CAPELLA-GUTIÉRREZ, Salvador; SILLA-MARTÍNEZ, José M.; GABALDÓN, Toni. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics (Oxford, England)*, PMID: 19505945, PMCID: PMC2712344, v. 25, n. 15, p. 1972–1973, 1 ago. 2009.

CHEN, Feng *et al.* OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Research*, PMID: 16381887, v. 34, n. suppl 1, p. D363–D368, 1 jan. 2006.

CHEN, Shu-Chuan; OGATA, Aaron. MixtureTree annotator: a program for automatic colorization and visual annotation of MixtureTree. *PloS One*, PMID: 25826378, PMCID: PMC4380466, v. 10, n. 3, p. e0118893, 2015.

CHEN, Wei-Hua; LERCHER, Martin J. ColorTree: a batch customization tool for phylogenic trees. *BMC research notes*, PMID: 19646243, PMCID: PMC2727521, v. 2, p. 155, 2009.

CONSORTIUM, The UniProt. UniProt: a hub for protein information. *Nucleic Acids Research*, PMID: 25348405, v. 43, n. D1, p. D204–D212, 28 jan. 2015.

COSTELLO, Mark J.; WILSON, Simon; HOULDING, Brett. Predicting Total Global Species Richness Using Rates of Species Description and Estimates of Taxonomic Effort. *Systematic Biology*, PMID: 21856630, v. 61, n. 5, p. 871–883, 1 out. 2012.

DARWIN, Charles; DE BEER, Sir Gavin. *The origin of species by means of natural selection: or, the preservation of favoured races in the struggle for life*. . [S.l.]: Oxford University Press, 1956.

DEGNAN, James H.; ROSENBERG, Noah A. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology & Evolution*, PMID: 19307040, v. 24, n. 6, p. 332–340, jun. 2009.

DELSUC, Frédéric; BRINKMANN, Henner; PHILIPPE, Hervé. Phylogenomics and the reconstruction of the tree of life. *Nature Reviews Genetics*, v. 6, n. 5, p. 361–375, maio 2005.

DOUZERY, Emmanuel J. P. *et al.* OrthoMaM v8: a database of orthologous exons and coding sequences for comparative genomics in mammals. *Molecular Biology and Evolution*, PMID: 24723423, v. 31, n. 7, p. 1923–1928, jul. 2014.

EDGAR, Robert C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, PMID: 15034147, v. 32, n. 5, p. 1792–1797, 1 mar. 2004.

GATESY, John; SPRINGER, Mark S. A Critique of Matrix Representation with Parsimony Supertrees. In: BININDA-EMONDS, OLAF R. P. (Org.). . *Phylogenetic Supertrees*. Computational Biology. [S.l.]: Springer Netherlands, 2004. p. 369–388.

GUINDON, Stéphane *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology*, PMID: 20525638, v. 59, n. 3, p. 307–321, maio 2010.

HARTFIELD, Matthew; MURALL, Carmen Lía; ALIZON, Samuel. Clinical applications of pathogen phylogenies. *Trends in Molecular Medicine*, PMID: 24794010, v. 20, n. 7, p. 394–404, 1 jul. 2014.

HELLMUTH, Marc *et al.* Phylogenomics with paralogs. *Proceedings of the National Academy of Sciences of the United States of America*, PMID: 25646426, PMCID: PMC4343152, v. 112, n. 7, p. 2058–2063, 17 fev. 2015.

HERRERO, Javier *et al.* Ensembl comparative genomics resources. *Database*, PMID: 26896847, v. 2016, p. bav096, 1 jan. 2016.

HILLIS, D. M.; HUELSENBECK, J. P.; CUNNINGHAM, C. W. Application and accuracy of molecular phylogenies. *Science*, PMID: 8171318, v. 264, n. 5159, p. 671–677, 29 abr. 1994.

HUERTA-CEPAS, Jaime *et al.* PhylomeDB: a database for genome-wide collections of gene phylogenies. *Nucleic Acids Research*, PMID: 17962297, PMCID: PMC2238872, v. 36, n. Database issue, p. D491–D496, jan. 2008.

HUERTA-CEPAS, Jaime; DOPAZO, Joaquín; GABALDÓN, Toni. ETE: a python Environment for Tree Exploration. *BMC Bioinformatics*, v. 11, p. 24, 2010.

KANEHISA, Minoru *et al.* KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, PMID: 26476454, p. gkv1070, 17 out. 2015.

KIMURA, Motoo. *The neutral theory of molecular evolution*. [S.l.]: Cambridge University Press, 1984.

KRIVENTSEVA, Evgenia V. *et al.* OrthoDB v8: update of the hierarchical catalog of orthologs and the underlying free software. *Nucleic Acids Research*, PMID: 25428351, PMCID: PMC4383991, v. 43, n. Database issue, p. D250-256, jan. 2015.

LANG, Jenna Morgan; DARLING, Aaron E.; EISEN, Jonathan A. Phylogeny of Bacterial and Archaeal Genomes Using Conserved Genes: Supertrees and Supermatrices. *PLOS ONE*, v. 8, n. 4, p. e62510, abr 2013.

LASSMANN, Timo; FRINGS, Oliver; SONNHAMMER, Erik L. L. Kalign2: high-performance multiple alignment of protein and nucleotide sequences allowing external features.

Nucleic Acids Research, PMID: 19103665, PMCID: PMC2647288, v. 37, n. 3, p. 858–865, fev. 2009.

LETUNIC, I.; BORK, P. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Research*, v. 39, n. Web Server, p. W475–W478, 5 abr. 2011.

LIN, Harris T.; BURLEIGH, J. Gordon; EULENSTEIN, Oliver. Triplet supertree heuristics for the tree of life. *BMC Bioinformatics*, v. 10, n. 1, p. 1–12, 2009.

LINNÉ, Carl Von; SALVIUS, Lars. *Systema naturae per regna tria naturae :secundum classes, ordines, genera, species, cum characteribus, differentiis, synonymis, locis*. Holmiae : Impensis Direct. Laurentii Salvii, 1758. v. v.1..

LÖYTYNOJA, Ari; GOLDMAN, Nick. webPRANK: a phylogeny-aware multiple sequence aligner with interactive alignment browser. *BMC Bioinformatics*, PMID: 21110866, v. 11, n. 1, p. 579, 26 nov. 2010.

MELO, HVF; ORTEGA, J. M. *Ferramentas e serviços online para a análise da origem cladística de genes e vias metabólicas*. 2014. Universidade Federal de Minas Gerais, Belo Horizonte - MG, Brasil, 2014.

MI, Huaiyu; MURUGANUJAN, Anushya; THOMAS, Paul D. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic acids research*, PMID: 23193289, PMCID: PMC3531194, v. 41, n. Database issue, p. D377-386, jan. 2013.

MORA, Camilo *et al.* How Many Species Are There on Earth and in the Ocean? *PLOS Biol*, v. 9, n. 8, p. e1001127, ago 2011.

MOSSES, Christopher; PEDERSEN, Christian Nørgaard Storm. *Triplet supertrees*. 2005. PhD Thesis, University of Aarhus, Aarhus, Denmark, 2005.

NGUYEN, Duc; XU, Tian. The expanding role of mouse genetics for understanding human biology and disease. *Disease Models & Mechanisms*, PMID: 19048054, PMCID: PMC2561976, v. 1, n. 1, p. 56–66, 2008.

NICHOLS, R. Gene trees and species trees are not the same. *Trends in Ecology & Evolution*, PMID: 11403868, v. 16, n. 7, p. 358–364, 1 jul. 2001.

NOVACEK, Michael J. Mammalian phylogeny: Genes and supertrees. *Current Biology*, v. 11, n. 14, p. R573–R575, 24 jul. 2001.

NUTTALL, George HF. *Blood immunity and blood relationship*. [S.l.]: Cambridge University Press, 1904.

OLAF R. P. BININDA-EMONDS; JOHN L. GITTLEMAN; STEEL, Mike A. The (Super)Tree of Life: Procedures, Problems, and Prospects. *Annual Review of Ecology and Systematics*, v. 33, n. 1, p. 265–289, 2002.

PALIDWOR, Gareth; REYNAUD, Emmanuel G.; ANDRADE-NAVARRO, Miguel A. Taxonomic colouring of phylogenetic trees of protein sequences. *BMC bioinformatics*, PMID: 16503967, PMCID: PMC1386715, v. 7, p. 79, 2006.

PHIFER-RIXEY, Megan *et al.* Adaptive Evolution and Effective Population Size in Wild House Mice. *Molecular Biology and Evolution*, PMID: 22490822, PMCID: PMC3457769, v. 29, n. 10, p. 2949–2955, out. 2012.

PHILIPPE, Hervé *et al.* Phylogenomics. *Annual Review of Ecology, Evolution, and Systematics*, v. 36, n. 1, p. 541–562, 2005.

PHILIPPE, Hervé *et al.* Resolving Difficult Phylogenetic Questions: Why More Sequences Are Not Enough. *PLOS Biol*, v. 9, n. 3, p. e1000602, 15 mar. 2011.

PIAGGIO-TALICE, Raul; BURLEIGH, J Gordon; EULENSTEIN, Oliver. Quartet supertrees. *Phylogenetic Supertrees*. [S.l.]: Springer, 2004. p. 173–191.

PIEL, William H *et al.* TreeBASE: a database of phylogenetic information. 2000, [S.l: s.n.], 2000.

POWELL, Sean *et al.* eggNOG v4.0: nested orthology inference across 3686 organisms. *Nucleic Acids Research*, PMID: 24297252, p. gkt1253, 1 dez. 2013.

PRICE, Morgan N.; DEHAL, Paramvir S.; ARKIN, Adam P. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE*, v. 5, n. 3, p. e9490, 10 mar. 2010.

PUMO, D. E. *et al.* Complete mitochondrial genome of a neotropical fruit bat, Artibeus jamaicensis, and a new hypothesis of the relationships of bats to other eutherian mammals. *Journal of Molecular Evolution*, PMID: 9847413, v. 47, n. 6, p. 709–717, dez. 1998.

QUEIROZ, Alan De; GATESY, John. The supermatrix approach to systematics. *Trends in Ecology & Evolution*, PMID: 17046100, 17046100, v. 22, n. 1, p. 34–41, 1 jan. 2007.

RAGAN, M. A. Phylogenetic inference based on matrix representation of trees. *Molecular Phylogenetics and Evolution*, PMID: 1342924, v. 1, n. 1, p. 53–58, mar. 1992.

RAMBAUT, A. *FigTree v1. 3.1: Tree figure drawing tool*. [S.l: s.n.], 2009. Disponível em: <http://tree.bio.ed.ac.uk/software/figtree/>.

RANWEZ, Vincent; CRISCUOLO, Alexis; DOUZERY, Emmanuel J. P. SuperTriplets: a triplet-based supertree approach to phylogenomics. *Bioinformatics*, PMID: 20529895, v. 26, n. 12, p. i115–i123, 15 jun. 2010.

REDDY, TBK *et al.* The Genomes OnLine Database (GOLD) v. 5: a metadata management system based on a four level (meta) genome project classification. *Nucleic acids research*, p. gku950, 2014.

REGIER, Jerome C. *et al.* Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. *Nature*, PMID: 20147900, v. 463, n. 7284, p. 1079–1083, 25 fev. 2010.

ROSKOV, Y *et al. Species 2000 & ITIS Catalogue of Life*. Disponível em: <http://www.catalogueoflife.org/>. Acesso em: 8 jul. 2016.

ROSS, Howard A.; RODRIGO, Allen G. An Assessment of Matrix Representation with Compatibility in Supertree Construction. In: BININDA-EMONDS, OLAF R. P. (Org.). . *Phylogenetic Supertrees*. Computational Biology. [S.l.]: Springer Netherlands, 2004. p. 35–63.

SAITOU, N.; NEI, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, PMID: 3447015, v. 4, n. 4, p. 406–425, 1 jul. 1987.

SALICHOS, Leonidas; ROKAS, Antonis. Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature*, PMID: 23657258, v. 497, n. 7449, p. 327–331, 16 maio 2013.

SANDERSON, M. J.; PURVIS, A.; HENZE, C. Phylogenetic supertrees: Assembling the trees of life. *Trends in Ecology & Evolution*, PMID: 21238221, v. 13, n. 3, p. 105–109, mar. 1998.

SAYERS, Eric W. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, PMID: 18940862, PMCID: PMC2686545, v. 37, n. Database issue, p. D5-15, jan. 2009.

SCHREIBER, Fabian *et al.* TreeFam v9: a new website, more species and orthology-on-the-fly. *Nucleic Acids Research*, PMID: 24194607, PMCID: PMC3965059, v. 42, n. Database issue, p. D922–D925, 1 jan. 2014.

SEOK, Junhee *et al.* Genomic responses in mouse models poorly mimic human inflammatory diseases. *Proceedings of the National Academy of Sciences*, PMID: 23401516, v. 110, n. 9, p. 3507–3512, 26 fev. 2013.

SIBLEY, C. G.; AHLQUIST, J. E. The phylogeny of the hominoid primates, as indicated by DNA-DNA hybridization. *Journal of Molecular Evolution*, PMID: 6429338, v. 20, n. 1, p. 2–15, 1984.

SIEVERS, Fabian; HIGGINS, Desmond G. Clustal Omega, accurate alignment of very large numbers of sequences. *Methods in Molecular Biology (Clifton, N.J.)*, PMID: 24170397, v. 1079, p. 105–116, 2014.

SNEL, Berend; HUYNEN, Martijn A.; DUTILH, Bas E. Genome Trees and the Nature of Genome Evolution. *Annual Review of Microbiology*, PMID: 16153168, v. 59, n. 1, p. 191–209, 2005.

SOKAL, Robert R. A statistical method for evaluating systematic relationships. *Univ Kans Sci Bull*, v. 38, p. 1409–1438, 1958.

SOKAL, Robert R. Numerical Taxonomy. *Scientific American*, v. 215, p. 106–116, 1966.

SONG, Sen *et al.* Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. *Proceedings of the National Academy of Sciences of the United States of America*, PMID: 22930817, PMCID: PMC3443116, v. 109, n. 37, p. 14942–14947, 11 set. 2012.

SONNHAMMER, Erik L. L.; ÖSTLUND, Gabriel. InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Research*, PMID: 25429972, p. gku1203, 27 nov. 2014.

STAMATAKIS, Alexandros. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, PMID: 16928733, v. 22, n. 21, p. 2688–2690, 1 nov. 2006.

TATUSOV, Roman L *et al.* The COG database: an updated version includes eukaryotes. *BMC bioinformatics*, PMID: 12969510 PMCID: PMC222959, v. 4, p. 41, 11 set. 2003.

The NCBI Handbook. 2nd. ed. [S.l.]: National Center for Biotechnology Information (US), 2013.

VANHOOREN, Valerie; LIBERT, Claude. The mouse as a model organism in aging research: usefulness, pitfalls and possibilities. *Ageing Research Reviews*, PMID: 22543101, v. 12, n. 1, p. 8–21, jan. 2013.

VON HAESELER, Arndt. Do we still need supertrees? *BMC Biology*, v. 10, p. 13, 2012.

WHITFIELD, James B.; KJER, Karl M. Ancient rapid radiations of insects: challenges for phylogenetic analysis. *Annual Review of Entomology*, PMID: 17877448, v. 53, p. 449–472, 2008.

WIENS, John J. The Role of Morphological Data in Phylogeny Reconstruction. *Systematic Biology*, PMID: 15371253, v. 53, n. 4, p. 653–661, 1 ago. 2004.