UNIVERSIDADE FEDERAL DE CIÊNCIAS DA SAÚDE DE PORTO ALEGRE

INFORMÁTICA BIOMÉDICA

MEISKI MARIÁ VEDOVATTO

GENÔMICA EVOLUTIVA DE BACTÉRIAS DO GÊNERO *MYCOPLASMA*: ANÁLISE TEÓRICO-COMPUTACIONAL E DESENVOLVIMENTO DE *PIPELINE*

PORTO ALEGRE
2018

Meiski Mariá Vedovatto

GENÔMICA EVOLUTIVA DE BACTÉRIAS DO GÊNERO *MYCOPLASMA*: análise teórico-computacional e desenvolvimento de *pipeline*

Monografia apresentada como requisito parcial para obtenção do grau de Bacharel em Informática Biomédica da Universidade Federal de Ciências da Saúde de Porto Alegre.

Orientador: Prof[a]. Dr.[a] Claudia Elizabeth Thompson

Porto Alegre
2018

Meiski Mariá Vedovatto

GENÔMICA EVOLUTIVA DE BACTÉRIAS DO GÊNERO *MYCOPLASMA*: análise teórico-computacional e desenvolvimento de *pipeline*

Monografia apresentada como requisito parcial para obtenção do grau de Bacharel em Informática Biomédica da Universidade Federal de Ciências da Saúde de Porto Alegre.

Data de Aprovação: 11/12/2018

Banca Examinadora

Profª. Drª. Claudia Elizabeth Thompson
Universidade Federal de Ciências da Saúde de Porto Alegre
Departamento de Farmacociências
Orientadora

Profª. Drª. Ana Trindade Winck
Universidade Federal de Ciências da Saúde de Porto Alegre
Departamento de Ciências Exatas e Sociais Aplicadas
Examinadora

Profª. Drª. Loreta Brandão de Freitas
Universidade Federal do Rio Grande do Sul
Departamento de Genética
Examinadora

AGRADECIMENTOS

Gostaria de agradecer a quem me auxiliou de perto e de longe, direta e indiretamente no decorrer desse Trabalho de Conclusão de Curso. Em especial, agradecer à minha orientadora, Profª. Dr.ª Claudia E. Thompson que, além de me guiar pela apaixonante área que é a evolução, me orientou na execução desse trabalho e sempre incentivou o pensamento científico, crítico e metodológico.

Agradeço também ao Fábio C. Andreis, por colaborar enormemente com meu desenvolvimento tanto sob o ponto de vista técnico quanto pessoal durante esse período, demonstrando o caráter de um grande pesquisador e agregando muito em cada etapa. Adicionalmente, agradeço à Flavielle Marques e Eduardo Pooch, antes colegas de graduação, agora amigos de formação, da vida e parceiros da jornada frenética que é ser a primeira turma de Informática Biomédica do Estado.

Trabalhando com Bioinformática, área tão interdisciplinar, não posso deixar de agradecer o papel desempenhado por todos os meus professores, os quais contribuíram de diferentes formas em minha formação e ajudaram, portanto, para que hoje eu fosse capaz de desenvolver o Trabalho de Conclusão.

O agradecimento à minha família, por todo o suporte, é imensurável e espero, um dia, ser capaz de retribuir tanto amor e paciência que investiram em mim nessa fase (e em todas as outras).

RESUMO

O estudo da genômica utilizando ferramentas de filogenômica tem sido possibilitado devido à disponibilidade de um grande número de sequências depositadas nos bancos de dados, resultantes do desenvolvimento de novas tecnologias de sequenciamento. A importância desse tipo de abordagem reside no fato de permitir uma melhor compreensão dos processos evolutivos que moldam os genomas e do relacionamento evolutivo entre eles. Adicionalmente, os estudos *in silico* possibilitam a obtenção de respostas para diversas perguntas biológicas referentes à evolução molecular dos genomas e de seus genes e proteínas. Bactérias do gênero *Mycoplasma* possuem um genoma de tamanho relativamente pequeno e grande dependência dos nutrientes supridos pelo hospedeiro. Exercem os mais diversos estilos de vida, sendo que a maior parte das espécies são parasitas responsáveis por doenças em humanos, outros animais e plantas. Uma importante hipótese evolutiva sobre esse grupo indica que passou por um processo de evolução degenerativa que levou à perda da parede celular. Em 2013, foi realizado um estudo que analisou as relações evolutivas entre 31 espécies do gênero *Mycoplasma*. Atualmente, mais genomas estão disponíveis nos bancos de dados biológicos públicos e a análise das sequências depositadas pode levar a um melhor entendimento de aspectos evolutivos relacionados a essas bactérias. O principal objetivo deste trabalho foi o desenvolvimento de um *pipeline* de análise filogenômica focado na análise de genomas bacterianos e a caracterização genômica e evolutiva do gênero *Mycoplasma* a partir de 83 espécies utilizando ferramentas como Proteinortho, GUIDANCE, Prottest, PhyML and MrBayes. Nosso *pipeline* representa um importante avanço para a automatização de processos e análises relacionadas à genômica funcional evolutiva de espécies bacterianas.

Palavras-chave: Genômica Evolutiva. Filogenômica. *Mycoplasma*. *Pipelines*. Bioinformática

ABSTRACT

The study of genomics using phylogenetic tools has increased due to the availability of several sequences in databases, resulting from the development and application of new technologies of DNA sequencing. Phylogenomic approaches allow a better comprehension of the evolutionary processes occurring in the genomes and the understanding of the evolutionary history of genomes belonging to different species. Additionally, computational studies lead to answers to different biological questions related to the molecular evolution of genomes and their genes and corresponding proteins. Bacteria of the *Mycoplasma* genus have small genomes and are highly dependent on the nutrients supplied by the host. They show different lifestyles, with most of them being parasites and responsible for diseases in humans, other animals and even plants. An important evolutionary hypothesis about this group is that Mycoplasmas underwent a process of degenerative evolution resulting in the loss of the cell wall. In 2013, a study identified the evolutionary relationships among 31 species of *Mycoplasma*. Nowadays, more genomes are available in the biological databanks and we have performed new analyses including 89 genomes from *Mycoplasma* to obtain a more comprehensive understanding of the evolutionary aspects related to these bacteria. The main objective of this work was the development of a pipeline of phylogenomic analysis and the genomic characterization of the *Mycoplasma* genus including 83 species using tools such as Proteinortho, GUIDANCE, Prottest, PhyML and MrBayes. Our pipeline represents an important step to automatize processes and analyses related to evolutionary functional genomics of bacterial species.

Keywords: Evolutionary Genomics. Phylogenomics. *Mycoplasma*. *Pipelines*. Bioinformatics.

LISTA DE FIGURAS

LISTA DE ABREVIATURAS E SIGLAS

ARN	Ácido Ribonucleico
CARDS	Toxinas da Síndrome do Desconforto Respiratório Adquirida na Comunidade (do inglês *Community-Acquired Respiratory Distress Syndrome Toxin*)
GO	Grupos Ortólogos
NCBI	Centro Nacional de Informação Biotecnológica (do inglês *National Center for Biotechnology Information*)
NEAC	Atividade Enzimática de Neuraminidases (do inglês *Neuraminidase Enzymatic Activity*)
OTU	Unidade taxonômica operacional (do inglês *Operational Taxonomic Unit*)
SNP	*Single Nucleotide Polymorphism*
THG	Transferência Horizontal de Genes
wgMLST	*whole genome Multilocus Sequence Typing*
WGS	Sequenciamento Completo do Genoma *Whole Genome Sequencing*

SUMÁRIO

1 INTRODUÇÃO

A partir da descoberta do material genético (Dahm, 2005), o entendimento de sua estrutura (Watson e Crick, 1953) e o desenvolvimento de técnicas de sequenciamento, houve o início da corrida para que pudéssemos decifrar e conhecer completamente as sequências de DNA dos organismos. Em 1978, foi publicado o primeiro genoma por Sanger et. al. (SANGER et al., 1978), do bacteriófago $\phi X174$, seguido pela publicação de diversos outros genomas de procariotos, como o da bactéria *Haemophilus influenzae* (FLEISCHMANN et al., 1995), da primeira arquea *Methanococcus jannaschii* (BULT et al., 1996) e do primeiro genoma eucarioto, referente à levedura *Saccharomyces cerevisiae* (GOFFEAU et al., 1996).

A partir de 2001, com a publicação do *draft* do genoma humano (VENTER et al., 2001; CONSORTIUM et al., 2001) e com o subsequente surgimento de tecnologias denominadas *next-generation sequencing* (NGS), algumas das quais, atualmente, capazes de sequenciar moléculas inteiras (LU et al., 2016; DIJK et al., 2014), houve um enorme crescimento do número de genomas sequenciados, tanto de procariotos quanto de eucariotos. O desenvolvimento de métodos mais acurados de montagem e anotação de genomas, somado ao constante aprimoramento das técnicas de sequenciamento, resultou em um aumento exponencial da disponibilidade de genomas nas bases de dados biológicos. Esse avanço impactou diretamente e de forma positiva a prática clínica, permitindo avanços importantes na área de medicina personalizada (CROWGEY, 2016) junto à bioinformática translacional (TENENBAUM, 2016).

Entretanto, o crescimento das bases de dados caracteriza um desafio para a análise dessas informações, demandando técnicas computacionais que considerem o significado biológico intrínseco dos fragmentos depositados. Para tanto, são desenvolvidas ferramentas de montagem de genomas, anotação funcional de genes, identificação de similaridades, análises comparativas e evolutivas. A análise evolutiva, por sua vez, possui grande relevância no entendimento da genômica evolutiva, que leva em consideração as diferenças genéticas entre os organismos, permitindo identificar e descrever os processos de transferência horizontal gênica, duplicação, existência de elementos móveis (BROWN, 2002a), ancestralidade, função e compreensão dos processos que moldam a evolução dos organismos (MARTINEZ-URTAZA et al., 2017).

A filogenética, estudo da história evolutiva dos genes, depende de algoritmos para a identificação de ortólogos, alinhamento múltiplo de sequências, identificação e processamento de modelos evolutivos e reconstrução de árvores filogenéticas (CURRAT et al., 2015; ZHANG; LIN, 2015). Essa análise considera diversas características, entre elas estão a história evolutiva

do organismo, as taxas de substituição (BROWN, 2002b), informações sobre sítios conservados (LAING et al., 2017), identificação de regiões repetidas (ZAHA et al., 2014; LEI et al., 2017), sintenia (GUIMARAES et al., 2014), etc. Na literatura, são bem descritos estudos filogenéticos que inferem as relações evolutivas entre genes de diferentes organismos (HODGE et al., 2000), famílias gênicas e proteínas (GABALDÓN, 2005), identificando como suas sequências e funções sofreram alterações no decorrer do tempo.

De modo complementar, a filogenômica atua sobre o estudo de genomas, considerando, assim, um maior número de ortólogos entre os diferentes organismos e, portanto, possuindo o potencial de inferir com maior confiabilidade a história evolutiva de organismos pertencentes a diferentes níveis taxonômicos (e.g. gênero, espécies) (DELSUC et al., 2005; EISEN; FRASER, 2003). Marcadores biológicos (SALEMI et al., 2009), presença de polimorfismos de nucleotídeo único (SNP, do inglês *Single Nucleotide Polymorphism*) (LAING et al., 2017) e o número de genes resultante da identificação de ortólogos são variantes adicionais importantes utilizadas para descrever a história evolutiva do organismo. A evolução do *Candidatus Hepatoplasma crinochetorum*, por exemplo, foi inferida a partir de 127 genes da classe *Mollicutes*, identificando *H. chinochetorun* como um ramo irmão do clado Hominis de *Mycoplasma* (LECLERCQ et al., 2014). Evidências como essa e outras são relevantes para o diagnóstico de agentes patogênicos, classificação dos organismos (MAKIMURA et al., 1999), conhecimento dos componentes mínimos necessários à sobrevivência sob longos períodos de seleção natural e no rastreamento de eventos de transferência gênica (LIU et al., 2012).

A genômica evolutiva pode contribuir também para a identificação de padrões evolutivos, o que é relevante para o desenvolvimento de fármacos, identificação de resistência antimicrobiana (TOPRAK et al., 2012) e para o estudo de forças evolutivas em diferentes condições (CURRAT et al., 2015; FREED et al., 2015). Entretanto, a genômica evolutiva se apresenta como um desafio teórico-computacional para a comunidade científica devido a sua abordagem multivariada (ITAN et al., 2015) e magnitude de processamento (BAICHOO; OUZOUNIS, 2017). A inevitável interação com diversos domínios científicos aumenta o grau de complexidade das análises e exige um nível de automação cada vez maior, sem perda de desempenho e acurácia.

Por essa razão, é crescente o estudo de *pipelines* em bioinformática com ênfase em configuração e padronização de *software* (LEIPZIG, 2017), com o propósito principal de estabelecer parâmetros no desenvolvimento que propiciem a disseminação do conhecimento e o crescimento da área, ressaltando a necessidade de algoritmos robustos que compreendam as técnicas de manipulação de dados, possibilitando controle total sobre a análise.

1.1 MICOPLASMAS

Micoplasmas, pertencentes à classe *Mollicutes* e ordem *Mycoplasmatales*, são caracterizadas pela ausência de parede celular, genoma de tamanho relativamente pequeno, entre 0,58 Mb e 2 Mb (CITTI et al., 2018), resultando em uma grande dependência do hospedeiro para o suprimento de nutrientes devido à incapacidade de sintetizar aminoácidos (HIMMELREICH et al., 1996). Foram relatadas 160 espécies diferentes do gênero (MAY et al., 2014), as quais possuem os mais diversos estilos de vida, sendo a maior parte parasitas responsáveis por doenças em humanos, outros animais e plantas (RAZIN et al., 1998). O primeiro estudo do gênero, em 1898, foi com um agente de pleuropneumonia bovina, que foi considerado um organismo viral. Somente em 1960, a partir do método de hibridização de DNA, foi possível identificar esses organismos como bactérias que não possuem parede celular (RAZIN; HAYFLICK, 2010). Os termos mollicutes e micoplasma são, muitas vezes, utilizados de forma intercambiável para falar da classe; no presente estudo, o termo micoplasma será relacionado somente ao gênero dessas bactérias.

A primeira espécie identificada como agente patogênico em humanos foi *Mycoplasma pneumoniae*, causadora de doenças no trato respiratório inferior e superior. *Mycoplasma genitalium* (LJUBIN-STERNAK; MEŠTROVIĆ, 2014) e *Mycoplasma penetrans* (SASAKI et al., 2002) também são causadoras de infecções, já *Mycoplasma fermentans* e *Mycoplasma hominis*, por exemplo, foram encontradas somente em quadros clínicos pré-estabelecidos, sendo consideradas parasitas oportunistas, bem como outras espécies do gênero. Apesar da associação de micoplasmas com algumas doenças ainda não ser muito clara, é possível afirmar que há uma relação entre as ocorrências, pois diversas vezes foram identificadas em pacientes imunocomprometidos, como portadores de HIV (MAVEDZENGE; WEISS, 2009; WANG et al., 1992; HANNAN, 1998) e em células tumorais (BARYKOVA et al., 2011; XIE et al., 2017), demonstrando o perfil oportunista do micro-organismo.

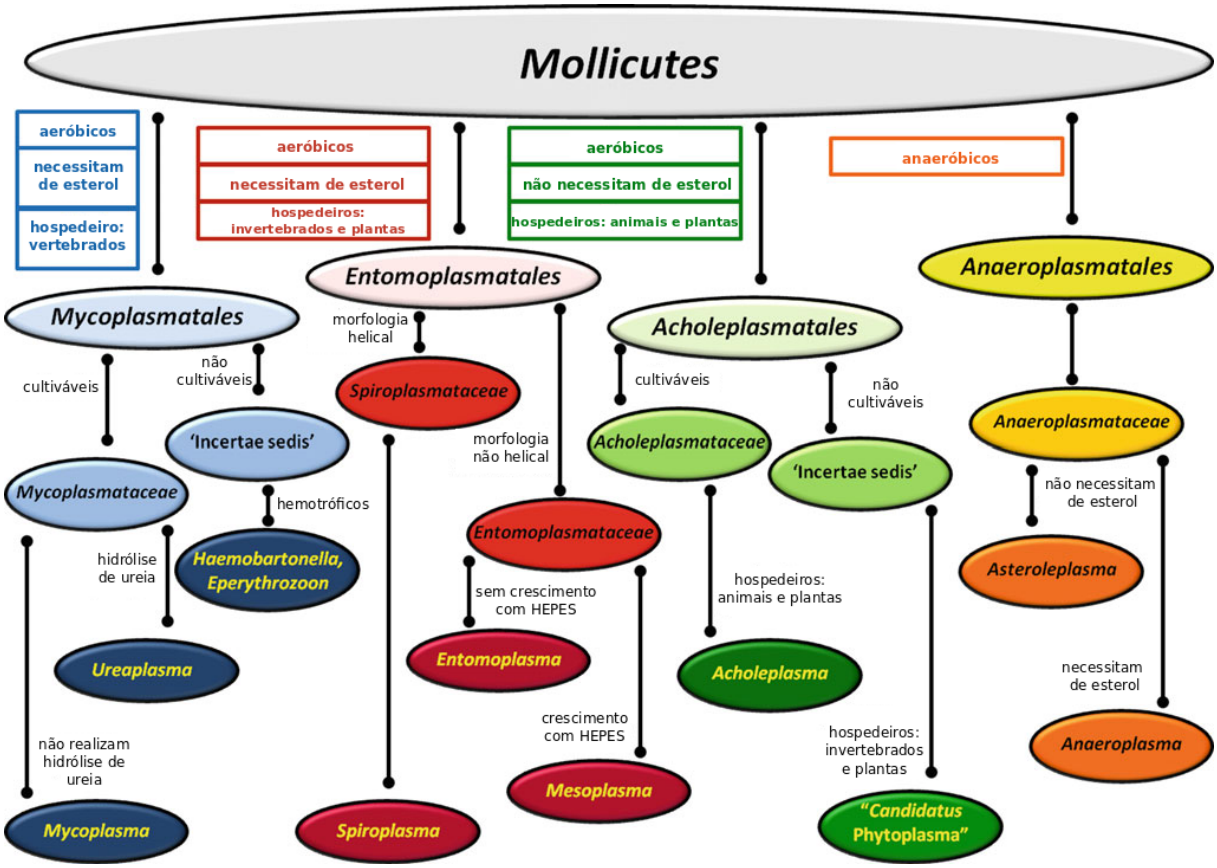
O trato respiratório e urogenital, glândulas mamárias, mucosas e olhos são os nichos preferenciais de micoplasmas em humanos. Entretanto, espécies do gênero já foram encontradas em células cerebrais de humanos (CHRISTO et al., 2010), focas e bovinos (ROSALES et al., 2017; TSIODRAS et al., 2005), vinculadas a quadros de encefalites, meningite asséptica e coinfecções afetando o Sistema Nervoso Central (TSIODRAS et al., 2005). Dentre os fatores de virulência das micoplasmas estão a liberação de toxinas responsáveis pela síndrome do desconforto respiratório adquirida na comunidade (CARDS, do inglês *Community-Acquired*

Respiratory Distress Syndrome Toxin) (KANNAN; BASEMAN, 2006), atividade de enzimática de neuraminidases (NEAC, do inglês *Neuraminidase Enzymatic Activity*) (BERČIČ et al., 2008) e de adesinas (ROTTEM, 2003), fundamentais para justificar a capacidade de adesão e evasão do sistema imune por infiltração no citoplasma celular (ROSENGARTEN et al., 2000).

A primeira hipótese evolutiva, proposta por Morowitz e Wallace (MOROWITZ; WALLACE, 1973), é de que micoplasmas são os seres mais primitivos existentes, anteriores ao desenvolvimento da camada celular de peptidoglicano. Entretanto, desde 1960, Neimark (NEIMARK, 1986) defendia a ideia de que, na verdade, o grupo passou por um processo de evolução degenerativa, confirmado por filogenia de RNA ribossomal (rRNA) (WOESE et al., 1980). A hipótese evolutiva melhor estabelecida, portanto, é de que essas bactérias teriam passado por um processo de evolução degenerativa e, adicionalmente, por uma série de ciclos que resultaram na diminuição de seu genoma e perda de parede celular. Atualmente, micoplasmas são considerados os menores e mais simples organismos capazes de autorreplicação (RAZIN; HAYFLICK, 2010), compartilhando um ancestral comum com bactérias gram-positivas, do ramo de *Streptoccocus* e estima-se que esse grupo divergiu há 600 milhões de anos (RAZIN et al., 1998).

Essa divergência resultou em dois ramos principais, os quais se diferenciaram há 400 milhões de anos. O primeiro (ramo AAA) é constituído por 3 gêneros, *Asteroleplasma*, *Anaeroplasma* e *Acholeplasma* e o segundo (SEM) é constituído por *Spiroplasma*, *Entomoplasma*, e *Mycoplasma* (MANILOFF, 1996). Na Figura 1, é possível observar características gerais de micoplasmas: são organismos aeróbicos, requerem esterol, estão presentes em vertebrados e não realizam hidrólise de ureia. O gênero é dividido em três grupos internos: Hominis, Pneumoniae e Mycoides (BROWN, 2010), sendo o último mais proximamente relacionado a *Mesoplasma* e *Entemoplasma*. Micoplasmas do grupo *mycoides* possuem uma história evolutiva diferente dos outros dois clados, pois seriam originários de um ancestral associado a insetos e se tornaram fenotipicamente semelhante a outras linhagens de *Mycoplasma* por meio de eventos independentes, por evolução convergente envolvendo transferência horizontal de genes (THG) (LO et al., 2018).

Figura 1 – Estrutura da classe *Mollicutes*.



Organização taxonômica e diferentes características fenotípicas e evolutivas dos grupos de molicutes. Com exceção das bactérias pertencentes aos gêneros *Mycoplasma* e *Ureaplasma*, todos os outros gêneros da classe não colonizam humanos (RAZIN, 2006). A variedade de hospedeiros atualmente conhecidos consiste de 49 espécies de mamíferos, incluindo humanos, 39 espécies de aves, 10 espécies de répteis, uma espécie de peixes e nenhuma espécie de anfíbio (MAY et al., 2014). HEPES é referente a solução tampão. Fonte: modificado de MAY et al., 2014.

Estudos de genômica comparativa de micoplasmas mostram a presença de regiões repetidas em diversas espécies, conferindo plasticidade fenotípica e contribuindo para os fatores de virulência e variação antigênica em *Mycoplasma hyopneumoniae* (BARATE et al., 2014), *M. pneumoniae* (MUSATOVOVA et al., 2012), *M. genitalium* (MA et al., 2010), *Mycoplasma hyorhinis* (YOGEV et al., 1991) e outros (ZHANG; WISE, 1996; SIMMONS et al., 2004; LIU et al., 2000; LYSNYANSKY et al., 1996). A presença dessas repetições é importante considerando o processo evolutivo do gênero que, mesmo sob pressão para redução do genoma, conservou esses "reservatórios evolutivos"(CATTANI, 2016), os quais muito provavelmente estariam sob pressão seletiva mais intensa (ROCHA; BLANCHARD, 2002).

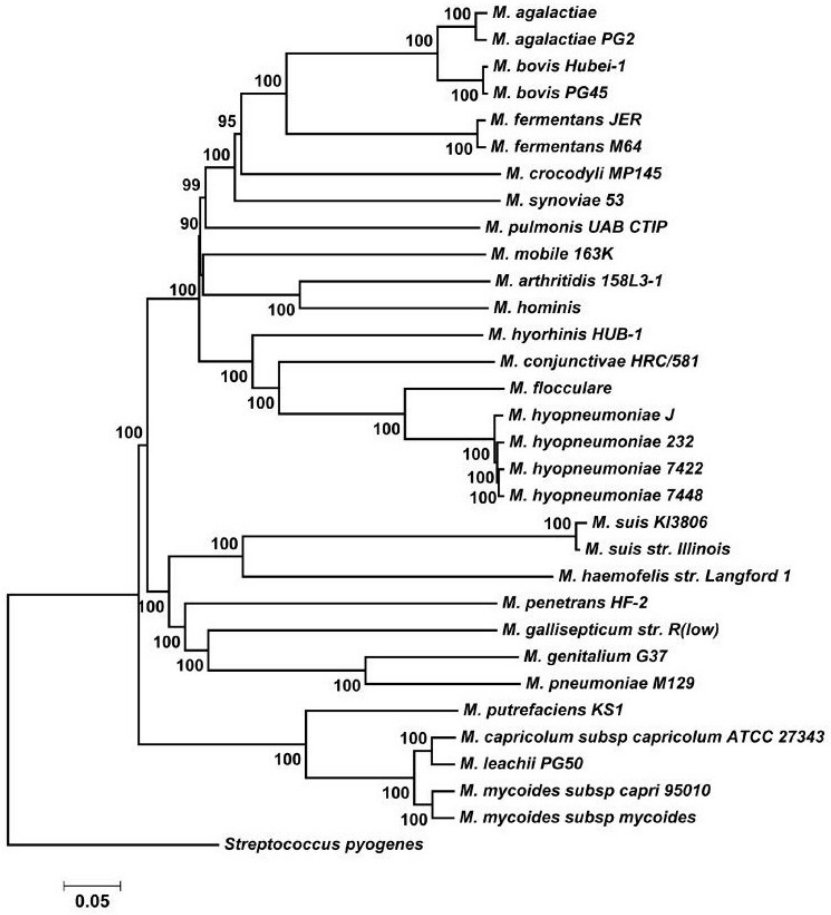
Os grupos polifiléticos *spiroplasma*, *mycoplasma*, *ureaplasma* e *mesoplasma* usam a

trinca UGA para codificar triptofano (INAMINE et al., 1990), não a utilizando como *stop codon*, como ocorre no código genético universal. Esses organismos utilizam, portanto, UAA e UAG para terminação (RAZIN et al., 1998). Essa mudança se dá pela pressão sobre o processo de tradução com uso preferencial de códons (MUTO; OSAWA, 1987), mantendo regiões gênicas conservadas com conteúdo G-C maior e apresentando regiões intergênicas ricas em A-T, podendo chegar a 90% de adenina e timina (RAZIN et al., 1998), principalmente porque a composição das trincas pode exibir até 93% de A-T na terceira posição em *Mycoplasma capricolum* (SHARP et al., 1993), por exemplo.

Os genomas de várias espécies de micoplasmas foram sequenciados nos últimos anos. Em 2007, havia apenas 13 genomas disponíveis (SIRAND-PUGNET et al., 2007), hoje já há 90 espécies sequenciadas e um total de 379 genomas depositados no Genbank. Dentre esses depósitos, estão diversas cepas patogênicas, como as espécies mais importantes identificadas no sistema respiratório de suínos: *M. hyopneumoniae*, *M. hyorhinis* e *Mycoplasma flocculare* (MARE, 1965; MEYLING; FRIIS, 1972). *M. hyopneumoniae* está associada à pneumonia nesses animais, enquanto *M. hyorhinis* está associada à poliserosite e artrite. Já *M. flocculare* não causa doença no hospedeiro e está restrita ao trato respiratório (FRIIS; FEENSTRA, 1994; KOBISCH; FRIIS, 1996).

Siqueira et al. realizaram o sequenciamento de *M. flocculare* e *M. hyopneumoniae* 7422 (SIQUEIRA et al., 2013), a análise evolutiva desses genomas e outras 29 espécies de micoplasmas (Figura 2), por meio de métodos de análise filogenômica, bem como descreveram a história evolutiva de alguns genes duplicados. Adicionalmente, neste estudo foram analisadas proteínas de superfície de *M. flocculare*, *M. hyopneumoniae* 7448 e *M. hyorhinis* HUB-1, sendo que muitas proteínas compartilhadas entre *M. hyopneumoniae* e *M. hyorhinis* foram identificadas como produtos de genes putativos relacionados à patogênese.

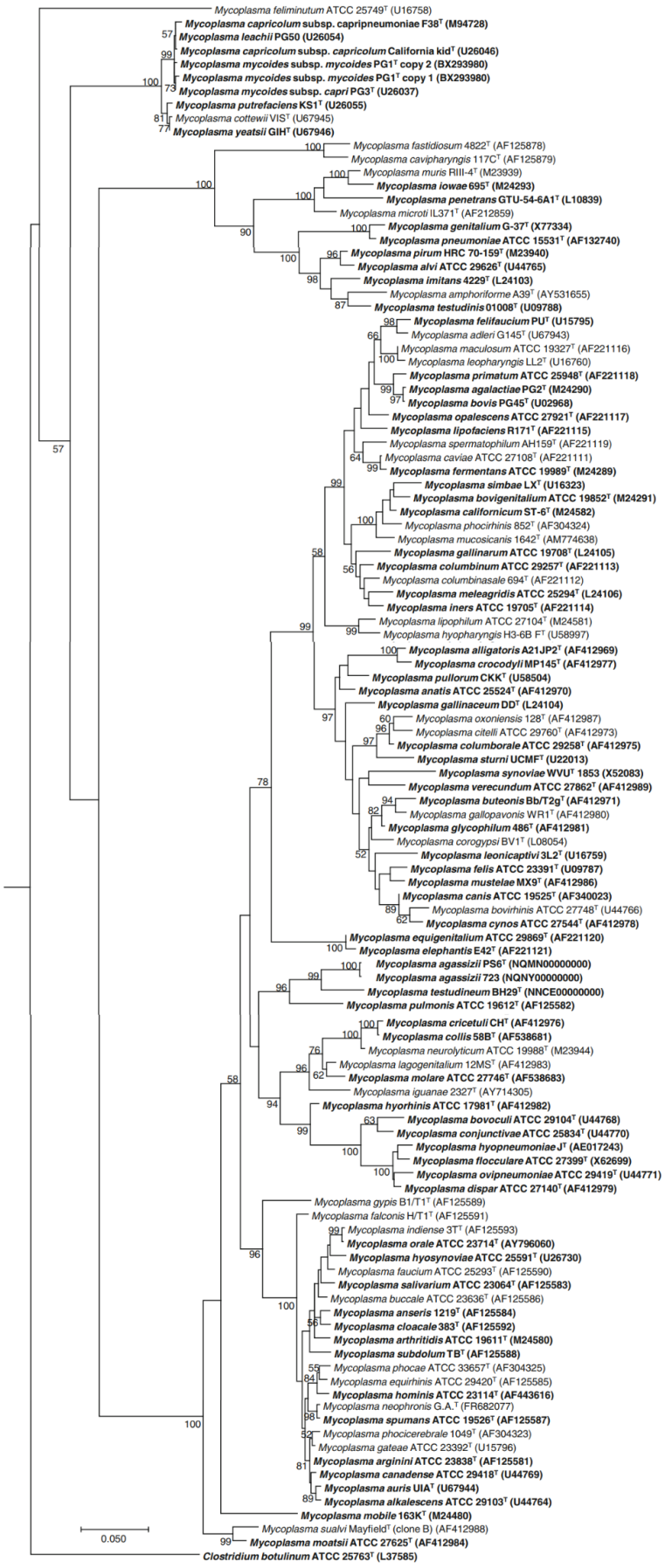
Figura 2 – História evolutiva de 31 espécies de micoplasmas baseada na análise filogenômica de 179 genes.



Árvore filogenômica de micoplasmas obtida com base na análise de 179 genes, utilizando o método da super-matriz, algoritmo *neighbor-joining*, distância p e *pairwise deletion* para o tratamento dos gaps. Fonte: Siqueira et al, 2013.

Em 2018, ALVAREZ-PONCE et al. utilizaram ARN (ácido ribonucleico) ribossomal (16S) para classificar duas novas cepas de *M. agassizii* isoladas de tartarugas do deserto, posicionando-as em um grupo monofilético juntamente a *M. testudineum* e *M. pulmonis* (Figura 3), em concordância com MAY et al. Apesar de o estudo ter alcançado seu objetivo em relação a *M. agassizii*, é interessante observarmos que aproximadamente 30 ramos da árvore não possuíram suporte estatístico acima de 50, o que é baixo. Portanto, podemos dizer que a árvore filogenética obtida possui uma resolução ruim, não conferindo confiabilidade ao resultado e impedindo inferências sobre as relações entre as demais espécies. O baixo suporte estatístico pode ser explicado pelo uso do 16S como marcador molecular, sendo que recentemente foi demonstrado que esse marcador não é capaz de distinguir internamente os grupos Pneumoniae e Hominis (GUPTA et al., 2018b), o que fica evidente, principalmente, no grupo Hominis na árvore de ALVAREZ-PONCE et al..

Figura 3 – História evolutiva de 115 micoplasmas baseada no marcador molecular 16S.

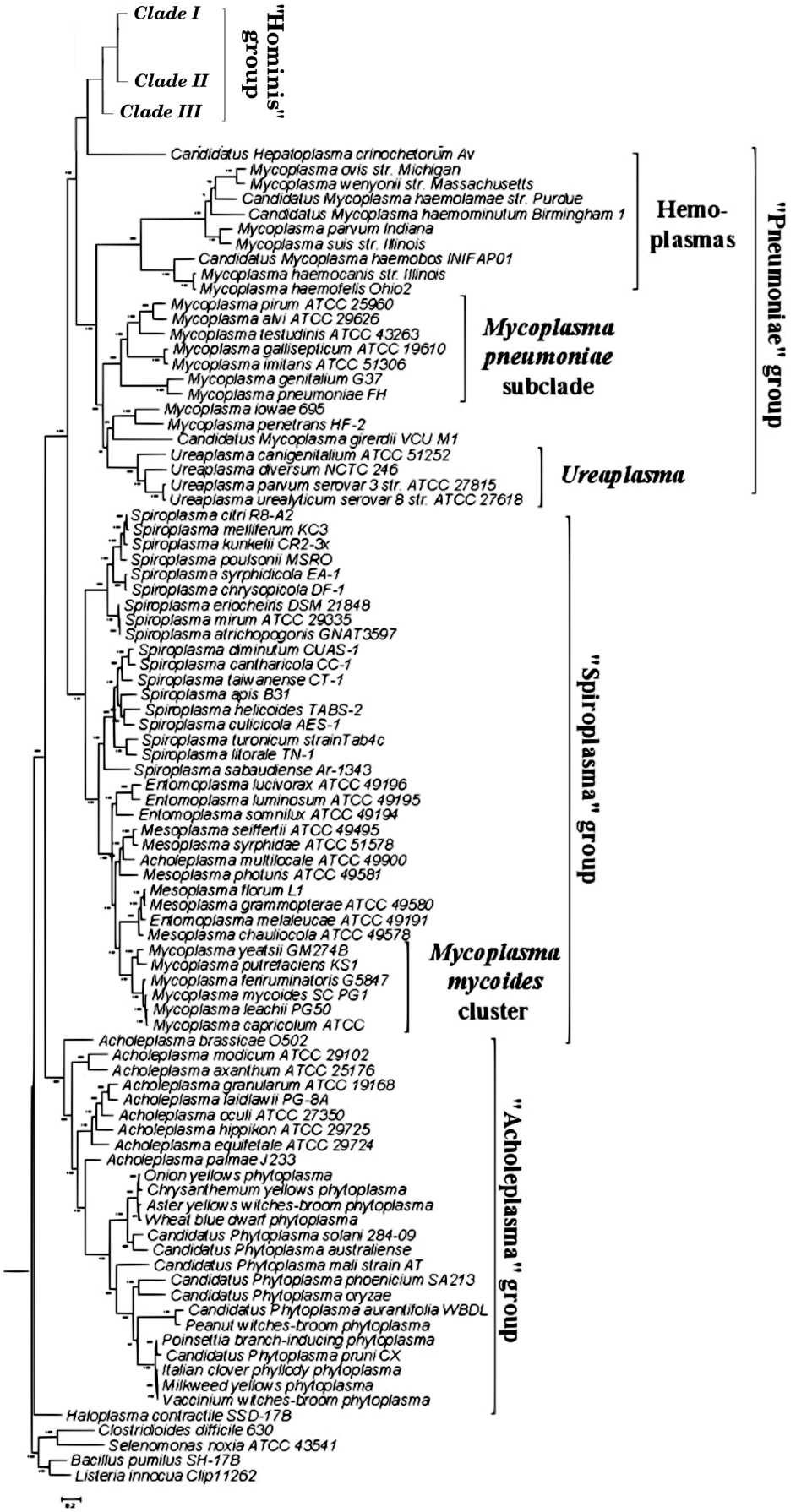


Com exceção do grupo externo (*Clostridium botulinum*) e as duas cepas de *M. agassizii*, as sequências de 112 micoplasmas foram obtidas a partir do banco de dados SILVA (QUAST et al., 2012). O alinhamento foi realizado com MUSCLE 3.8.31 (EDGAR, 2004), utilizando o software MEGA7 (KUMAR et al., 2016) para inferência por máxima verossimilhança com 1000 replicatas. As unidades taxonômicas em negrito estão depositadas em bancos públicos. Somente clados com *bootstrap* superior a 50 têm o suporte estatístico apresentado. Adaptado de Alvarez-Ponce et al., 2018.

GUPTA et al., em 2018, utilizaram perfis proteicos de marcadores moleculares de Tenericutes (WANG; WU, 2013) a fim de solucionar as relações descritas como conturbadas do filo (Figuras 4 e 5). Uma das árvores foi obtida com o uso de 63 proteínas conservadas, uma segunda árvore obtida a partir de 45 proteínas ribossomais, uma terceira utilizando as três principais subunidades da ARN Polimerase concatenadas e, por fim, utilizaram o marcador molecular 16S. Dentre outros resultados, os autores propuseram, a partir das árvores, uma reclassificação das espécies de micoplasmas sugerindo que a ordem *Mycoplasmatales* abrange também as espécies pertencentes ao gênero *Spiroplasma* e que uma nova ordem, *Mycoplasmoidales* ord. nov. fosse criada para abranger as outras espécies do gênero. Também propuseram que os grupos Hominis e Pneumoniae formam duas novas famílias. Essas investigações levaram em consideração aspectos moleculares robustos (mais de 100 deleções e inserções conservadas e 14 proteínas de assinatura conservadas) e a nova classificação ajuda a compreender aspectos biológicos e clínicos desses micro-organismos.

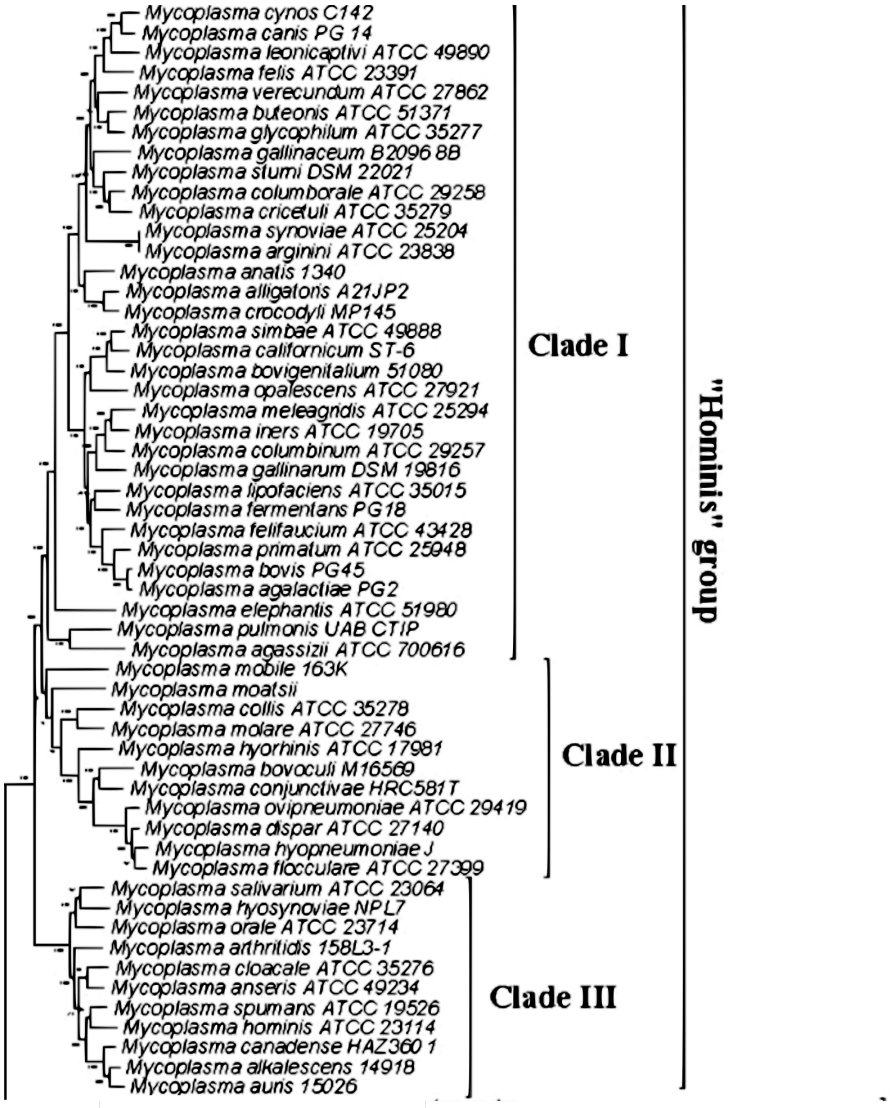
Esses novos estudos só foram possíveis em virtude da maior disponibilidade de genomas depositados nos bancos de dados biológicos públicos, enriquecendo o conhecimento sobre os aspectos evolutivos dos organismos. A compreensão dos aspectos funcionais e evolutivos relevantes na interação patógeno-hospedeiro, bem como a identificação do padrão evolutivo de espécies patogênicas e não patogênicas, processo esse que ainda não está muito delineado (apesar da clara importância da THG na patogenicidade), permanecem entre os objetivos mais intrigantes na biologia evolutiva.

Figura 4 – História evolutiva de micoplasmas do filo Tenericutes - Grupos Pneumoniae, Spiroplasma e Acholeplasma



Árvore filogenética do filo Tenericutes inferida por máxima verossimilhança com base em 63 proteínas conservadas do filo Firmicutes - PhyEco (WANG; WU). Árvores preliminares de máxima verossimilhança foram construídas usando FastTree2 (PRICE et al., 2010) e, subsequentemente, submetidas ao RAxML, com uso do modelo evolutivo LG (LE; GASCUEL, 2008). O suporte foi calculado por SH (GUINDON et al., 2010). Adaptado de Gupta et al., 2018.

Figura 5 – História evolutiva de micoplasmas do filo Tenericutes - Grupo Hominis



Árvore filogenética do filo Tenericutes inferida por máxima verossimilhança baseada em 63 proteínas conservadas marcadores do filo Firmicutes - PhyEco WANG; WU. Árvores preliminares de máxima verossimilhança foram construídas usando FastTree2 (PRICE et al., 2010) e, subsequentemente, submetidas ao RAxML com uso do modelo evolutivo LG (LE; GASCUEL, 2008). Suporte calculado por SH (GUINDON et al., 2010). Adaptado de Gupta et al., 2018.

1.2 MÉTODOS UTILIZADOS EM FILOGENÉTICA E FILOGENÔMICA

A maioria das aplicações na área de evolução se baseiam em métodos estatísticos para soluções relacionadas à reconstrução filogenética, inferência de modelos evolutivos e avaliação de alinhamentos. Os problemas estatísticos encontrados são, frequentemente, pouco padronizados, porque os modelos geralmente envolvem processos estocásticos ao longo das unidades taxonômicas operacionais (OTU, do inglês *Operational Taxonomic Unit*) de uma árvore (NIELSEN, 2006). Os métodos também atuam de forma a suprir a necessidade de suporte estatístico da filogenia, possibilitando o controle da qualidade, ajudando a estabelecer parâmetros e mitigando a incerteza da inferência. Nesses aspectos, foram desenvolvidas algumas técnicas de reamostragem como *bootstrap* (FELSENSTEIN, 1985), *jackknife* (EFRON, 1982) e a probabilidade posterior (HUELSENBECK et al., 2001), utilizadas em métodos de máxima verossimilhança e Inferência Bayesiana, respectivamente. Ademais, existem diversas estimativas para modelos evolutivos e relógios moleculares, os quais são melhor explorados na seção 1.2.2.

De modo geral, os métodos em análise filogenética podem ser divididos em duas categorias, a serem diferenciadas entre si de acordo com a abordagem utilizada em relação aos dados. Como classificado por Delsuc, os métodos são (1) baseados em sequências e (2) em características do genoma completo (DELSUC et al., 2005). O primeiro considera os genes e proteínas homólogas e ortólogas entre as sequências, utilizando-os para a construção do alinhamento múltiplo e possui duas abordagens: (i) construção da supermatriz (FELSENSTEIN; FELENSTEIN, 2004) e (ii) da superárvore (RAGAN, 1992). O segundo explora características intrínsecas aos genomas, buscando estabelecer comparativos entre os mecanismos por meio dos quais os genomas evoluem: conteúdo gênico, sintenia, presença de oligonucleotídeos e sua estrutura nos genomas. No presente estudo será utilizada a abordagem baseada em sequências utilizando o método de supermatriz.

1.2.1 **Reconstrução Filogenética**

A seguir será feita uma breve descrição dos métodos de reconstrução filogenética, considerando os quatro mais comumente utilizados pela comunidade científica. Inicialmente é importante destacar que a construção das árvores filogenéticas pode ser feita baseada em caracteres e com modelo de evolução explícito (como nos métodos probabilísticos de Máxima Verossimilhança, em inglês *Maximum Likelihood* [ML], e Inferência Bayesiana [IB]), não baseada em caracteres e com uso de modelo de evolução (métodos de Distância) ou baseada em caracteres mas não baseada em modelo de evolução (Máxima Parcimônia [MP]), classificação

representada na Figura 6, que tem origem do latim *percere*, de "poupar"ou "economizar".

O mais simples dos quatro métodos é o baseado em distância (FELSENSTEIN, 1984), que é capaz de analisar enormes quantidades de dados de maneira muito rápida para inferir a relação entre todos os táxons, considerando para o cálculo de distância os alinhamentos em pares de sequências. Esse método pode utilizar diferentes modelos evolutivos para o cálculo da divergência genética entre as seqências, tais como a "distância p", que considera o número de diferenças entre duas sequências em relação ao número total de sítios no alinhamento múltiplo.

Figura 6 – Classificação dos métodos de análise filogenética.

	Baseado em caracteres	Não baseado em caracteres
Modelo de evolução explícito	Máxima Verossimilhança Inferência Bayesiana	Métodos de Distância
Não baseado em modelo de evolução	Máxima Parcimônia	

Classificação dos principais métodos de inferência filogenética com base no uso caracteres e de modelos evolutivos.

Fonte: SALEMI et al., 2009.

No método de distância, o objetivo principal é construir uma árvore que represente as distâncias entre as sequências a partir da matriz calculada. Os principais algoritmos de clusterização que utilizam esse método são *unweighted-pair group method with arithmetic mean* (UPGMA) (SOKAL, 1958) e *neighbor-joining* (NJ) (SAITOU; NEI, 1987), caracterizados pela rapidez mesmo com um grande número de sequências (>50 ou centenas) e pela capacidade de recuperar árvores que se assemelham com as construídas com o uso de métodos baseados em caracteres (PEVSNER, 2015). Ademais, mostra-se eficiente para clusterização de dados de *microarray*, por exemplo (SALEMI et al., 2009).

No método de MP, o critério para encontrar a melhor árvore é analisar o comprimento dos ramos e considerando como "verdadeiros"os mais curtos, de forma a conservar as relações o mais simples possíveis seguindo o princípio da parcimônia. Assumindo que todos os sítios são independentes, a maior parte dos modelos estatísticos de substituição não são aplicáveis, por essa razão é um método não baseado em modelos evolutivos. Como consequência, o método de MP apresenta algumas limitações e desvantagens, como a subestimação da divergência evolutiva entre as sequências, não sendo, portanto, muito utilizado para reconstrução de árvores filogenéticas (PEVSNER, 2015). Exige ainda que exista uma taxa mutacional constante pela

extensão de todas as sequências, restringindo seu uso dependendo da heterogeneidade dos dados em análise (BROCCHIERI, 2001).

Os últimos dois métodos são probabilísticos e baseados na análise de máxima verossimilhança, frequentemente usada para estimar parâmetros de um modelo estatístico, dadas algumas obervações. É utilizada quando não há conhecimento sobre os parâmetros dos dados e visa encontrar a topologia de maior probabilidade que os represente, considerando o modelo. Assim, quando aplicamos o princípio de ML à filogenética, nossos dados são representados pelo alinhamento múltiplo e desconhecemos os parâmetros de topologia e tamanho de ramos da árvore filogenética, por exemplo. É necessária a determinação da distribuição, dentre todas aquelas definidas pelos possíveis valores de seus parâmetros, com maior possibilidade de ter gerado os dados. Ou seja, desejamos obter uma estimativa dos valores dos parâmetros desconhecidos e para resolver tal problema de estimação, precisamos detectar aqueles que maximizem a função de verossimilhança.

O método de ML (FELSENSTEIN, 1981) permite a aplicação de modelos evolutivos. Assim, são realizados cálculos probabilísticos para cada sítio, que levam a um produtório ($\prod$) que é, posteriormente, transformado em um somatório ($\sum$) por meio do uso de logaritmos. É utilizado o logaritmo natural da função de verossimilhança (*ln* L) porque maximizar o logaritmo natural de uma função é, em geral, mais simples e leva aos mesmos resultados da maximização da função original. O objetivo é identificar a árvore que melhor representa os dados em estudo e possui, portanto, o maior logaritmo natural. As vantagens do método de máxima verossimilhança são o uso de modelos evolutivos para sítios e ramos, trazendo melhores soluções no caso de análises com sequências de táxons divergentes entre si.

Análises realizadas com Inferência Bayesiana, por sua vez, utilizam o Teorema de Bayes (BAYES et al., 1763) para estimar a probabilidade posterior da árvore. Esse é um método interessante visto que para obtenção do suporte estatístico da topologia não são necessários cálculos adicionais, pois este é inferido a partir do próprio método, representando a probabilidade de cada clado representado na árvore (HUELSENBECK et al., 2001). A Inferência Bayesiana funciona a partir de *priors* (distribuições prováveis e pré-definidas) de parâmetros como topologia, comprimento dos ramos, variação entre os sítios e outros, que nos exemplos citados seguem as distribuições probabilísticas dos tipos uniforme, exponencial e gama, respectivamente (PEVSNER, 2015). Os *priors* estão diretamente relacionados ao conhecimento específico do pesquisador realizando a análise, sendo necessário respaldo na literatura para configurá-los ou optando-se por uma análise mais conservadora a partir do uso das distribuições de probabilidade

mais usuais. Esse cuidado é necessário porque a introdução de parâmetros pode levar a uma distribuição posterior enviesada. Um dos *softwares* mais conhecidos e utilizados que implementa essa técnica é o MrBayes (RONQUIST; HUELSENBECK, 2003), que utiliza Cadeias de Markov e o método de Monte Carlo (*Markov Chain Monte Carlo*), bem como o algoritmo *Metropolis Coupling*, em uma técnica conhecida como MCMCMC (HUELSENBECK, RONQUIST, 2001).

1.2.2 **Modelos Evolutivos**

Zuckerkandl e Pauling (ZUCKERKANDL; PAULING, 1962), em 1962, a partir da análise de hemoglobinas, propuseram que há uma taxa constante de alteração em todas as sequências codificantes de um mesmo gene em diferentes espécies. Puderam constatar, por meio de evidências fósseis, que a divergência da sequência é linear em relação ao tempo de diferenciação das duas espécies. A partir disso, foi estabelecido que há um relógio molecular variável entre as proteínas, permitindo estimar e calcular qual a divergência em relação ao tempo entre as sequências de organismos diferentes. Esse achado abriu campo para estudos filogenéticos capazes de estimar qual o ancestral comum entre as sequências, há quanto tempo elas divergiram, qual a taxa de mutação de cada sequência e, por fim, permitiu uma construção mais acurada da filogenia dos organismos.

Diversos modelos evolutivos foram desenvolvidos com o intuito de responder estas perguntas. O mais simples (JUKES et al., 1969) considera frequências iguais das bases e taxas mutacionais também iguais. Já no modelo de Kimura (KIMURA, 1980), há distinção entre as probabilidades de transversão e transição. Visando permitir a variação da frequência de bases nas sequências, Felsenstein (FELSENSTEIN, 1981) desenvolveu o modelo F81, logo seguido pelo modelo HKY85 (HASEGAWA et al., 1985), no qual todas as propriedades anteriormente citadas são variáveis. Para lidar com o conteúdo GC bem divergente entre genomas, o modelo T92 (TAMURA, 1992) estendeu o modelo de Kimura dois parâmetros (K2P) e tornou mais ajustadas as taxas de substituição. Dentre estes e muitos outros modelos (ZHARKIKH, 1994; KIMURA, 1981; POSADA, 2003), o mais complexo é o GTR (*general time-reversible*) (TAVARÉ, 1986), que considera taxas e frequências de bases desiguais. Adicionalmente, foram desenvolvidos modelos evolutivos para análises baseadas em aminoácidos, os quais tendem a ser mais conservadores.

Dada a diversidade de matrizes de substituição, é importante ressaltar que a escolha do modelo a ser utilizado pode ter grande impacto na análise filogenética. Alguns softwares como JModelTest (POSADA, 2008) e ProtTest (ABASCAL et al., 2005) ajudam a eliminar essa

insegurança relacionada à escolha, pois realizam testes estatísticos comparativos dos modelos indicando qual é o mais adequado para os dados em análise. Dependendo do caso em estudo, é interessante também o desenvolvimento de modelos específicos, que podem ainda minimizar problemas de viés composicional (STEEL et al., 1993) e de THG (LAKE; RIVERA, 2004). Em contrapartida, escolher um modelo que não represente as características dos dados pode levar a resultados de difícil interpretação, como árvores com topologias equivocadas mas que apresentam excelente suporte estatístico (DELSUC et al., 2005).

Os modelos evolutivos desempenham papel fundamental no que tange à representação do padrão evolutivo do gene. As análises assumem que o modelo escolhido é verdadeiro e que é, portanto, o que melhor representa os dados, tornando o resultado passível de erro se violado esse princípio, já que os modelos falham em compreender o enredo do processo evolutivo (DELSUC et al., 2005). Isso se dá, principalmente, porque os modelos são construídos de forma a minimizar a complexidade e assumem, por exemplo, que os sítios possuem independência entre si, mas alguns casos demonstram não seguir esse comportamento. Em proteínas, as interações físico-químicas entre sítios vizinhos ou a estrutura proteica interferem em como os outros sítios são modificados (ROBINSON et al., 2003; CHOI et al., 2007), necessitando modelos dependentes do contexto (MORRISON, 2013).

O uso de diversos modelos evolutivos associados a análises particionadas, como possibilitado pelo MrBayes, conservam a heterogeneidade das sequências e respeitam o processo natural de diferentes pressões evolutivas sobre diferentes porções do genoma, contribuindo para aumentar a confiabilidade ao aproximar a simulação dos eventos reais. Essa abordagem se mostra muito relevante em análises filogenômicas, já que os ortólogos estão sob diferentes taxas evolutivas e possuem características heterogêneas entre si.

Em conjuntos de dados muito grandes, cenário comum em filogenômica, etapas como a de particionamento dos dados e configuração de modelos evolutivos dependem majoritariamente de *scripts* desenvolvidos pelo grupo de pesquisa, que utiliza alguma linguagem de programação com bibliotecas para tais ações. Como descrito por LEIPZIG, o processo de transformação dos dados em bioinformática é usualmente constituído de muitos passos que estão, geralmente, poucos integrados, e que quando colocados em forma de *pipelines*, carecem de recursos de paralelização, de conferência das dependências e de compatibilidade entre elas, de ajuste de parâmetros por parte do usuário, consistência nos formatos de arquivos e rastreamento do progresso da análise. Já existem iniciativas que visam mitigar esses problemas (AMSTUTZ et al., 2016), principalmente por meio de *pipelines* utilizáveis em diferentes plataformas, tornando-os escaláveis, modificáveis

e, assim, contribuindo para o desenvolvimento científico e possibilitando o uso por diferentes grupos de pesquisa em todo mundo.

1.3 APLICAÇÕES EM FILOGENÔMICA

O emprego da ciência da computação no processamento de dados biológicos é fundamentalmente norteado por questões biológicas, bioquímicas, funcionais, evolutivas, entre outras. A diversidade de soluções, portanto, é essencial e gera ferramentas em igual proporção para responder eficientemente cada uma dessas perguntas. Aplicações na filogenômica, por exemplo, compreendem, no mínimo, três áreas: predição de função gênica (BROWN; SJÖLANDER, 2006), compreensão de eventos de THG (WHITAKER et al., 2009) e inferência de relações evolutivas.

Apesar de todos os programas se basearem em modelos e testes estatísticos para análise filogenética, são construídos com diferentes abordagens para o processamento dos dados. Algumas ferramentas de filogenômica, discutidas abaixo, recebem ênfase na interface gráfica, na robustez e qualidade dos dados. Algumas outras não exploradas aqui (PIEL; VOS, 2018; DUNN et al., 2013; ROBBERTSE et al., 2011; PETERS et al., 2011; STAJICH et al., 2002; JUNIER; ZDOBNOV, 2010), contribuem para a grande diversidade de ferramentas disponíveis para reconstrução filogenética.

Os primeiros *softwares* de evolução molecular que se popularizaram devido a sua usabilidade facilitada foram o PAUP (SWOFFORD, 1985), MacClade (MADDISON, 2008), MEGA (KUMAR et al., 2012) e TREECON (PEER; WACHTER, 1993), os quais implementam métodos de máxima parcimônia, máxima verossimilhança e diferentes algoritmos de reconstrução de árvores. Possuem, entretanto, limitações quanto à escalabilidade. Com o aprimoramento desses algoritmos, é possível um rápido processamento e escalabilidade para análise de dados produzidos em massa, levando ao desenvolvimento de ferramentas mais robustas.

Em 2004, a linguagem de programação e também ambiente de desenvolvimento R (GENTLEMAN, 2008) recebeu a biblioteca APE (PARADIS et al., 2004). Em Bioinformática, o R é amplamente utilizado na análise de expressão gênica e tem como objetivo geral realizar análises estatísticas, possuindo, portanto, uma vantagem para a reconstrução filogenética, que é fundamentalmente baseada em métodos estatísticos. Isso torna necessário apenas o desenvolvimento de bibliotecas para evolução molecular, como a Analysis of Phylogenetics and Evolution (APE).

Desenvolvido em 2008, o programa Phyutility (SMITH; DUNN, 2008), baseado em duas bibliotecas Java e disponibilizado para uso por linha de comando, oferece, além das

funções padrões de análise filogenética, opções interessantes e que não estão presentes em outras aplicações, como a busca e *download* de sequências diretamente do NCBI. Esse programa possui uma abordagem mais integrativa, justificada na adoção do uso para filoinformática (CRACRAFT, 2002), e possibilita a análise de diversidade, filogenia, biogeografia e classificação. A ferramenta Dendropy (SUKUMARAN; HOLDER, 2010), para python (ROSSUM et al., 2007), também busca essa abordagem, oferecendo mais opções de simulação para genética de populações.

A vantagem do crescimento dos estudos em filogenética é o acúmulo de conhecimento no meio acadêmico. A fim de explorar isso, o *workframe* Mesquite foi construído visando a incorporação de módulos desenvolvidos pela comunidade, contribuindo para a integridade das análises e estabelecendo um *software* flexível (mas não exaustivo em métodos).

Em razão da necessidade de automatização, aplicações que podem ser inseridas no fluxo de *pipelines* recebem destaque, como é o caso da Phyx (BROWN et al., 2017), projetada para uso no terminal linux. Essa aplicação permite a simulação de dados por meio de um compilado de ferramentas para diversos fins, tais como a filtragem de alinhamentos, edição de nomes e remoção de sítios, e possui, pelo menos, 14 funcionalidades que visam, principalmente, o processamento de grande quantidade de dados.

A fim de solucionar esse mesmo problema, o *framework* ETE (HUERTA-CEPAS et al., 2010) é considerado um dos mais completos, pois realiza a maior parte das análises filogenéticas e filogenômicas e permite que o usuário configure seu próprio *pipeline* com mais de 12 opções de ferramentas, tais como, de análise de modelo evolutivo, construção de árvores e análise de seleção positiva, gerando, inclusive, imagens resultantes dos processos de análise. ETE3 auxilia também na análise de menor granularidade dos dados em virtude de alternativas de refinamento personalizado do alinhamento utilizando o trimAl (CAPELLA-GUTIÉRREZ et al., 2009), com ajustes dos parâmetros.

Já em python, somente em 2016 foi disponibilizada a primeira biblioteca projetada para análise de evolução molecular. Mais completa do que a Bio.Phylo (TALEVICH et al., 2012), a MEvoLib (ÁLVAREZ-JARRETA; RUIZ-PESINI, 2016) engloba 20 modelos evolutivos e pode ser incorporada em *pipelines* de reconstrução de forma mais eficiente, teoricamente tornando desnecessário o uso de outras ferramentas. Entretanto, é importante observar que as aplicações em R e python ainda carecem de métodos para preparação como *trimming*, seleção e controle de qualidade, por meio de estatísticas, na etapa de alinhamento, por exemplo.

O alinhamento, por sua vez, é considerado o gargalo da filogenômica, sendo frequentemente alvo de estudos de otimização de algoritmos, que resultam em ferramentas diversas

de alinhamento. Além disso, considerando o uso de genomas de bactérias (ricos em regiões de recombinação e elementos móveis) ou o uso de genomas maiores, como os de eucariotos, a complexidade do alinhamento aumenta exponencialmente. Recentemente, o uso do *core genome* em estudos evolutivos se mostrou uma forma de contornar e desconsiderar essa plasticidade genômica (TALEVICH et al., 2012), tornando o processo mais escalável.

Visando esclarecer a capacidade de obtenção de informação a partir do *core genome*, Tsang (TSANG et al., 2017) compararam árvores filogenéticas de 10 espécies de bactérias construídas baseadas em SNP, utilizando o programa Parsnp (TREANGEN et al., 2014), e construídas a partir do sequenciamento completo do genoma (WGS, do inglês *Whole Genome Sequencing*). Os autores comprovaram que a análise por SNP reproduziu de maneira satisfatória o mesmo resultado das árvores utilizando todo o genoma ou aquelas obtidas por meio de *whole genome Multilocus Sequence Typing* (wgMLST) (KATZ et al., 2017). Entretanto, ressaltaram que o uso de SNPs para reconstrução filogenética é um método alternativo e deve ser utilizado apenas quando não há recursos computacionais suficientes. Outra desvantagem da análise baseada em SNPs é a necessidade de utilização de genomas completos e intraespecíficos, limitando o escopo da questão a ser respondida.

1.3.1 **Problemáticas**

A maioria dos pipelines são idealizados no meio acadêmico, já que seu desenvolvimento necessita de denso conteúdo teórico e há ainda escassa aplicação comercial da genômica evolutiva. O desenvolvimento de *scripts* próprios do grupo de pesquisa, por sua vez, permite o rastreamento completo de transformação da informação, preservando a integridade dos dados, evitando a inserção de ruído e garantindo controle da qualidade. Essa prática se justifica dado que, apesar de haver uma vasta disponibilidade de ferramentas e *pipelines*, perguntas biológicas que dependem da busca e identificação de ortólogos em linhagens muito antigas e trabalhos com organismos não modelo (GRANT; KATZ, 2014), entre outros exemplos, exigem o desenvolvimento de módulos específicos que melhor lidem com o conjunto de dados (SMITH et al., 2009). Entretanto, esses *scripts*, chamados *inhouse*, muitas vezes apresentam redundância de código, escassez de testes (DARRIBA et al., 2018), pouca documentação do fluxo da informação e ausência de controle de versionamento (LEIPZIG, 2017) e carecem, principalmente, de automatização.

Ferramentas de aprimoramento da análise filogenética, como GUIDANCE (SELA et al., 2015) e trimAl (CAPELLA-GUTIÉRREZ et al., 2009), que atuam sobre o alinhamento e diminuem o ruído filogenético, são importantes para o controle da qualidade e resultam em

árvores filogenéticas com melhor resolução (ROKAS et al., 2003). Entretanto, tais aplicações de inspeção complicam o processo de automatização, pois o critério de decisão sobre os parâmetros, por exemplo, é arbitrário e, até então, não replicável para qualquer conjunto de sequências em estudo. Esses fatos justificam a incompleta automatização e microgerenciamento, em nível de processo, exigidos pela análise filogenética e endossam, portanto, a importância da busca por padrões que possam ser aplicados em questões evolutivas de diversos gêneros. Isso inclui também o desenvolvimento de ferramentas adaptáveis a parâmetros diferentes, correspondendo às necessidades dos dados e da pergunta biológica norteadora.

Com base nas referências aqui exemplificadas, propõe-se o uso de ferramentas de análise filogenômica para a construção de um *pipeline* a ser aplicado no gênero *Mycoplasma* como estudo piloto, a fim de identificar o perfil evolutivo de 89 genomas representantes de cada uma das espécies depositadas no NCBI, primando pela automatização, controle de qualidade e que levando em consideração aspectos evolutivos para a escolha das ferramentas a serem integradas.

2 OBJETIVO

Análise de aspectos evolutivos dos genomas de bactérias do gênero *Mycoplasma* por meio de métodos teórico-computacionais e desenvolvimento de um *pipeline* de análise filogenômica focado na análise de genomas bacterianos, tendo como estudo de caso as bactérias do gênero *Mycoplasma*.

2.1 OBJETIVOS ESPECÍFICOS

1. Análise Filogenômica de bactérias do gênero *Mycoplasma*.
2. Desenvolvimento de um *pipeline* focado na análise evolutiva de genomas bacterianos, a partir de alguns *scripts* já desenvolvidos pelo grupo de pesquisa coordenado pela Profa. Dra. Claudia E. Thompson.
3. Desenvolvimento do *pipeline* de acordo com boas práticas de programação.
4. Aplicação de *cut-offs* descritos na literatura para automatização do processo de análise sem perda de informação biologicamente relevante.
5. Validação e comparação dos resultados do *pipeline* com estudos recentes publicados na literatura.

3 ARTIGO CIENTÍFICO

O artigo científico resultante desse Trabalho de Conclusão de Curso será submetido para a revista *eLIFE* (Fator de Impacto 2017 = 7,616) sob o título de

Análise Filogenômica de Micoplasmas: estudo de caso e desenvolvimento de *pipeline*.

Autores

Meiski Mariá Vedovatto (1)

Fábio Carrer Andreis (3)

Claudia Elizabeth Thompson (1,2,3,4)

Afiliação

1. Informática Biomédica, Universidade Federal de Ciências da Saúde de Porto Alegre
2. Departamento de Farmacociências, Universidade Federal de Ciências da Saúde de Porto Alegre
3. Unidade de Biologia Teórica e Computacional, Centro de Biotecnologia, Universidade Federal do Rio Grande do Sul
4. Programa de Pós-Graduação em Ciências da Saúde, Universidade Federal de Ciências da Saúde de Porto Alegre

E-mail dos autores:

meiski@ufcspa.edu.br;

fabio.andreis@ufrgs.br;

cthompson@ufcspa.edu.br;

Análise Filogenômica de Micoplasmas: estudo de caso e desenvolvimento de *pipeline*

Meiski Mariá Vedovatto[1*], **Fábio Carrer Andreis**[3†*], **Claudia Elizabeth Thompson**[1,2,3,4†*]

***For correspondence:**
meiski@ufcspa.edu.br (1);
cthompson@ufcspa.edu.br (2);
fabio.andreis@ufrgs.br; (3)

[†]Todos os autores contribuíram igualmente para esse trabalho

[1]Informática Biomédica, Universidade Federal de Ciências da Saúde de Porto Alegre; [2]Departamento de Farmacociências, Universidade Federal de Ciências da Saúde de Porto Alegre; [3]Unidade de Biologia Teórica e Computacional, Centro de Biotecnologia, Universidade Federal do Rio Grande do Sul; [4]Programa de Pós-Graduação em Ciências da Saúde, Universidade Federal de Ciências da Saúde de Porto Alegre

Abstract *The study of genomics using phylogenetic tools has increased due to the availability of several sequences in databases, resulting from the development and application of new technologies of DNA sequencing. Phylogenomic approaches allow a better comprehension of the evolutionary processes occurring in the genomes and the understanding of the evolutionary history of genomes belonging to different species. Additionally, computational studies lead to answers to different biological questions related to the molecular evolution of genomes and their genes and corresponding proteins. Bacteria of the Mycoplasma genus have small genomes and are highly dependent on the nutrients supplied by the host. They show different lifestyles, with most of them being parasites and responsible for diseases in humans, other animals and even plants. An important evolutionary hypothesis about this group is that mycoplasmas underwent a process of degenerative evolution resulting in the loss of the cell wall. In 2013, a study identified the evolutionary relationships among 31 species of Mycoplasma using a phylogenomic approach. Nowadays, more genomes are available in databanks and we have performed new analyses including all these to obtain a more comprehensive understanding of the evolutionary aspects related to these bacteria. The main objective of this work was the development of a pipeline of phylogenomic analysis and the genomic characterization of the Mycoplasma genus, including 83 species and 89 genomes, and using tools such as Proteinortho, GUIDANCE, ProtTest, PhyML and MrBayes. Our pipeline represents an important step to automatize processes and analyses related to evolutionary genomics of bacterial species.*

Introdução

Com o rápido desenvolvimento de novas tecnologias e a diminuição dos custos para o sequenciamento de genomas, tem crescido vertiginosamente a quantidade de genomas completos disponíveis em bancos de dados públicos (***Benson et al., 2015***). Isso permite com que possamos analisar comparativa e evolutivamente um grande número de genomas, possibilitando a formulação de novas hipóteses de pesquisa e obtenção de respostas a perguntas biológicas de forma mais eficiente. Estudos sobre a diversidade da vida, o conteúdo, estrutura e função de genes e sua variabilidade têm incluído cada vez mais organismos, permitindo inferências evolutivas mais abrangentes e confiáveis. Em relação aos procariotos, houve um aumento significativo na disponibilidade de genomas. Para o gênero *Mycoplasma*, por exemplo, havia 13 depósitos em 2007

(***Sirand-Pugnet et al., 2007***) e, atualmente, identificamos 84 espécies com genomas sequenciados, isso sem considerar aqueles disponíveis para diferentes cepas de cada espécie. Isso possibilita um estudo evolutivo bem mais abrangente desse gênero, que é necessário tendo em vista que seus pequenos genomas e a falta de características únicas dificultam o entendimento da relação entre as diferentes espécies de micoplasmas e os gêneros relacionados (***Gupta et al., 2018***).

Micoplasmas, pertencentes à classe *Mollicutes* e ordem *Mycoplasmatales*, são organismos parasitas obrigatórios que não apresentam parede celular e são encontrados em mamíferos, aves, répteis e peixes, mas estão ausentes em anfíbios (***Razin et al., 1998***). Nos mamíferos, apresentam patogenicidade em humanos, bovinos, ovinos, caprinos e outros, sendo considerados agentes etiológicos de doenças respiratórias, artrite, agalaxia e pleuropneumonia, respectivamente (***May et al., 2014***). A presença de espécies de *Mycoplasma* em pacientes imunocomprometidos, como portadores de HIV (***Mavedzenge and Weiss, 2009***; ***Wang et al., 1992***; ***Hannan, 1998***) e em células tumorais (***Barykova et al., 2011***; ***Xie et al., 2017***), caracteriza o perfil oportunista do gênero, que abrange 160 espécies (***May et al., 2014***).

Essas bactérias são consideradas os menores organismos capazes de autorreplicação, possuindo dependência de nutrientes por parte do hospedeiro devido à incapacidade de sintetizar aminoácidos (***Himmelreich et al., 1996***). A ausência de regiões para essa síntese se deve ao tamanho reduzido dos seus genomas, que varia entre 0,58 Mb e 2 Mb (***Citti et al., 2018***). Essa propriedade única das micoplasmas teria surgido por meio do processo de evolução degenerativa (***Razin et al., 1998***) e, adicionalmente, por uma série de ciclos que resultaram na diminuição de seu genoma e perda de parede celular. ***Maniloff*** (***1996***) estimou que a classe *Mollicutes* compartilha um ancestral comum com bactérias gram-positivas, do ramo de *Streptoccocus*, e divergiu há 600 milhões de anos. Essa divergência resultou em dois ramos principais, os quais se diferenciaram há 400 milhões de anos. O primeiro (ramo AAA) é constituído por 3 gêneros, *Asteroleplasma*, *Anaeroplasma* e *Acholeplasma*, e o segundo (SEM) é constituído por *Spiroplasma*, *Entomoplasma* e *Mycoplasma*.

O gênero *Mycoplasma* é dividido em três grupos internos: Hominis, Pneumoniae e Mycoides (***Brown, 2010***), sendo o último mais proximamente relacionado a *Mesoplasma* e *Entemoplasma*. O grupo *mycoides* possui uma história evolutiva diferente dos outros dois clados, pois seriam originários de um ancestral associado a insetos e se tornaram fenotipicamente semelhantes a outras linhagens de *Mycoplasma* por meio de eventos independentes de evolução convergente envolvendo transferência horizontal de genes (THG) (***Lo et al., 2018***). Adicionalmente, ***Citti et al.*** (***2018***) apontam que esses eventos de THG desempenham papel fundamental na aquisição de resistência nessas bactérias. A presença de repetições nesses genomas também é um resquício evolutivo e ***Rocha and Blanchard*** (***2002***) ressaltam a importância dessas regiões, já que conferem plasticidade fenotípica e foram preservadas em micro-organismos tão pequenos, conservando a variabilidade gênica mesmo tendo passado por um processo de simplificação genômica.

Estudos recentes sobre a evolução de micoplasmas têm utilizado marcadores moleculares, como o 16S, para melhor entender a classificação e relacionamento evolutivo das espécies já conhecidas e as que têm sido identificadas (***Alvarez-Ponce et al., 2018***; ***Kamminga et al., 2017***). Outros estudos visam solucionar as relações dentro da classe *Mollicutes* (***Citti et al., 2018***) e também em nível da família *Mycoplasmataceae* (***May et al., 2014***). Importante ressaltar que nos últimos anos, análises filogenômicas têm ganhado relevância na solução de problemas relacionados à resolução da árvore das espécies (***Eisen and Fraser, 2003***), já que a disponibilidade de dados genômicos permite a identificação de sequências ortólogas entre diferentes espécies com base no conteúdo de seus genomas, o que tende à redução das incongruências na topologia e ao aumento do suporte estatístico (***Rokas et al., 2003***). Tais estudos são importantes para o esclarecimento da relação entre espécies com potencial patogênico e aquelas não patogênicas, bem como para o entendimento dos processos que levam à patogenicidade, permitindo uma melhor compreensão de aspectos evolutivos, adaptativos e das relações dos patógenos com seus respectivos hospedeiros.

Alguns autores, utilizando uma abordagem filogenômica, conseguiram inferir a taxonomia de novos organismos em relação aos *Mollicutes*. ***Leclercq et al.*** (***2014***) utilizou 127

ortólogos de 48 genomas, sendo 35 de micoplasmas, para identificar o relacionamento evolutivo desses micro-organismos com o genoma recém sequenciado de *Candidatus Hepatoplasma*, classificando-o como grupo proximamente relacionado do clado Hominis. Já **Guimaraes et al. (2014)**, analisou 54 genomas, sendo 33 de micoplasmas, identificando um total de 32 ortólogos. Nesse estudo se verificou que a diferenciação entre gêneros carece de métricas classificatórias, tendo em vista o grande impacto gerado por eventos de recombinação entre classes e gêneros, que não são tão facilmente identificáveis. Esses estudos, entretanto, não analisaram toda a diversidade de *Mycoplasma* agora disponível.

O mais recente e completo estudo evolutivo de micoplasmas foi desenvolvido por **Gupta et al. (2018)** e utilizou 140 genomas, sendo 80 desses pertencentes à *Mycoplasma*. A árvore foi inferida com base em 63 proteínas identificadas a partir do perfil HMM (*Hidden Markov Model*), construído por **Wang and Wu (2013)** visando a obtenção de marcadores para o filo Tenericutes. Entretanto, é importante ressaltar que mesmo com os resultados obtidos nos estudos citados até aqui, nenhum deles contempla todas as peculiaridades das micoplasmas em nível de gênero, demostrando ser interessante o desenvolvimento de métodos alternativos que possam analisar sua evolução de forma mais acurada e sistemática e que busque identificar as características intrínsecas ao gênero *Mycoplasma* por meio de marcadores que sejam mais representativos.

A grande maioria dos estudos publicados e disponíveis na literatura na área de filogenômica é desenvolvida com base em *scripts* de algum grupo de pesquisa utilizando alguma ferramenta que automatiza parcialmente o processo, como bibliotecas em python (**Talevich et al., 2012**; **Álvarez-Jarreta and Ruiz-Pesini, 2016**). Considerando isso, foi desenvolvido um *pipeline* visando o estudo evolutivo de genomas de bactérias, tendo como estudo de caso micoplasmas, primando pela automatização das análises e considerando aspectos evolutivos na escolha das ferramentas a serem integradas a fim de apresentar o resultado mais fidedigno possível. Em nível de automação, a maioria das ferramentas hoje disponíveis não integram passos fundamentais da reconstrução filogenética, como a etapa de alinhamento e seu controle (filtro). Dessa forma, um dos objetivos foi a busca e estabelecimento de parâmetros que permitam a transformação dos da-

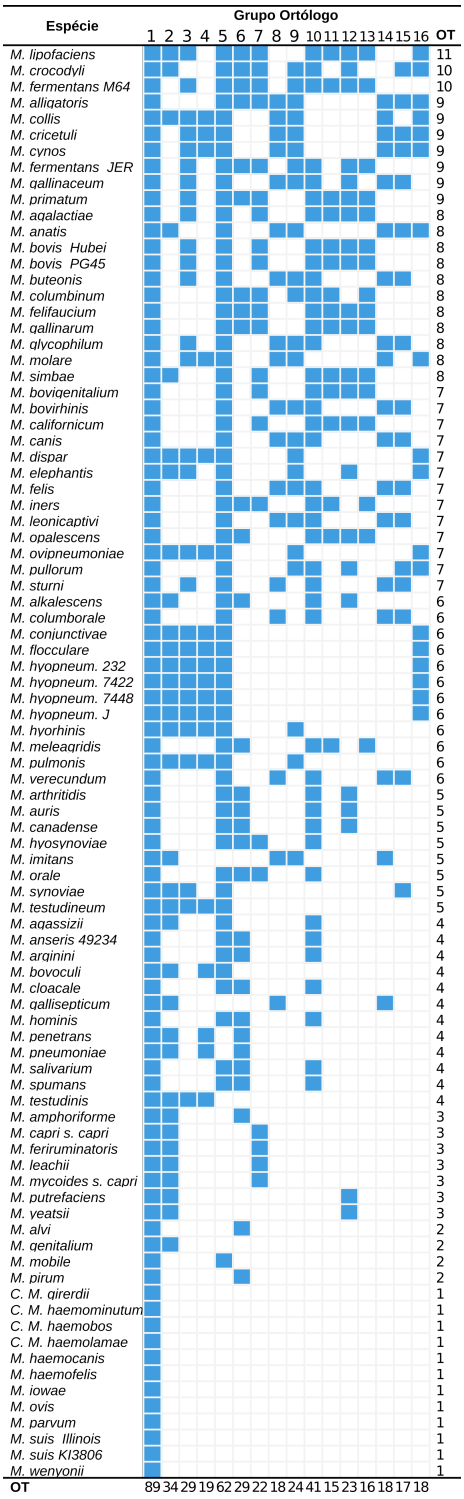


Figure 1. Distribuição dos GO em cada genoma. É possível observar as espécies representadas por somente 1 GO. Coluna e linha OT são as ocorrências totais dos *clusters* por espécies e vice-versa.

dos de forma ininterrupta, utilizando genomas de micoplasmas como estudo piloto.

Resultados e Discussão

Seleção de Genomas e Identificação de Ortólogos

Nesse estudo, foram incluídos 89 genomas de micoplasmas, sendo 83 pertencentes a espécies diferentes, com três cepas adicionais de *M. hyopneumoniae* e uma adicional de cada uma das espécies *M. fermentans*, *M. suis* e *M. bovis*, disponíveis no *National Center for Biotechnology Information* (NCBI). Na seção "Genome" do NCBI, buscou-se pela palavra-chave "*Mycoplasma*". No Material Suplementar se encontram listados todos os genomas incluídos no estudo, com informação referente à espécie e cepa a qual pertence, o número de acesso (ID), o *status* do genoma (se completo, *scaffold*, etc), seu tamanho, % GC, número de genes e proteínas e informações sobre patogenicidade.

A identificação dos ortólogos foi realizada por meio do *software* Proteinortho (***Lechner et al., 2011***), tendo como critérios de identidade e cobertura o valor de 60%. Essa análise de genes ortólogos resultou na identificação de 6.292 *clusters*, dos quais somente 16 foram identificados em, no mínimo, 33% dos genomas. Os *clusters* identificados, o número de espécies onde se encontram e suas respectivas funções estão descritos na Tabela 1. Na Figura 1, estão representados os 16 grupos de ortólogos (GO), numerados de 1 a 16, a identificação das espécies onde são encontrados, bem como o número total de genomas para cada conjunto de ortólogos. É possível observar as espécies representadas por somente um GO. Se utilizássemos presença em, no mínimo, 50% das espécies como critério, a inferência seria baseada em somente dois *clusters*, referentes ao fator de alongamento Tu e à proteína ribossomal 50S L16 (*clusters* 1 e 2, Tabela 1). Esse resultado reduzido de número de ortólogos se deve aos parâmetros de identidade e cobertura utilizados, 60% em ambos os casos, que se mostraram muito restritivos. Nessa Tabela 1, também é possível observar que não houve representatividade de todos os *clusters* em todas as OTUs. De fato, nenhuma espécie foi representada por todos os *clusters*, sendo *M. lipofaciens* a espécie com maior número de genes ortólogos, um total de 11.

A partir dos arquivos gerados pelo Proteinortho, foi desenvolvido um *pipeline* para automatização das demais etapas de análise evolutiva, a ser descrito na próxima seção. Importante ressaltar que o único *cluster* presente em todos os genomas foi o 1, que é do gene *tuf*, já descrito na literatura como sendo um bom marcador molecular (***Nichio et al., 2017***) e que possui função como fator de alongamento no processo de síntese proteica (***Zaha et al., 2014***). Nessa análise posterior o relacionamento evolutivo de 12 espécies foi inferido baseado em somente um gene, o *tuf*. Além disso, as relações entre quatro micoplasmas (quatro genomas) foram baseadas em somente um gene além do *tuf*, são eles: *Otc* para *M. alvi* e *M. pirum*, *nrdF* para *M. genitalium* e *rplP* para *M. mobile*, que possuem as funções de catalisador da reação de formação de citrulina e fosfato, de catalisar a formação de deoxirribonucleotídeos e, por fim, de ligar-se ao RNA ribossômico 23S atuando de forma essencial na subunidade de transcrição, respectivamente.

Finalmente, houve outros sete genomas cuja história evolutiva resultante da análise por meio de nosso *pipeline* foi baseada em 2 GOs além do *tuf*: *Otc* para *M. amphoriforme*, a família de proteínas transportadoras que facilitam a entrada e saída de substâncias pela membrana para *M. capricolum subsp capricolum, M. feriruminatoris, M. leachii* e *M. mycoides subsp capri*, e proteínas da família álcool desidrogenase para *M. putrefaciens* e *M. yeatsii*. Em todas as últimas 7 espécies citadas, também foi utilizado o GO *nrdf*.

Pipeline

No *pipeline* foram integrados *scripts* desenvolvidos pelo nosso grupo de pesquisa que eram executados individualmente e utilizam diversas ferramentas externas, visando a automatização. Este quesito tem grande peso na aplicação, pois a análise filogenômica comumente depende de acompanhamento e conferência manual dos resultados.

Table 1. Caracterização dos *clusters* encontrados com a ferramenta Proteinortho que foram, posteriormente, utilizados para inferir a história evolutiva de *Mycoplasma*

Cluster	Função	N° de espécies
1	Fator de alongamento Tu	89
2	Classe 1b subunidade beta ribonucleosídeo-difosfato redutase	34
3	Arildialquilfosfatase	29
4	Subunidade do transportador de ascorbato PTS IIB	19
5	Proteína ribossomal 50S L16	62
6	Ornitina transcarbamilase	29
7	Permease de transporte ABC	22
8	Permease de transporte ABC	18
9	Transportadores de cassetes de ligação de ATP	24
10	Fosfoquetolase	41
11	ATPase do tipo P com translação de magnésio	15
12	Álcool desidrogenase dependente de zinco	23
13	Hidrolase da família *Cof-type* HAD-IIB	16
14	Transportadores de cassetes de ligação de ATP	18
15	Meteniltetrahidrofolato ciclohidrolase	17
16	Subunidade do transportador de galactitol PTS IIB	18

Na tentativa de mitigar o micro-gerenciamento, processos decisórios em etapas fundamentais, como o alinhamento, foram programadas para ocorrerem de modo transparente para o usuário baseando-se em *cut-offs* de filtros estabelecidos na literatura, como aqueles testados e avaliados por ***Tan et al.*** (***2015***) em relação aos escores a serem utilizados pelo GUIDANCE (***Sela et al., 2015***), que é um *software* que realiza o controle de qualidade dos alinhamentos. Essa abordagem de implementação reflete diretamente na qualidade da árvore filogenética ou filogenômica, como no nosso estudo de caso.

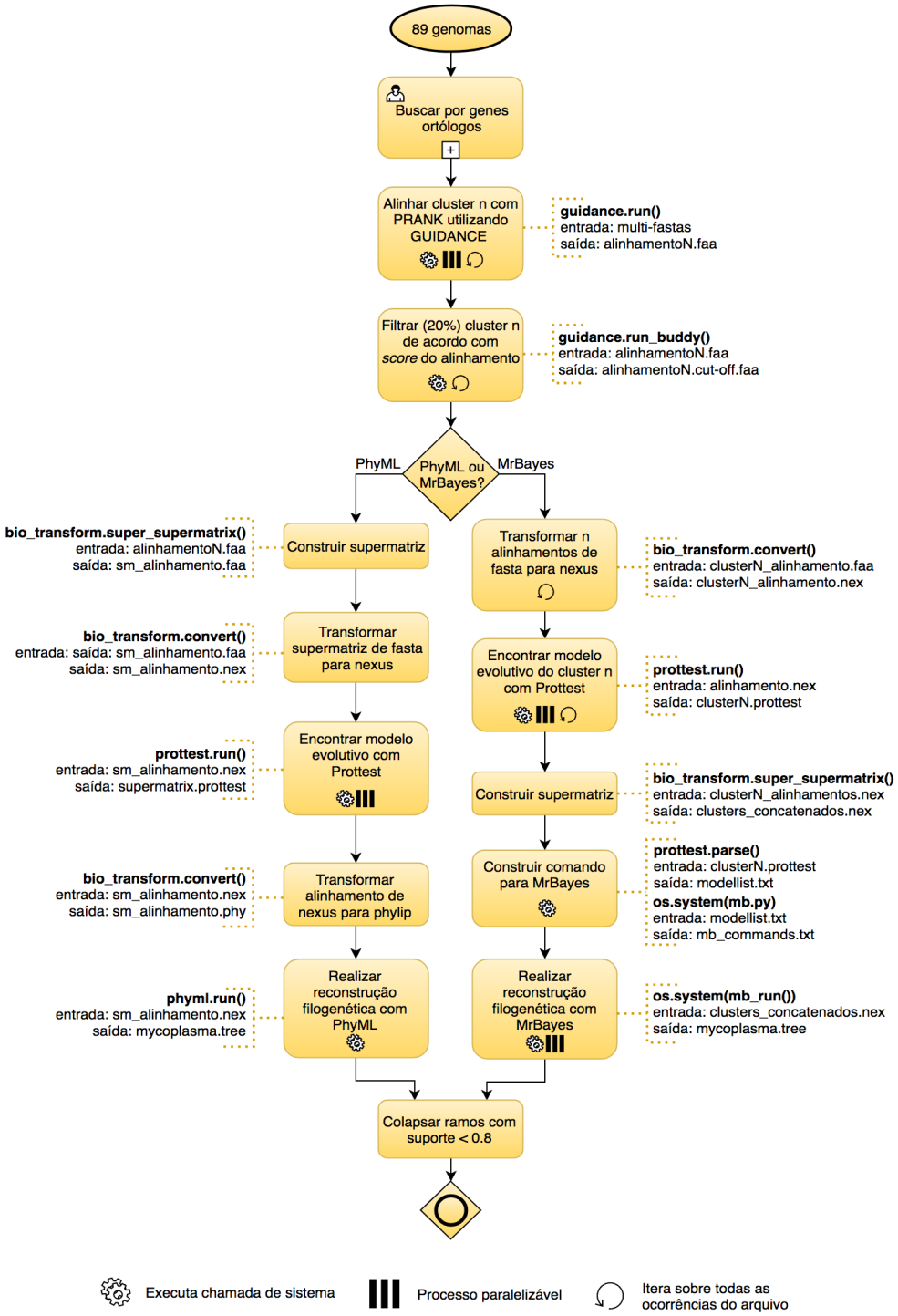


Figure 2. Fluxograma dos passos implementados no *pipeline*. Esse esquema visa exemplificar como ocorre o fluxo de dados abstraindo detalhes da busca por genes ortólogos, mas partindo do princípio de que o usuário já possui multi-fastas com ocorrências únicas das sequências. O pseudocódigo está disponível no Apêndice 1.

Na Figura 2 está representado o fluxo de transformação dos dados que foi implementado no *pipeline*. Partindo de um controle principal da aplicação, cujo pseudocódigo está disponível no material suplementar, as ferramentas são ordenadas e suas respectivas entradas e saídas direcionadas para os passos seguintes, visando o controle e a alteração dos formatos de arquivos específicos que são requisitos das ferramentas como, por exemplo, o alinhamento no formato

nexus para ProtTest (***Abascal et al., 2005***) e em formato phylip para PhyML (***Guindon et al., 2010***). A estratégia utilizada para manipulação dos dados conta com o que já está desenvolvido em Unix, pois facilita o rastreamento dos arquivos gerados ao realizar a busca por padrões já definidos no código para a saída de cada programa, o que está representado no pseudocódigo com o "*", significando que pode haver qualquer variação de nome, desde que a extensão do arquivo esteja atrelada à ferramenta que o gerou. Complementarmente, as ocorrências de *clusterN* ou *alinhamentoN*, indicam que a tarefa é realizada iterando sobre todas as *n* instâncias, havendo uma porção variável no nome do arquivo e outra invariável, o que permite o rastreamento da informação.

Para uma execução eficiente do *pipeline*, a recuperação de informações referentes às características da máquina na qual a análise está sendo processada se mostra muito importante em virtude da quantidade de dados utilizados e da complexidade algorítmica de algumas etapas da análise evolutiva. Por essa razão, por chamada de sistema, a informação referente ao número de *cores* disponíveis na máquina é adquirida e passada por parâmetro para os *softwares* GUIDANCE, ProtTest e MrBayes (***Huelsenbeck and Ronquist, 2001***), possibilitando que essas ferramentas que permitem paralelização sejam melhor utilizadas. Por padrão, foi estabelecido que metade da capacidade computacional da máquina seria empregada para execução dessas ferramentas.

Em comparação a outros *pipelines* descritos na literatura, o que é proposto aqui distingue-se no que tange à transparência da execução e fluidez do processo para o usuário, não sendo necessário qualquer modo de intervenção. Outro aspecto interessante é o viés evolutivo na escolha das ferramentas e de parâmetros, pois apesar de qualquer alinhador compatível com o GUIDANCE poder ser escolhido, por exemplo, por padrão, o PRANK (***Löytynoja and Goldman, 2005***; ***Löytynoja, 2014***) é utilizado e, como descrito por seus autores, ele não é genérico, mas sim utiliza diferentes métodos de alinhamento projetados para uso em estudos filogenéticos. Seguindo esse mesmo objetivo, a execução de testes com o ProtTest remove a arbitrariedade na escolha do modelo evolutivo e dos parâmetros adicionais, conferindo maior confiabilidade à inferência e caracterizando uma análise que melhor se adapta aos dados. Por fim, o filtro realizado com os dados do GUIDANCE tornam o alinhamento mais acurado, removendo ruído e levando à manutenção do sinal filogenético.

Entretanto, há pontos que podem ser melhorados na aplicação, os quais são explorados a seguir. Uma das limitações é encontrada na etapa de finalização do *pipeline*, quando realizado por inferência bayesiana. Durante a execução do MrBayes são gerados diversos arquivos muito informativos para fins de acompanhamento dos resultados e identificação de convergência, mesmo que parciais, indicando se a análise está ocorrendo de maneira adequada. As trocas (*swaps*) entre cadeias, a análise de convergência e o tamanho efetivo de amostragem são alguns desses registros, que constituem medidas importantes a serem avaliadas a fim de definirmos se a análise pode ser finalizada. Entretanto, em nosso *pipeline*, o único fator considerado foi o *average standard deviation of split frequencies*, sendo que quando esse parâmetro alcançava o valor igual ou menor do que 0.01, a inferência era finalizada e, subsequentemente, era realizada a sumarização dos resultados. Programas como o Tracer (***Rambaut et al., 2014***), proporcionam uma forma de avaliação desses dados, mas é dependente do usuário para inserção e atualização dos arquivos de *log*. Esse problema poderia ser contornado com a implementação de um módulo adicional que, executado concomitantemente ao MrBayes, acessa essas informações de maneira dinâmica e mantém o usuário ciente sobre o processamento.

Outro fator observado que é ligeiramente impeditivo é a pluralidade de modos com que um programa pode estar instalado na máquina que executará a análise. Como todas as ferramentas externas são utilizadas por chamada de sistema, caso essas não estejam configuradas no caminho do sistema, será necessário que o usuário informe a localização dos arquivos binários, dificultando a etapa de configuração inicial. Isso se mostra relevante, pois as ferramentas externas são requisitos funcionais do *pipeline*. Pode-se dizer também que o *pipeline* não alcança um dos objetivos mais apreciados nas aplicações de bioinformática (***Leipzig, 2017***), que é estimar o tempo

de processamento para cada etapa, informação essa que depende de fatores como a forma com que as ferramentas estão implementadas e o poder computacional, pontos que fogem um pouco do escopo do *pipeline* desenvolvido nesse estudo.

Análise Filogenômica de *Mycoplasma*

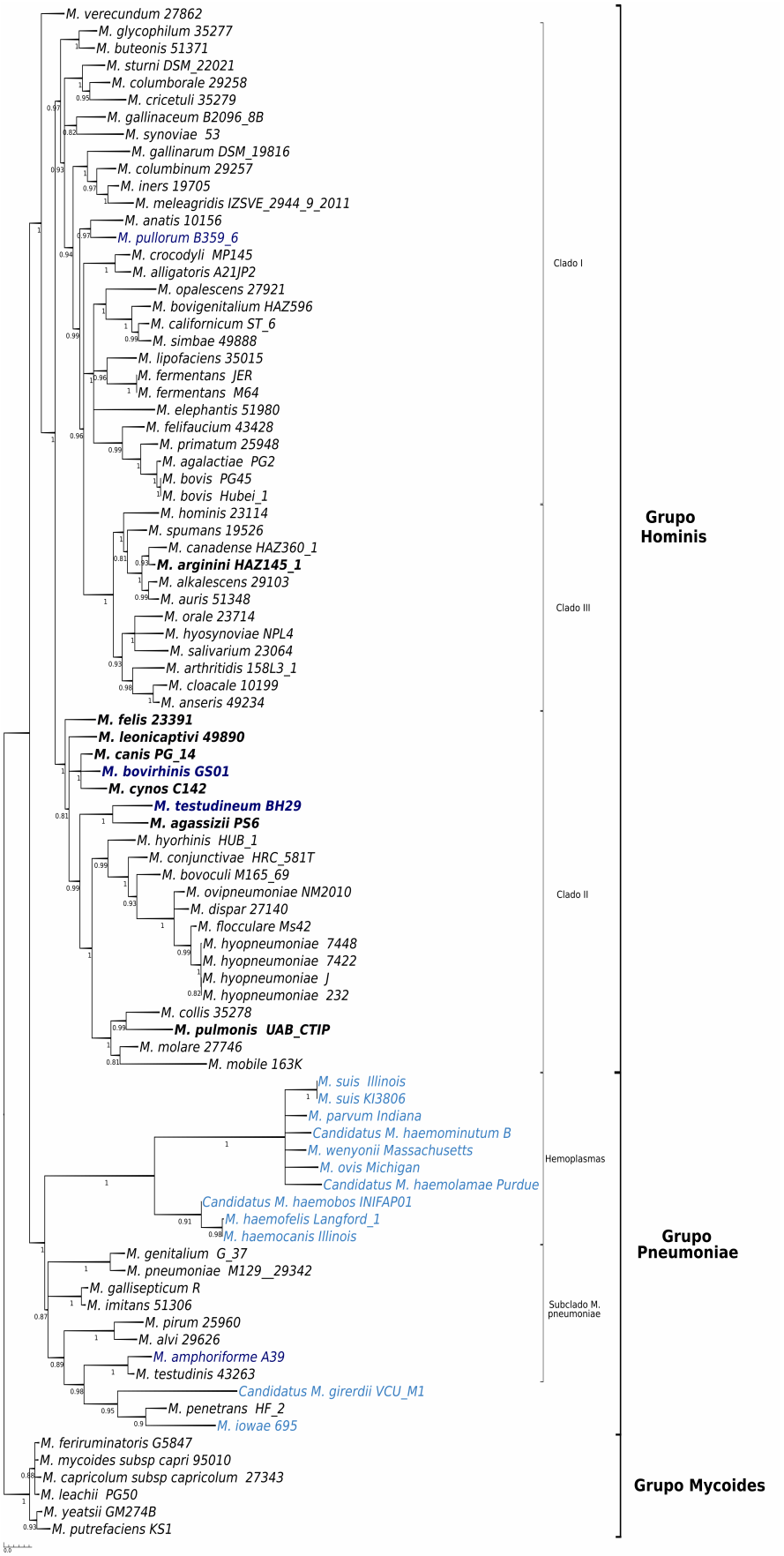
Os dados dos genomas utilizados para a análise filogenômica foram processados sem a inclusão de um grupo externo, pois assim como no trabalho realizado por **Yotoko and Bonatto** (**2007**), nosso intuito era identificar o maior número possível de marcadores moleculares e evitar a inserção de ruído que pudesse enviesar os resultados.

Para fins comparativos e de validação dos resultados, utilizamos as árvores de **Gupta et al.** (**2018**), obtidas com o uso de 63 proteínas, escolhidas tanto por possuírem bom suporte estatístico como por apresentarem maior número de ortólogos em relação a outros estudos. Importante observar que a árvore obtida por **Gupta et al.** (**2018**) incluiu 80 genomas de micoplasmas, mas não possui algumas espécies que usamos em nossa análise, como *M. amphoriforme*, *M. testudineum*, *M. bovirhinis* e *M. pullorum*. Dessa forma, realizamos também uma análise comparativa em relação à árvore de **Alvarez-Ponce et al.** (**2018**), baseada em 16S extraídos do banco de dados SILVA (**Quast et al., 2012**) e que abrange 113 genomas de micoplasmas, bem como em relação à árvore publicada por **Siqueira et al.** (**2013**), com base em 179 ortólogos, onde foram incluídas 31 micoplasmas.

A árvore resultante de nossa análise (Figura 3) apresentou similaridades em relação aos dados de **Gupta et al.** (**2018**) quando consideramos os grandes grupos: Hominis, Pneumoniae e Mycoides. Dentro de Hominis, é possível a identificação de três clados, tal como encontrado por **Gupta et al.** (**2018**). Entre as principais diferenças, podemos citar *M. felis*, *M. leonicaptivi*, *M. canis*, *M. bovirhinis*, *M.cynos*, *M. testudineum*, *M. agassizii* e *M. pulmonis* como pertencentes ao clado I em **Gupta et al.** (**2018**), enquanto que pertencentes ao clado II em nosso estudo. *M. verecundum* ficou basalmente posicionada em relação aos três clados de Hominis e, ao compararmos com as demais árvores, não foi encontrado nenhum consenso, pois **Gupta et al.** (**2018**) a agrupa com *M. buteonis* e *M. glycophilum* (tanto por 16S como por filogenômica) no clado I, e em **Alvarez-Ponce et al.** (**2018**) é colocada junto a *M. synoviae*. **Siqueira et al.** (**2013**) não utiliza essa espécie. As demais espécies de micoplasmas possuem um padrão de distribuição similar ao encontrado por **Gupta et al.** (**2018**) em cada clado. No entanto, algumas OTUs estão com o posicionamento diferente nos clados mais internos.

Ainda no grupo Hominis, uma espécie não utilizada por **Gupta et al.** (**2018**), *M. pullorum*, forma um grupo monofilético com *M. anatis*, resultado que se assemelha ao de **Alvarez-Ponce et al.** (**2018**), no qual *M. anatis* é basal em relação a *M. pullorum*. Quanto a *M. testudineum*, **Siqueira et al.** (**2013**) e **Gupta et al.** (**2018**) não a utilizaram na análise, mas na árvore de **Alvarez-Ponce et al.** (**2018**), essa espécie está próxima de *M. agassizii*, o que foi corroborado por nossos resultados. *M. bovirhinis*, também não utilizado por **Gupta et al.** (**2018**), apesar de estar em um clado politômico, está agrupado similarmente em **Alvarez-Ponce et al.** (**2018**). Adicionalmente, *M. arginini*, aqui posicionada no clado III, é apresentada no I em **Gupta et al.** (**2018**) e **Alvarez-Ponce et al.** (**2018**).

Quanto ao grupo Pneumoniae, é interessante observar que mais da metade (12) das OTUs foram inferidas somente a partir do gene *tuf*. O grupo monofilético *Candidatus M. haemobos*, *M. haemofelis* e *M. haemocanis* está idêntico ao encontrado por **Gupta et al.** (**2018**) e **Meli et al.** (**2010**), que a partir de 16S e utilizando outros 3 genomas de *M. haemobus*, encontraram a mesma relação. **Alvarez-Ponce et al.** (**2018**) e **Siqueira et al.** (**2013**) não utilizaram essas espécies. De modo geral, o grupo Pneumoniae foi o que manteve mais claramente as relações entre as espécies em todos os estudos comparados, como entre *M. genitalium* e *M. pneumoniae*, que estavam de acordo com o esperado. O padrão evolutivo de todas as OTUs em nosso trabalho está bastante similar ao descrito em **Gupta et al.** (**2018**), com exceção da relação entre algumas espécies posicionadas nos ramos mais internos e entre *M. testudinis* e *M. amphoriforme*, isso porque **Gupta et al.** (**2018**) não utiliza essa última espécie. Entretanto, **Alvarez-Ponce et al.** (**2018**) a utiliza e podemos confirmar, portanto, que a relação *M. testudinis* e *M. amphoriforme* é corroborada.


M. verecundum 27862
M. glycophilum 35277
M. buteonis 51371
M. sturni DSM_22021
M. columborale 29258
M. cricetuli 35279
M. gallinaceum B2096_8B
M. synoviae 53
M. gallinarum DSM_19816
M. columbinum 29257
M. iners 19705
M. meleagridis IZSVE_2944_9_2011
M. anatis 10156
M. pullorum B359_6
M. crocodyli MP145
M. alligatoris A21JP2
M. opalescens 27921
M. bovigenitalium HAZ596
M. californicum ST_6
M. simbae 49888
M. lipofaciens 35015
M. fermentans JER
M. fermentans M64
M. elephantis 51980
M. felifaucium 43428
M. primatum 25948
M. agalactiae PG2
M. bovis PG45
M. bovis Hubei_1
M. hominis 23114
M. spumans 19526
M. canadense HAZ360_1
M. arginini HAZ145_1
M. alkalescens 29103
M. auris 51348
M. orale 23714
M. hyosynoviae NPL4
M. salivarium 23064
M. arthritidis 158L3_1
M. cloacale 10199
M. anseris 49234
M. felis 23391
M. leonicaptivi 49890
M. canis PG_14
M. bovirhinis GS01
M. cynos C142
M. testudineum BH29
M. agassizii PS6
M. hyorhinis HUB_1
M. conjunctivae HRC_581T
M. bovoculi M165_69
M. ovipneumoniae NM2010
M. dispar 27140
M. flocculare Ms42
M. hyopneumoniae 7448
M. hyopneumoniae 7422
M. hyopneumoniae J
M. hyopneumoniae 232
M. collis 35278
M. pulmonis UAB_CTIP
M. molare 27746
M. mobile 163K
M. suis Illinois
M. suis KI3806
M. parvum Indiana
Candidatus M. haemominutum B
M. wenyonii Massachusetts
M. ovis Michigan
Candidatus M. haemolamae Purdue
Candidatus M. haemobos INIFAP01
M. haemofelis Langford_1
M. haemocanis Illinois
M. genitalium G_37
M. pneumoniae M129__29342
M. gallisepticum R
M. imitans 51306
M. pirum 25960
M. alvi 29626
M. amphoriforme A39
M. testudinis 43263
Candidatus M. girerdii VCU_M1
M. penetrans HF_2
M. iowae 695
M. feriruminatoris G5847
M. mycoides subsp capri 95010
M. capricolum subsp capricolum 27343
M. leachii PG50
M. yeatsii GM274B
M. putrefaciens KS1
Clado I
Clado III
Clado II
Hemoplasmas
Subclado M. pneumoniae
Grupo Hominis
Grupo Pneumoniae
Grupo Mycoides

306

Figure 3. Árvore filogenômica de micoplasmas obtida a partir da análise de 89 genomas. Inferência evolutiva realizada a partir da saída do *pipeline* desenvolvido nesse estudo. OTUs em azul claro representam as espécies inferidas somente com o gene *tuf*. OTUs em negrito são as que em comparação com ***Gupta et al.*** (***2018***), estão em clados diferentes. Espécies em azul escuro são as que ***Gupta et al.*** (***2018***) não utilizou. Ramos com suporte estatístico inferior a 0.8 foram colapsados e seu valor de suporte omitido. O alinhamento foi obtido com o programa PRANK. O modelo evolutivo utilizado foi LG, com sítios invariáveis, distribuição gama e frequência empírica e a máxima verossimilhança como método de reconstrução filogenética, implementada no programa PhyML. A árvore foi gerada e manipulada utilizando o *software* TreeGraph (***Stöver and Müller, 2010***).

O grupo de Hemoplasmas, por sua vez, foi completamente inferido a partir do gene *tuf*, sendo impossível deduzir qualquer relação para cinco de seus dez genomas, pois estão representadas por politomias. É possível perceber que isto tem relação com o fato de todos utilizarem o fator de alongamento Tu como único marcador, demonstrando que ele é bem conservado entre essas espécies, mas que por outro lado isso leva à perda do sinal filogenético e incapacidade de distinção entre elas. O mesmo aconteceu com o grupo Mycoides, monofilético, onde é possível ver o relacionamento próximo entre *M. yeatsii* e *M. putrefaciens*. No entanto, nesse grupo, apesar de ter havido a identificação de mais de um marcador além do *tuf*, ainda sim houve a formacão de uma politomia.

Devido aos genes ortólogos utilizados na análise e considerando que todas as espécies são igualmente representadas por somente um gene e tiveram, portanto, as relações estabelecidas majoritariamente a partir dele, é impreterível realizar a análise da topologia da árvore pesando esse fator. De acordo com ***Kamla et al.*** (***1996***), a informação inferida com o gene *tuf* é referente a aspectos fenotípicos de micoplasmas, separando os grupos de acordo com a presença ou ausência de proteínas de adesão e por atributos metabólicos. A fim de considerar os dados aqui utilizados sobre o *tuf*, seria necessário analisar em maior profundidade ponderando as propriedades fisiológicas e bioquímicas de micoplasmas. Entretanto, esse tipo de análise explora mais a história evolutiva desse gene, o *tuf*, e pode não representar a história da espécie, que é o foco desse estudo. Como também concluído por ***Kamla et al.*** (***1996***), é necessário adicionar mais marcadores a fim de atingir esse objetivo.

É possível observar que há incongruências nas árvores obtidas com esses 16 *clusters*, pois apesar de serem bem conservados, não há uma distribuição uniforme desses por todas as OTUs, levando a uma comparação não regular entre as espécies. Assim, duas espécies distantemente relacionadas, A e B, por exemplo, podem ser posicionadas mais proximamente se houver o que possa ser comparado entre elas, e podem não agrupar com uma terceira que sabe-se que agrupa ou com A ou B, mas que não possui os mesmos genes ortólogos a serem utilizados na inferência. Essa lógica, associada ao fato de que o gene *tuf* é extremamente conservado (***Razin et al., 1998***), explica porque houve alguns ramos com suporte estatístico baixo, prejudicando a inferência evolutiva. Como explicado na metodologia, esses ramos com suporte inferior a 0.8 foram colapsados, representado nas árvores por politomias. Como resultado, mais da metade (7) das espécies representadas somente pelo gene *tuf* estão em ramos politômicos, o que é justificável devido à natureza do gene e que leva à consequente ausência de sinal filogenético para a distinção dentre elas.

Avaliação da Função de Filtro do Alinhamento

Na etapa de alinhamento, utilizamos um filtro atestado por ***Tan et al.*** (***2015***). O intuito desse estudo foi examinar o impacto do uso de ferramentas de filtro em estudos filogenéticos, chegando à conclusão de que remover até 20% dos sítios de um alinhamento é um bom limite, ainda sendo considerado conservativo. Dessa forma, ao aplicarmos um filtro similar a todos os *clusters* antes da construção da supermatriz, evitamos, teoricamente, problemas como a perda do sinal filogenético. A fim de avaliarmos a importância do filtro utilizado na etapa de alinhamento, realizamos uma comparação das árvores obtidas a partir de alinhamentos com e sem filtro, que estão exibidas na Figura 4.

As árvores filogenômicas resultantes apresentaram topologias muito semelhantes, onde todos os grupos principais se mantiveram: Hominis, Pneumoniae e Mycoides. No entanto, é possível

observar discrepâncias nos ramos mais internos, bem como um suporte estatístico maior na árvore filtrada, como já era esperado de acordo com ***Tan et al.*** (***2015***).

A maior diferença entre essas árvores está no posicionamento de *Candidatus M. girerdii*, que está posicionada junto a *M. penetrans* e *M. iowae* na árvore obtida a partir do alinhamento filtrado, mas está basal em relação a todos os genomas do grupo Pneumoniae na árvore obtida a partir do alinhamento integral. Entretanto, as análises de ***Gupta et al.*** (***2018***), tanto de 16S como de filogenômica, apresentam o mesmo resultado da árvore filtrada, enfatizando a importância da utilização de ferramentas de filtro de alinhamento na análise filogenômica. ***Alvarez-Ponce et al.*** (***2018***) não utilizou essa espécie, não sendo possível estabelecer uma comparação.

A árvore com filtro também apresentou resultados positivos em alguns outros grupos quando comparamos àquela sem o filtro, onde há, por exemplo, dois nodos com baixo suporte estatístico. O clado formado pelas cepas de *M. hyopneumoniae*, na árvore com filtro, mostra uma relação próxima entre as cepas J e 232. Além disso, há um grupo formado por *M. salivarum, M. hyosynoviae* e *M. orale*, também com aumento do suporte estatístico. No entanto, esses dados devem ser avaliados com cuidado, visto que ***Siqueira et al.*** (***2013***) não encontraram essa relação entre *M. hyopneumoniae* J e *M. hyopneumoniae* 232.

Esses dados devem ser reavaliados considerando diferentes valores de *cut-off* para o GUIDANCE. Assim, um estudo específico focado em micoplasmas pode contribuir para estabelecer um melhor critério para o uso dos filtros de alinhamento múltiplo nessas espécies. Também é importante considerarmos que o uso de critérios de identidade e cobertura menos conservadores para a busca de ortólogos pode levar a um número maior de grupos de ortólogos identificados nos genomas em análise. Consequentemente, um número maior de ortólogos levará a um alinhamento maior e, nesse caso, a importância da ferramenta de filtro dos alinhamentos poderá ter uma relevância ainda maior.

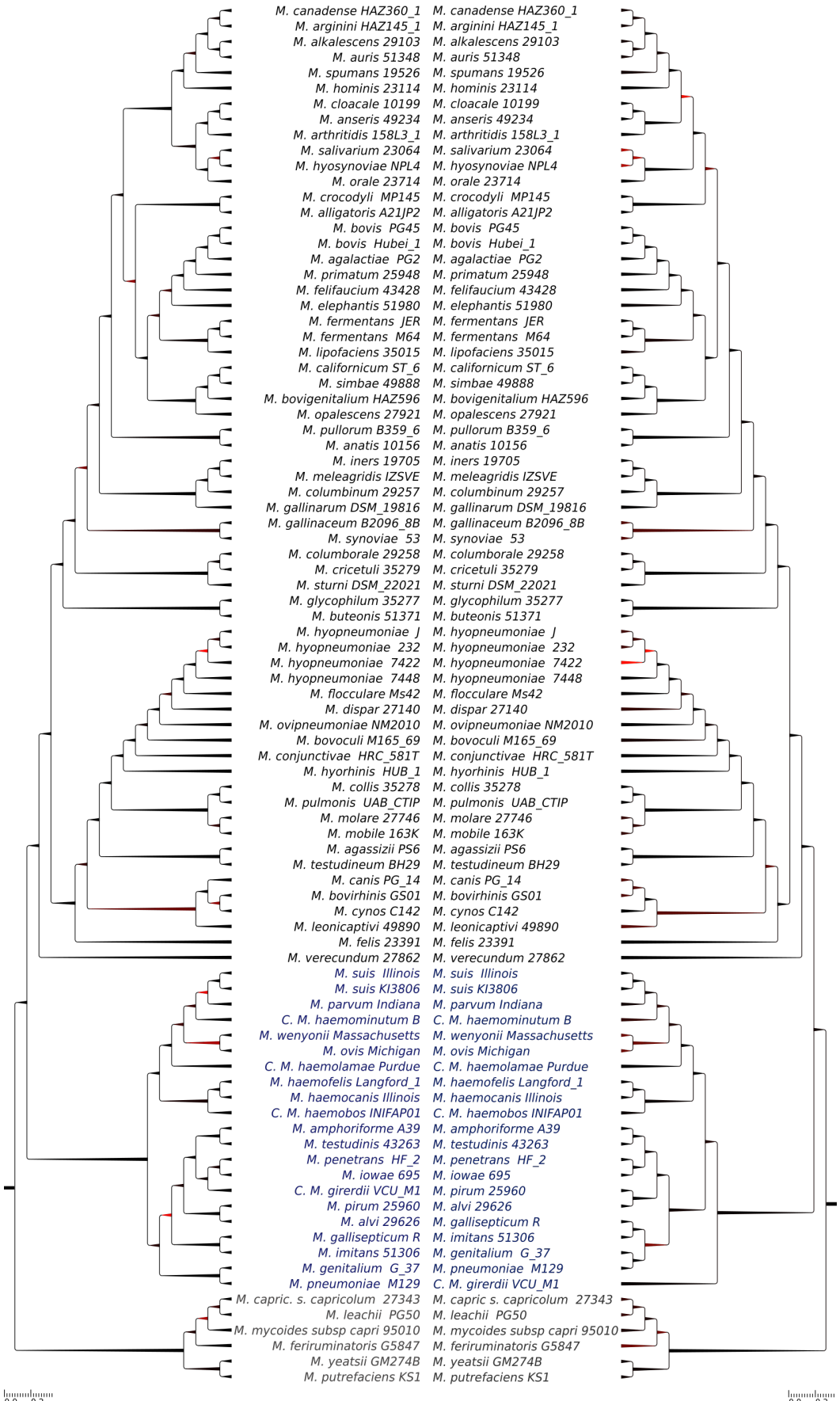

M. canadense HAZ360_1
M. arginini HAZ145_1
M. alkalescens 29103
M. auris 51348
M. spumans 19526
M. hominis 23114
M. cloacale 10199
M. anseris 49234
M. arthritidis 158L3_1
M. salivarium 23064
M. hyosynoviae NPL4
M. orale 23714
M. crocodyli MP145
M. alligatoris A21JP2
M. bovis PG45
M. bovis Hubei_1
M. agalactiae PG2
M. primatum 25948
M. felifaucium 43428
M. elephantis 51980
M. fermentans JER
M. fermentans M64
M. lipofaciens 35015
M. californicum ST_6
M. simbae 49888
M. bovigenitalium HAZ596
M. opalescens 27921
M. pullorum B359_6
M. anatis 10156
M. iners 19705
M. meleagridis IZSVE
M. columbinum 29257
M. gallinarum DSM_19816
M. gallinaceum B2096_8B
M. synoviae 53
M. columborale 29258
M. cricetuli 35279
M. sturni DSM_22021
M. glycophilum 35277
M. buteonis 51371
M. hyopneumoniae J
M. hyopneumoniae 232
M. hyopneumoniae 7422
M. hyopneumoniae 7448
M. flocculare Ms42
M. dispar 27140
M. ovipneumoniae NM2010
M. bovoculi M165_69
M. conjunctivae HRC_581T
M. hyorhinis HUB_1
M. collis 35278
M. pulmonis UAB_CTIP
M. molare 27746
M. mobile 163K
M. agassizii PS6
M. testudineum BH29
M. canis PG_14
M. bovirhinis GS01
M. cynos C142
M. leonicaptivi 49890
M. felis 23391
M. verecundum 27862
M. suis Illinois
M. suis KI3806
M. parvum Indiana
C. M. haemominutum B
M. wenyonii Massachusetts
M. ovis Michigan
C. M. haemolamae Purdue
M. haemofelis Langford_1
M. haemocanis Illinois
C. M. haemobos INIFAP01
M. amphoriforme A39
M. testudinis 43263
M. penetrans HF_2
M. iowae 695
C. M. girerdii VCU_M1
M. pirum 25960
M. alvi 29626
M. gallisepticum R
M. imitans 51306
M. genitalium G_37
M. pneumoniae M129
M. capric. s. capricolum 27343
M. leachii PG50
M. mycoides subsp capri 95010
M. feriruminatoris G5847
M. yeatsii GM274B
M. putrefaciens KS1
M. canadense HAZ360_1
M. arginini HAZ145_1
M. alkalescens 29103
M. auris 51348
M. spumans 19526
M. hominis 23114
M. cloacale 10199
M. anseris 49234
M. arthritidis 158L3_1
M. salivarium 23064
M. hyosynoviae NPL4
M. orale 23714
M. crocodyli MP145
M. alligatoris A21JP2
M. bovis PG45
M. bovis Hubei_1
M. agalactiae PG2
M. primatum 25948
M. felifaucium 43428
M. elephantis 51980
M. fermentans JER
M. fermentans M64
M. lipofaciens 35015
M. californicum ST_6
M. simbae 49888
M. bovigenitalium HAZ596
M. opalescens 27921
M. pullorum B359_6
M. anatis 10156
M. iners 19705
M. meleagridis IZSVE
M. columbinum 29257
M. gallinarum DSM_19816
M. gallinaceum B2096_8B
M. synoviae 53
M. columborale 29258
M. cricetuli 35279
M. sturni DSM_22021
M. glycophilum 35277
M. buteonis 51371
M. hyopneumoniae J
M. hyopneumoniae 232
M. hyopneumoniae 7422
M. hyopneumoniae 7448
M. flocculare Ms42
M. dispar 27140
M. ovipneumoniae NM2010
M. bovoculi M165_69
M. conjunctivae HRC_581T
M. hyorhinis HUB_1
M. collis 35278
M. pulmonis UAB_CTIP
M. molare 27746
M. mobile 163K
M. agassizii PS6
M. testudineum BH29
M. canis PG_14
M. bovirhinis GS01
M. cynos C142
M. leonicaptivi 49890
M. felis 23391
M. verecundum 27862
M. suis Illinois
M. suis KI3806
M. parvum Indiana
C. M. haemominutum B
M. wenyonii Massachusetts
M. ovis Michigan
C. M. haemolamae Purdue
M. haemofelis Langford_1
M. haemocanis Illinois
C. M. haemobos INIFAP01
M. amphoriforme A39
M. testudinis 43263
M. penetrans HF_2
M. iowae 695
M. pirum 25960
M. alvi 29626
M. gallisepticum R
M. imitans 51306
M. genitalium G_37
M. pneumoniae M129
C. M. girerdii VCU_M1
M. capric. s. capricolum 27343
M. leachii PG50
M. mycoides subsp capri 95010
M. feriruminatoris G5847
M. yeatsii GM274B
M. putrefaciens KS1
376
0.0 0.2
0.0 0.3

Figure 4. Árvores filogenômicas de micoplasmas obtidas a partir de alinhamentos que sofreram até 20% de corte em seu conteúdo (árvore à direita) e inferida a partir do conteúdo integral dos alinhamentos (árvore à esquerda), as quais possuem, respectivamente, 5.276 e 6.426 sítios. Em preto: grupo Hominis, em azul: grupo Pneumoniae, em cinza: grupo Mycoides. Estão representados em vermelho os ramos com suporte estatístico inferior a 0.8. O alinhamento foi obtido com o programa PRANK. Os parâmetros utilizados na inferência são os mesmos da Figura 3.

Como perspectiva desse trabalho, pretendemos futuramente incluir novas funções no *pipeline* como, por exemplo, o método de reconstrução filogenética por distância, que funcionará de maneira paralela juntamente aos demais métodos de reconstrução, e a parte inicial de identificação de ortólogos. Adicionalmente, faremos a análise de todos os genomas de micoplasmas disponíveis no NCBI, incluindo todas as cepas de micoplasmas. Atualmente, nesse banco de dados se encontram disponíveis 90 espécies, sendo que se considerarmos todas as diferentes cepas o número total de genomas é de 379. Nosso estudo considerou um genoma representativo de cada espécie de micoplasma, com exceção de *M. moatsii* cujo genoma foi removido da análise em função de haver sinalização de contaminação, conforme descrito na seção "Materiais e Métodos". Paralelamente, realizaremos a análise evolutiva considerando um grupo externo ao gênero *Mycoplasma*.

Materiais e Métodos

As sequências foram obtidas no NCBI e as cepas de cada um dos 89 genomas foram escolhidas de acordo com os seguintes critérios: (i) genoma completo ou com menor número de *scaffolds*, (ii) se a cepa possui atividade, descrita na literatura, como patógeno e (iii) qualidade da montagem (N50) quando mais de um genoma cumpria os outros requisitos. Para determinados micro-organismos, como *M. hyopneumoniae*, *M. bovis*, *M. fermentans* e *M. suis*, foram escolhidos mais de um genoma de diferentes cepas, a fim de incluir mais cepas patogênicas e melhor entender a relação evolutiva entre elas e as não patogênicas, tendo um total de quatro genomas de *M. hyopneumoniae* e dois de cada uma das outras cepas anteriormente citadas. *M. moatsii* foi removido da análise, pois a única montagem disponível estava assinalada como contaminada.

A busca por genes ortólogos foi realizada com a ferramenta Proteinortho (***Lechner et al., 2011***), pois de acordo com ***Nichio et al.*** (***2017***) é a que apresenta melhor desempenho e acurácia quando é utilizada uma grande quantidade de dados. A manipulação dos resultados, bem como sua transformação no *pipeline*, foi realizada por meio de comandos Unix, *scripts* nas linguagens de programação perl (***Wall et al., 1999***) e python (***Van Rossum et al., 2007***), sendo python a linguagem mais utilizada. A biblioteca utilizada em python, já projetada para bioinformática, foi a BioPython (***Cock et al., 2009***). Adicionalmente, as ferramentas externas definidas a seguir, foram todas implementadas por chamada de sistema, com a biblioteca *os* (*operating system*) do python.

As principais funcionalidades do *pipeline* são desempenhadas por programas externos, majoritariamente projetados para uso por linha de comando em Unix, os quais possuem suas respectivas dependências e não dispõem de interface de programação de aplicações (APIs) para integração com python. Por esse motivo, chamadas de sistema configurando os comandos com os parâmetros desejados foram utilizadas para a ferramenta GUIDANCE 2.02, ProtTest 2.1, para *scritps* internos de transformação, PhyML 20120412 e MrBayes 3.2.6. O GUIDANCE foi utilizado para o alinhamento com o PRANK 140110, sendo que o suporte do alinhador foi aplicado aos dados com parâmetros *default*. Seu resultado (pontuação e alinhamento originais) foram, subsequentemente, utilizados por *scripts* internos para filtrar o alinhamento excluindo até 20% do tamanho dos *clusters*, limite estabelecido de acordo com ***Tan et al.*** (***2015***).

O ProtTest foi utilizado para cálculo dos modelos evolutivos e identificação de qual o melhor modelo para cada grupo de ortólogos, executado com os seguintes parâmetros: testes incluindo variação de taxa e número de categoria entre sítios, matrizes JTT, LG, Dayhoff, WAG e Blosum62 e inclusão de modelos com estimativa empírica de frequências. Como resultado o melhor modelo evolutivo para os dados sendo analisados foi LG, com os parâmetros adicionais +I+G+F. Em relação aos métodos de reconstrução filogenética foram utilizados dois probabilísticos: (i) máx-

ima verossimilhança, implementado pelo PhyML e (ii) inferência bayesiana, implementado pelo MrBayes. Para o MrBayes foi utilizado o critério de 25% de *burn-in*, parada automatizada quando o parâmetro *average standard deviation of split frequencies* atingisse um valor igual ou menor que 0.01, frequência de amostragem de 100, quatro cadeias e código genético configurado para *Mycoplasma*. No PhyML foram utilizados parâmetros para sequências de aminoácido (*-d aa*), com suporte estatístico calculado pelo método de ***Guindon et al.*** (***2010***) (*-b -4*), matriz de substituição LG (***Le and Gascuel, 2008***) (*-m LG*), distribuição gama (*-a e*), sítios invariáveis (*-v e*) e frequência de caracteres (*-f e*), todos estimados.

Agradecimentos

Os autores gostariam de agradecer Rangel Celso Souza e Ana Tereza Ribeiro de Vasconcelos, do Laboratório Nacional de Computação Científica (LNCC), pela disponibilidade de servidores e simulações para identificação dos ortólogos. Agradecemos também à Universidade Federal de Ciências da Saúde de Porto Alegre (UFCSPA) e Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelas bolsas de Iniciação Científica e de Doutorado, de M.M.V. e F.C.A., respectivamente, e ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo apoio financeiro e recursos computacionais obtidos por meio do projeto aprovado no Edital Universal 2014 (processo nº 458160/2014-8) sob coordenação da professora Claudia E. Thompson.

Material Suplementar

Material suplementar disponível em: https://goo.gl/91KCj8.

References

Abascal F, Zardoya R, Posada D. ProtTest: selection of best-fit models of protein evolution. Bioinformatics. 2005; 21(9):2104–2105.

Álvarez-Jarreta J, Ruiz-Pesini E. MEvoLib v1. 0: the first molecular evolution library for Python. BMC bioinformatics. 2016; 17(1):436.

Alvarez-Ponce D, Weitzman CL, Tillett RL, Sandmeier FC, Tracy CR. High quality draft genome sequences of Mycoplasma agassizii strains PS6 T and 723 isolated from Gopherus tortoises with upper respiratory tract disease. Standards in genomic sciences. 2018; 13(1):12.

Barykova YA, Logunov DY, Shmarov MM, Vinarov AZ, Fiev DN, Vinarova NA, Rakovskaya IV, Baker PS, Shyshynova I, Stephenson AJ, et al. Association of Mycoplasma hominis infection with prostate cancer. Oncotarget. 2011; 2(4):289.

Benson DA, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. Nucleic acids research. 2015; 43(Database issue):D30.

Brown DR. Phylum XVI. Tenericutes Murray 1984a, 356 VP (Effective publication: Murray 1984b, 33.). In: *Bergey's Manual® of Systematic Bacteriology* Springer; 2010.p. 567–723.

Citti C, Dordet-Frisoni E, Nouvel L, Kuo C, Baranowski E. Horizontal Gene Transfers in Mycoplasmas (Mollicutes). Current issues in molecular biology. 2018; 29:3–22.

Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009; 25(11):1422–1423.

Eisen JA, Fraser CM. Phylogenomics: intersection of evolution and genomics. Science. 2003; 300(5626):1706.

Guimaraes AM, Santos AP, do Nascimento NC, Timenetsky J, Messick JB. Comparative genomics and phylogenomics of hemotrophic mycoplasmas. PloS one. 2014; 9(3):e91445.

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Systematic biology. 2010; 59(3):307–321.

Gupta RS, Son J, Oren A. A phylogenomic and molecular markers based taxonomic framework for members of the order Entomoplasmatales: proposal for an emended order Mycoplasmatales containing the family Spiroplasmataceae and emended family Mycoplasmataceae comprised of six genera. Antonie van Leeuwenhoek. 2018; p. 1–28.

Hannan P. Comparative susceptibilities of various AIDS-associated and human urogenital tract mycoplasmas and strains of Mycoplasma pneumoniae to 10 classes of antimicrobial agent in vitro. Journal of medical microbiology. 1998; 47(12):1115–1122.

Himmelreich R, Hilbert H, Plagens H, Pirkl E, Li BC, Herrmann R. Complete sequence analysis of the genome of the bacterium Mycoplasma pneumoniae. Nucleic acids research. 1996; 24(22):4420–4449.

Huelsenbeck JP, Ronquist F. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics. 2001; 17(8):754–755.

Kamla V, Henrich B, Hadding U. Phylogeny based on elongation factor Tu reflects the phenotypic features of mycoplasmas better than that based on 16S rRNA. Gene. 1996; 171(1):83–87.

Kamminga T, Koehorst JJ, Vermeij P, Slagman SJ, Martins dos Santos VA, Bijlsma JJ, Schaap PJ. Persistence of functional protein domains in mycoplasma species and their role in host specificity and synthetic minimal life. Frontiers in cellular and infection microbiology. 2017; 7:31.

Le SQ, Gascuel O. An improved general amino acid replacement matrix. Molecular biology and evolution. 2008; 25(7):1307–1320.

Lechner M, Findeiß S, Steiner L, Marz M, Stadler PF, Prohaska SJ. Proteinortho: detection of (co-) orthologs in large-scale analysis. BMC bioinformatics. 2011; 12(1):124.

Leclercq S, Dittmer J, Bouchon D, Cordaux R. Phylogenomics of “Candidatus Hepatoplasma crinochetorum,” a lineage of mollicutes associated with noninsect arthropods. Genome biology and evolution. 2014; 6(2):407–415.

Leipzig J. A review of bioinformatic pipeline frameworks. Briefings in bioinformatics. 2017; 18(3):530–536.

Lo WS, Gasparich GE, Kuo CH. Convergent evolution among ruminant-pathogenic Mycoplasma involved extensive gene content changes. Genome biology and evolution. 2018; 10(8):2130–2139.

Löytynoja A. Phylogeny-aware alignment with PRANK. In: *Multiple sequence alignment methods* Springer; 2014.p. 155–170.

Löytynoja A, Goldman N. An algorithm for progressive multiple alignment of sequences with insertions. Proceedings of the National Academy of Sciences. 2005; 102(30):10557–10562.

Maniloff J. The minimal cell genome:" on being the right size". Proceedings of the National Academy of Sciences of the United States of America. 1996; 93(19):10004.

Mavedzenge SN, Weiss HA. Association of Mycoplasma genitalium and HIV infection: a systematic review and meta-analysis. Aids. 2009; 23(5):611–620.

May M, Balish MF, Blanchard A. The Order Mycoplasmatales. In: *The Prokaryotes* Springer; 2014.p. 515–550.

Meli ML, Willi B, Dreher UM, Cattori V, Knubben-Schweizer G, Nuss K, Braun U, Lutz H, Hofmann-Lehmann R. Identification, molecular characterization, and occurrence of two bovine hemoplasma species in Swiss cattle and development of real-time TaqMan quantitative PCR assays for diagnosis of bovine hemoplasma infections. Journal of clinical microbiology. 2010; 48(10):3563–3568.

Nichio BT, Marchaukoski JN, Raittz RT. New Tools in Orthology Analysis: A Brief Review of Promising Perspectives. Frontiers in genetics. 2017; 8:165.

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic acids research. 2012; 41(D1):D590–D596.

Rambaut A, Drummond A, Suchard M. Tracer v1. 6 http://beast. bio. ed. ac. uk. Tracer (visited on 2017-06-12). 2014; .

Razin S, Yogev D, Naot Y. Molecular biology and pathogenicity of mycoplasmas. Microbiology and molecular biology reviews. 1998; 62(4):1094–1156.

Rocha EP, Blanchard A. Genomic repeats, genome plasticity and the dynamics of Mycoplasma evolution. Nucleic acids research. 2002; 30(9):2031–2042.

Rokas A, Williams BL, King N, Carroll SB. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature. 2003; 425(6960):798.

Sela I, Ashkenazy H, Katoh K, Pupko T. GUIDANCE2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. Nucleic Acids Research. 2015; 43(W1):W7–W14.

Siqueira FM, Thompson CE, Virginio VG, Gonchoroski T, Reolon L, Almeida LG, da Fonsêca MM, de Souza R, Prosdocimi F, Schrank IS, et al. New insights on the biology of swine respiratory tract mycoplasmas from a comparative genome analysis. BMC genomics. 2013; 14(1):175.

Sirand-Pugnet P, Citti C, Barré A, Blanchard A. Evolution of mollicutes: down a bumpy road with twists and turns. Research in microbiology. 2007; 158(10):754–766.

Stöver BC, Müller KF. TreeGraph 2: combining and visualizing evidence from different phylogenetic analyses. BMC bioinformatics. 2010; 11(1):7.

Talevich E, Invergo BM, Cock PJ, Chapman BA. Bio. Phylo: a unified toolkit for processing, analyzing and visualizing phylogenetic trees in Biopython. BMC bioinformatics. 2012; 13(1):209.

Tan G, Muffato M, Ledergerber C, Herrero J, Goldman N, Gil M, Dessimoz C. Current methods for automated filtering of multiple sequence alignments frequently worsen single-gene phylogenetic inference. Systematic biology. 2015; 64(5):778–791.

Van Rossum G, et al. Python Programming Language. In: *USENIX Annual Technical Conference*, vol. 41; 2007. p. 36.

Wall L, Christiansen T, Schwartz RL. Programming perl. . 1999; .

Wang RH, Hayes MM, Wear D, Lo S, Shih JK, Alter H, Grandinetti T, Pierce P. High frequency of antibodies to Mycoplasma penetrans in HIV-infected patients. The Lancet. 1992; 340(8831):1312–1316.

Wang Z, Wu M. A phylum-level bacterial phylogenetic marker database. Molecular biology and evolution. 2013; 30(6):1258–1262.

Xie X, Yang M, Ding Y, Chen J. Microbial infection, inflammation and epithelial ovarian cancer. Oncology letters. 2017; 14(2):1911–1919.

Yotoko KS, Bonatto SL. A phylogenomic appraisal of the evolutionary relationship of Mycoplasmas. Genetics and Molecular Biology. 2007; 30(1):270–276.

Zaha A, Ferreira HB, Passaglia LM. Biologia Molecular Básica-5. Artmed Editora; 2014.

Appendix 1

Pseudocódigo

Função controle

```
INÍCIO;

   cpus ← CHAMA sistema.cpus();
   argumentos ← [arquivo de entrada com parâmetros;
   CHAMA montaClusters(diretório);
   CHAMA guidance.run(argumentos[parametros_guidance],cpus);
   CHAMA guidance.aplica_filtro();

   SE argumentos[parametros_alinhador] == PhyML;
      CHAMA transforma_seq.SuperMatriz();
      CHAMA tranforma_seq.Converte(fasta, faa, nexus);
      CHAMA prottest.run(argumentos[parametros_prottest],cpus);
      CHAMA transforma_seq.Converte(nexus, nex, phylip);
      CHAMA Filo.run(PhyML);

   SENÃO;
      CHAMA tranforma_seq(fasta, faa, nexus);
      CHAMA prottest.run(argumentos[parametros_prottest],cpus);
      CHAMA transforma_seq.SuperMatriz();
      CHAMA prottest.parse();
      CHAMA Filo.run(MrBayes)

   CHAMA transforma_seq.Colapsar();
FIM;
```

Módulo *bio-transform*

```
INÍCIO;
   CARREGA BioPython, os, subprocess;

   FUNÇÃO converte(formato_inicial, extensão, formato_final);
      arquivos ← sistema.listaArquivos(*.extensao);
      para cada arquivo em arquivos;
         Bio.converte(arquivo, formato_inicial, formato_final);

   FUNÇÃO SuperMatriz();
      arquivos ← sistema.listaArquivos(*.extensao);
      supermatriz ← Bio.Nexus.combine(arquivos);
      supermatriz.salva();
```

Módulo GUIDANCE

```
INÍCIO;
FUNÇÃO run;
pastas ← sistema.listaDiretórios();
        para cada pasta em pastas;
         nome_arquivo ← pasta[nome];
```

```
         CHAMA sistema(guidance -nome_arquivo -parametro1 -parametroN -cpus)

FUNÇÃO aplica_filtro;
     pastas ← sistema.listaDiretórios();
        para cada pasta em pastas;
         alinhamento ← pasta[nome]+MSA.PRANK.aln.With_Names;
            scores ← pasta[nome]+MSA.PRANK.Guidance2_col_col.scr;
            aln_filtrado ← sistema(python aplica_filtro.py -alinhamento -score);
         ENQUANTO aln_filtrado[excluído] > 0.2;
         aln_filtrado ← sistema(python aplica_filtro.py -alinhamento -score);
FIM;
```

Módulo ProtTest

```
INÍCIO;
FUNÇÃO parse;
     arquivo de saída <- modellist.out;
        arquivos ← sistema.listaArquivos(*.prottest);
        modellist ← sistema.encontraPadrão(arquivos, modelo);

    FUNÇÃO run;
     arquivos ← sistema.listaArquivos(*.nex);
        para cada arquivo em arquivos:
        CHAMA sistema(prottest -arquivo -parametro1 -parametro2 -
        parametroN -cpus -saída=arquivo.prottest)
FIM;
```

Módulo Filo

```
INÍCIO;
FUNÇÃO run(programa);
    SE programa == PhyML;
     parametros[] ← sistema.encontraPadrão(*.prottest, modelo,
parâmetros_adicionais);
matriz ← sistema.listaArquivos(*matriz.phy);
        CHAMA sistema(phyml - matriz -parametro1 -parametroN);
    SENÃO;
     matriz ← sistema.listaArquivos(*matriz.nex);
        parâmetros ← sistema(python mb.py);
        matriz.append(parâmetros);
        CHAMA sistema(MrBayes -matriz);
FIM;
```

4 DISCUSSÃO

O trabalho aqui desenvolvido envolve diferentes áreas do conhecimento, tais como genética, evolução molecular, estatística, métodos em computação e programação, demonstrando assim a importância da interdisciplinaridade para a resolução de problemas em biologia. A complexidade dos sistemas biológicos requer que utilizemos estratégias sofisticadas para a análise de dados e incentiva o desenvolvimento de novos algoritmos, bem como novas hipóteses para explicar a natureza dos dados. No *pipeline* desenvolvido foram aplicadas boas práticas de desenvolvimento de *software* e gerenciamento de projeto, que na prática são registros de entrada e saída de dados e seus respectivos formatos, reutilização de código pelo uso de funções bem definidas e documentação de todas as funcionalidades e como estas se relacionam, visando aprimoramentos futuros evitando problemas tais como a troca de pesquisadores envolvidos com a aplicação.

É importante dizer que os resultados atingidos não esgotaram todos os objetivos inicialmente pretendidos com o desenvolvimento desse *pipeline*. Como perspectivas futuras e analisando os dados obtidos com o estudo de caso, melhorias em nível de implementação poderiam ser aplicadas. Dentre elas, está a meta de colocá-lo conforme exigido pela *Commom Workflow Language* (CWL) (AMSTUTZ et al., 2016), que é uma série de especificações para descrição de *workflows* de análise de dados visando a portabilidade das plataformas, iniciativa crescente entre as ciências de uso intensivo de dados, como processamento de imagens, física, química e bioinformática. Esse tipo de complementação coloca a aplicação em um nível mais acessível tanto para desenvolvedores como para usuários, visto que há uma documentação descrita que recebe suporte da comunidade.

Outros aspectos que devem ser melhor avaliados posteriormente são referentes às etapas do processamento. No presente estudo, o *pipeline* foi desenvolvido a partir de ortólogos já previamente identificados e selecionados, não havendo preocupação com os parâmetros de escolha nem com a automatização desse processo. A ideia é futuramente explorar essa etapa na tentativa de incluí-la na análise, permitindo que o único requisito para o usuário seja configurar o caminho do diretório de genomas, caso a automatização seja possível. Importante salientarmos que no caso em estudo foram utilizados critérios que se mostraram muito conservadores para a identidade e cobertura definidos no proteinortho, 60% para ambos. Dessa forma, houve a identificação de poucos grupos de ortólogos (GOs) para os 89 genomas sendo analisados, um total de 16 GOs.

A etapa de aquisição de dados de ortólogos teve grande impacto nos resultados. Alguns clados inteiros foram inferidos a partir de somente um gene, enviesando a análise, já que diversos ramos acabaram não sendo classificados corretamente, de acordo com outros estudos disponíveis na literatura. Assim, estão em andamento novas análises de identificação de ortólogos, considerando critérios menos restritivos que, possivelmente, permitirão a inclusão de mais genes na análise, sendo todos uniformemente distribuídos pelas unidades taxonômicas em estudo. Todas as inferências evolutivas serão refeitas utilizando também esses novos resultados, a fim de avaliarmos quais os melhores critérios para esse conjunto de dados.

O baixo número de grupos ortólogos identificados, um total de 16 abrangendo no máximo 33% das espécies, se deve ao parâmetro utilizado para a busca dos *clusters*, 60% de identidade e cobertura. É verdadeiro dizer que esses são parâmetros bem conservativos, visto que diversos outros autores utilizam entre 30 e 50% de identidade e cobertura para encontrar sequências homólogas. Novas estimativas do número de GOs, resultantes de análises preliminares com alterações dos parâmetros acima citados no proteinortho, indicam a presença de 25 GOs com 60% de cobertura e 40% de identidade e 40 GOs também com 60% de cobertura e 35% de identidade. Entretanto, apesar desses resultados preliminares serem positivos, o intuito é encontrar um número ainda maior, pois YOTOKO; BONATTO, SIQUEIRA et al., LO et al., VASCONCELOS et al., OSHIMA; NISHIDA e outros autores, identificaram 227, 179, 161, 146 e 143 GOs, respectivamente, analisando um número diferente de genomas de micoplasmas.

Além da mudança de parâmetros, pretendemos realizar uma extensa revisão bibliográfica a fim de buscar outras ferramentas de identificação de ortólogos que possam ser avaliadas, testadas e comparadas ao proteinortho, em termos de rapidez de análise e acurácia dos resultados. Uma ferramenta promissora é o programa programa OrthoMCL (LI et al., 2003). No entanto, é conhecido que o tempo de processamento de dados genômicos com OrthoMCL é superior ao do proteinortho. Levando em consideração que todos os estudos filogenômicos necessitam passar por essa etapa de análise, é fundamental que possamos avaliar de maneira adequada as melhores ferramentas disponíveis e incluir em nosso *pipeline* aquela com melhor desempenho.

Um ponto crucial para qualquer inferência evolutiva é o alinhamento múltiplo. Existem algumas ferramentas descritas na literatura que aplicam filtros nos alinhamentos, como já comentado anteriormente, sendo que o GUIDANCE (SELA et al., 2015) é uma das que possui melhor performance e, por essa razão, foi implementada em nosso *pipeline*. Por meio dessa ferramenta é possível excluirmos regiões do alinhamento múltiplo que possuam baixa confiabilidade, cujos sítios não apresentem homologia posicional. A exclusão dessas regiões aumenta a confiabilidade

dos alinhamentos e das posteriores análises evolutivas. Os resultados de nossas análises mostram que o *cut-off* de 20% foi adequado.

Em relação aos métodos de reconstrução filogenética, pretendemos incluir o método de distância em nosso *pipeline*. Tal método é capaz de analisar muito rapidamente um grande número de sequências biológicas, o que se torna muito interessante quando analisamos muitos genomas. Em especial, será importante a inclusão desse método quando estivermos utilizando o *pipeline* para análise de genomas de eucariotos. De fato, no futuro, pretendemos desenvolver e aprimorar uma ferramenta capaz de lidar de maneira rápida e precisa com esse tipo de dado. Além dos aprimoramentos citados até aqui, seria interessante a inclusão de ferramentas para a edição de árvores filogenéticas, já que se trata de etapa final e muito importante para que possamos discutir os principais achados e compartilhar os resultados de maneira gráfica com a comunidade acadêmica.

O método de inferência filogenética a ser utilizado deve ser analisado cuidadosamente. A história evolutiva das micoplasmas incluídas no presente estudo foi inferida a partir do método de máxima verossimilhança, mas no *pipeline* desenvolvido também está implementado o método por inferência bayesiana. A análise com este método probabilístico alternativo está em andamento e seus resultados serão comparados aos do PhyML para o mesmo conjunto de dados. A divergência ou coerência entre esses resultados levará a uma análise referente à confiabilidade nas relações evolutivas identificadas entre os micro-organismos e permitirá a análise da presença de politomias e dos grupos monofiléticos encontrados em ambas árvores, possibilitando uma discussão mais ampla acerca dos resultados do trabalho.

Ainda com relação à análise filogenética, seria interessante e informativa a inclusão de um grupo externo, como *Bacillus subtilis*, que foi utilizado por GUPTA et al., KAMMINGA et al. e CITTI et al. No entanto, a escolha de um grupo externo deverá ser avaliada com cuidado, já que a inclusão de um grupo muito distantemente relacionado a micoplasmas poderá levar à identificação de um número muito pequeno de genes ortólogos presentes em todos as espécies consideradas. Nesse caso, será importante considerar para fins de análise filogenômica a inclusão de todos os GOs identificados a fim de que não percamos sinal filogenético. Esses passos e estratégias adicionais contribuirão para a compreensão da história evolutiva de micoplasmas.

Nosso estudo de caso foi baseado na análise e inferência evolutiva do relacionamento entre 89 genomas pertencentes a 83 espécies diferentes de micoplasmas. O fato de serem bactérias com pequeno tamanho de genoma as torna ideal como conjunto de teste do nosso *pipeline* inicial. Além disso, em virtude de haver estudos recentes publicados na literatura demonstrando o

relacionamento evolutivo entre diferentes espécies de micoplasmas, tivemos a oportunidade de analisar de forma comparativa os resultados obtidos.

É importante ressaltar que a maioria das filogenias obtidas com o uso do marcador molecular 16S e baseadas em informações obtidas em bancos de dados biológicos acurados, como o SILVA (QUAST et al., 2012), não utiliza algumas cepas denominadas *Candidatus Mycoplasma* spp na análise. Nesses casos, para fins de comparação com nossos resultados foi necessário utilizarmos estudos bem recentes (GUPTA et al., 2018a) e que não foram baseados exclusivamente no tradicional marcador molecular 16S. Isso demonstra também que ainda há discussões abertas na literatura a respeito da classificação filogenética e taxonômica de micoplasmas e reforça a necessidade da abordagem aqui desenvolvida, bem como o desenvolvimento de *pipelines* capazes de rapidamente analisar a crescente quantidade de dados genômicos sendo disponibilizados nos bancos de dados.

Essa análise comparativa de árvores filogenéticas obtidas em diferentes estudos, por sua vez, é um processo minucioso, principalmente se realizado para muitas OTUs. Alguns *softwares* podem ser usados para inferir a similaridade das relações obtidas por cada uma das árvores. O programa Meta-Tree (NYE, 2008), por exemplo, oferece uma estimativa baseada em distância simples do tamanho dos ramos entre grupos topologicamente iguais. Futuramente, a fim de apresentarmos dados qualitativos melhor embasados sobre as árvores, programas como o Meta-Tree serão implementados em nosso *pipeline* a fim de reduzir a possibilidade de erro na análise comparativa realizada de maneira manual. Isso levará também a uma redução no tempo de análise.

Ainda sobre os métodos de inferência filogenética é muito relevante questionarmos se são os mais coerentes para os organismos para os quais se deseja inferir a hipótese evolutiva. Bactérias do gênero *Mycoplasma*, por exemplo, possuem clados bem caracterizados quanto à ocorrência de THG e regiões de recombinação, como já descrito na fundamentação teórica deste trabalho. O grupo Mycoides, apesar de ser monofilético em todas as árvores obtidas nesse trabalho, e em demais estudos da literatura, possui conteúdo gênico semelhante a outros gêneros, o que demonstra ser uma relação importante para a patogenicidade desses micro-organismos (LO et al., 2018).

Entretanto, tais eventos de recombinação podem ser um problema para o alinhamento e, por consequência, para todo o seguimento do trabalho. Algumas ferramentas, como as descritas por BROMBERG et al. e FAN et al. foram desenvolvidas para lidar exatamente com as regiões recombinantes dos genomas. Importante ressaltarmos que regiões recombinantes podem levar a

conclusões a respeito do relacionamento evolutivo entre as diferentes espécies que não condizem com sua verdadeira história evolutiva. Dessa forma, quando estamos estudando procariotos, em especial, precisamos estar atentos à avaliação de eventos de recombinação que possam ter ocorrido. Não há na literatura um estudo filogenômico que utilize essas técnicas de avaliação de recombinação para micoplasmas. Assim, até o momento, não sabemos em detalhe como esses aspectos afetam a evolução dessas espécies e o quanto podem levar a confusões na inferência de suas histórias evolutivas. Isso configura, por fim, mais um cenário que também poderá ser explorado em relação à evolução desses micro-organismos em estudos futuros.

5 CONCLUSÕES

Compreender as relações evolutivas entre os organismos revela muito sobre sua ecologia e adaptação, bem como ajuda a entender características específicas, como os fatores que levam à patogenicidade, por exemplo. Desde antes de Charles Darwin, mundialmente conhecido por sua teoria da evolução, buscava-se delinear e levantar hipóteses sobre esses processos. Em relação a organismos procarióticos, tais estudos receberam um estimulo muito grande com o advento de novas técnicas de sequenciamento, que permitiram a disponibilização de uma quantidade massiva de dados a serem explorados.

Avaliar bactérias do gênero *Mycoplasma*, por sua vez, mostra-se relevante devido a uma série de fatores que dificultam a distinção entre elas. Esses organismos, de genoma bem pequeno, são um excelente estudo piloto na implementação de um *pipeline* a fim de validá-lo, permitindo comparar resultados anteriores com os encontrados no estudo em questão e, consequentemente, expandindo o conhecimento sobre tais organismos. Sendo assim, o presente estudo caracterizou evolutivamente o gênero e resultou em um *pipeline* a ser utilizado também para outros conjuntos de dados.

As considerações finais desse trabalho estão alinhadas com os objetivos propostos, a saber que o desenvolvimento de *software* seguindo boas práticas de programação e de gerenciamento de projeto, juntamente à utilização de bases científicas na resolução dos problemas relacionados à biologia é positivo e auxilia no processo de testes de hipóteses evolutivas, bem como permite a continuidade dos projetos por diferentes pesquisadores.

A filogenômica, tema norteador das discussões aqui desenvolvidas, mostrou ser uma estratégia que carece de padrões a serem utilizados na automatização do processo de análise, o que não deve ser visto de forma negativa, mas sim como um aspecto a ser explorado no sentido de desenvolver *pipelines* que permitam a adaptação a todos os estágios do processo a fim de particularizar a análise para cada conjunto de dados. Dados esses que, quando avaliados e aprimorados por algoritmos de filtro na etapa de alinhamento, tendem a ter um resultado positivo sobre a inferência filogenética. Entretanto, como observado, é necessário conhecer bem as sequências, o que representam na biologia do organismo e qual seu nível de conservação, identificando a presença ou ausência de sinal filogenético ou presença de muito ruído, por exemplo.

Nosso estudo também permitiu concluir que a maioria das espécies de micoplasmas são posicionadas corretamente nos clados se utilizados poucos *clusters*. Entretanto, isso se reflete no

posicionamento de algumas OTUs, mostrando ser necessário incorporar mais genes ortólogos na análise. É interessante ressaltar que muitos dos grupos encontrados nas árvores tiveram suas relações mais próximas mantidas.

Por fim, há uma série de melhorias que podem ser implementadas no fluxo de dados e inferência, módulos a serem desenvolvidos, novas ferramentas para serem testadas e, principalmente, explorar o uso do *pipeline* em outros organismos, a fim de verificar os resultados e adaptação dos parâmetros em outros conjuntos de dados. Considerando o atual estado da arte e as melhorias a serem futuramente incorporadas, a aplicação aqui desenvolvida pode contribuir para o crescimento da área ao simplificar a sistemática da inferência evolutiva e possibilitas o uso por outros pesquisadores que não são tão familiarizados com o desenvolvimento desse tipo de aplicação.

REFERÊNCIAS

ABASCAL, F.; ZARDOYA, R.; POSADA, D. Prottest: selection of best-fit models of protein evolution. **Bioinformatics**, Oxford University Press, v. 21, n. 9, p. 2104–2105, 2005.

ÁLVAREZ-JARRETA, J.; RUIZ-PESINI, E. Mevolib v1. 0: the first molecular evolution library for python. **BMC bioinformatics**, BioMed Central, v. 17, n. 1, p. 436, 2016.

ALVAREZ-PONCE, D.; WEITZMAN, C. L.; TILLETT, R. L.; SANDMEIER, F. C.; TRACY, C. R. High quality draft genome sequences of mycoplasma agassizii strains ps6 t and 723 isolated from gopherus tortoises with upper respiratory tract disease. **Standards in genomic sciences**, BioMed Central, v. 13, n. 1, p. 12, 2018.

AMSTUTZ, P.; ANDEER, R.; CHAPMAN, B.; CHILTON, J.; CRUSOE, M. R.; GUIMERA, R. V.; HERNANDEZ, G. C.; IVKOVIC, S.; KARTASHOV, A.; KERN, J. et al. Common workflow language, draft 3. 2016.

BAICHOO, S.; OUZOUNIS, C. A. Computational complexity of algorithms for sequence comparison, short-read assembly and genome alignment. **Biosystems**, Elsevier, v. 156, p. 72–85, 2017.

BARATE, A. K.; CHO, Y.; TRUONG, Q. L.; HAHN, T.-W. Immunogenicity of ims 1113 plus soluble subunit and chimeric proteins containing mycoplasma hyopneumoniae p97 c-terminal repeat regions. **FEMS microbiology letters**, Blackwell Publishing Ltd Oxford, UK, v. 352, n. 2, p. 213–220, 2014.

BARYKOVA, Y. A.; LOGUNOV, D. Y.; SHMAROV, M. M.; VINAROV, A. Z.; FIEV, D. N.; VINAROVA, N. A.; RAKOVSKAYA, I. V.; BAKER, P. S.; SHYSHYNOVA, I.; STEPHENSON, A. J. et al. Association of mycoplasma hominis infection with prostate cancer. **Oncotarget**, Impact Journals, LLC, v. 2, n. 4, p. 289, 2011.

BAYES, T.; PRICE, R.; CANTON, J. An essay towards solving a problem in the doctrine of chances. C. Davis, Printer to the Royal Society of London London, U. K, 1763.

BERČIČ, R. L.; SLAVEC, B.; LAVRIČ, M.; NARAT, M.; ZORMAN-ROJS, O.; DOVČ, P.; BENČINA, D. A survey of avian mycoplasma species for neuraminidase enzymatic activity. **Veterinary microbiology**, Elsevier, v. 130, n. 3-4, p. 391–397, 2008.

BROCCHIERI, L. Phylogenetic inferences from molecular sequences: review and critique. **Theoretical population biology**, Elsevier, v. 59, n. 1, p. 27–40, 2001.

BROMBERG, R.; GRISHIN, N. V.; OTWINOWSKI, Z. Phylogeny reconstruction with alignment-free method that corrects for horizontal gene transfer. **PLoS computational biology**, Public Library of Science, v. 12, n. 6, p. e1004985, 2016.

BROWN, D.; SJÖLANDER, K. Functional classification using phylogenomic inference. **PLoS computational biology**, Public Library of Science, v. 2, n. 6, p. e77, 2006.

BROWN, D. R. Phylum xvi. tenericutes murray 1984a, 356 vp (effective publication: Murray 1984b, 33.). In: **Bergey's Manual® of Systematic Bacteriology**. [S.l.]: Springer, 2010. p. 567–723.

BROWN, J. W.; WALKER, J. F.; SMITH, S. A. Phyx: phylogenetic tools for unix. **Bioinformatics**, Oxford University Press, v. 33, n. 12, p. 1886–1888, 2017.

BROWN, T. A. How genomes evolve. In: **Genomes. 2nd edition**. [S.l.]: Wiley-Liss, 2002.

BROWN, T. A. Molecular phylogenetics. Wiley-Liss, 2002.

BULT, C. J.; WHITE, O.; OLSEN, G. J.; ZHOU, L.; FLEISCHMANN, R. D.; SUTTON, G. G.; BLAKE, J. A.; FITZGERALD, L. M.; CLAYTON, R. A.; GOCAYNE, J. D. et al. Complete genome sequence of the methanogenic archaeon, methanococcus jannaschii. **Science**, American Association for the Advancement of Science, v. 273, n. 5278, p. 1058–1073, 1996.

CAPELLA-GUTIÉRREZ, S.; SILLA-MARTÍNEZ, J. M.; GABALDÓN, T. trimal: a tool for automated alignment trimming in large-scale phylogenetic analyses. **Bioinformatics**, Oxford University Press, v. 25, n. 15, p. 1972–1973, 2009.

CATTANI, A. M. Elementos repetitivos na regulação da transcrição de mycoplasma hyopneumoniae. 2016.

CHOI, S. C.; HOBOLTH, A.; ROBINSON, D. M.; KISHINO, H.; THORNE, J. L. Quantifying the impact of protein tertiary structure on molecular evolution. **Molecular biology and evolution**, Oxford University Press, v. 24, n. 8, p. 1769–1782, 2007.

CHRISTO, P. P.; SILVA, J. S. P. d.; WERNECK, I. V.; DIAS, S. L. Rhombencephalitis possibly caused by mycoplasma pneumoniae. **Arquivos de neuro-psiquiatria**, SciELO Brasil, v. 68, n. 4, p. 656–658, 2010.

CITTI, C.; DORDET-FRISONI, E.; NOUVEL, L.; KUO, C.; BARANOWSKI, E. Horizontal gene transfers in mycoplasmas (mollicutes). **Current issues in molecular biology**, v. 29, p. 3–22, 2018.

CONSORTIUM, I. H. G. S. et al. Initial sequencing and analysis of the human genome. **Nature**, Nature Publishing Group, v. 409, n. 6822, p. 860, 2001.

CRACRAFT, J. The seven great questions of systematic biology: an essential foundation for conservation and the sustainable use of biodiversity. **Annals of the Missouri Botanical Garden**, JSTOR, p. 127–144, 2002.

CROWGEY, E. L. **Applied genomics: development of bioinformatics pipelines for analyzing clinical pediatric genomic data**. Tese (Doutorado) — University of Delaware, 2016.

CURRAT, M.; GERBAULT, P.; DI, D.; NUNES, J. M.; SANCHEZ-MAZAS, A. Forward-in-time, spatially explicit modeling software to simulate genetic lineages under selection. **Evolutionary Bioinformatics**, SAGE Publications Sage UK: London, England, v. 11, p. EBO–S33488, 2015.

DARRIBA, D.; FLOURI, T.; STAMATAKIS, A. The state of software for evolutionary biology. **Molecular biology and evolution**, Oxford University Press, v. 35, n. 5, p. 1037–1046, 2018.

DELSUC, F.; BRINKMANN, H.; PHILIPPE, H. Phylogenomics and the reconstruction of the tree of life. **Nature Reviews Genetics**, Nature Publishing Group, v. 6, n. 5, p. 361, 2005.

DIJK, E. L. V.; AUGER, H.; JASZCZYSZYN, Y.; THERMES, C. Ten years of next-generation sequencing technology. **Trends in genetics**, Elsevier, v. 30, n. 9, p. 418–426, 2014.

DUNN, C. W.; HOWISON, M.; ZAPATA, F. Agalma: an automated phylogenomics workflow. **BMC bioinformatics**, BioMed Central, v. 14, n. 1, p. 330, 2013.

EDGAR, R. C. Muscle: multiple sequence alignment with high accuracy and high throughput. **Nucleic acids research**, Oxford University Press, v. 32, n. 5, p. 1792–1797, 2004.

EFRON, B. **The jackknife, the bootstrap, and other resampling plans**. [S.l.]: Siam, 1982. v. 38.

EISEN, J. A.; FRASER, C. M. Phylogenomics: intersection of evolution and genomics. **Science**, The American Association for the Advancement of Science, v. 300, n. 5626, p. 1706, 2003.

FAN, H.; IVES, A. R.; SURGET-GROBA, Y.; CANNON, C. H. An assembly and alignment-free method of phylogeny reconstruction from next-generation sequencing data. **BMC genomics**, BioMed Central, v. 16, n. 1, p. 522, 2015.

FELSENSTEIN, J. Evolutionary trees from dna sequences: a maximum likelihood approach. **Journal of molecular evolution**, Springer, v. 17, n. 6, p. 368–376, 1981.

FELSENSTEIN, J. Distance methods for inferring phylogenies: a justification. **Evolution**, Wiley Online Library, v. 38, n. 1, p. 16–24, 1984.

FELSENSTEIN, J. Confidence limits on phylogenies: an approach using the bootstrap. **Evolution**, Wiley Online Library, v. 39, n. 4, p. 783–791, 1985.

FELSENSTEIN, J.; FELENSTEIN, J. **Inferring phylogenies**. [S.l.]: Sinauer associates Sunderland, MA, 2004. v. 2.

FLEISCHMANN, R. D.; ADAMS, M. D.; WHITE, O.; CLAYTON, R. A.; KIRKNESS, E. F.; KERLAVAGE, A. R.; BULT, C. J.; TOMB, J.-F.; DOUGHERTY, B. A.; MERRICK, J. M. et al. Whole-genome random sequencing and assembly of haemophilus influenzae rd. **Science**, American Association for the Advancement of Science, v. 269, n. 5223, p. 496–512, 1995.

FREED, E. F.; WINKLER, J. D.; WEISS, S. J.; GARST, A. D.; MUTALIK, V. K.; ARKIN, A. P.; KNIGHT, R.; GILL, R. T. Genome-wide tuning of protein expression levels to rapidly engineer microbial traits. **ACS synthetic biology**, ACS Publications, v. 4, n. 11, p. 1244–1253, 2015.

FRIIS, N.; FEENSTRA, A. Mycoplasma hyorhinis in the etiology of serositis among piglets. **Acta Veterinaria Scandinavica**, v. 35, n. 1, p. 93–98, 1994.

GABALDÓN, T. Evolution of proteins and proteomes: a phylogenetics approach. **Evolutionary Bioinformatics**, SAGE Publications Sage UK: London, England, v. 1, p. 117693430500100004, 2005.

GENTLEMAN, R. **R programming for bioinformatics**. [S.l.]: Chapman and Hall/CRC, 2008.

GOFFEAU, A.; BARRELL, B. G.; BUSSEY, H.; DAVIS, R.; DUJON, B.; FELDMANN, H.; GALIBERT, F.; HOHEISEL, J.; JACQ, C.; JOHNSTON, M. et al. Life with 6000 genes. **Science**, American Association for the Advancement of Science, v. 274, n. 5287, p. 546–567, 1996.

GRANT, J. R.; KATZ, L. A. Building a phylogenomic pipeline for the eukaryotic tree of life-addressing deep phylogenies with genome-scale data. **PLoS currents**, Public Library of Science, v. 6, 2014.

GUIMARAES, A. M.; SANTOS, A. P.; NASCIMENTO, N. C. do; TIMENETSKY, J.; MESSICK, J. B. Comparative genomics and phylogenomics of hemotrophic mycoplasmas. **PloS one**, Public Library of Science, v. 9, n. 3, p. e91445, 2014.

GUINDON, S.; DUFAYARD, J.-F.; LEFORT, V.; ANISIMOVA, M.; HORDIJK, W.; GASCUEL, O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of phyml 3.0. **Systematic biology**, Oxford University Press, v. 59, n. 3, p. 307–321, 2010.

GUPTA, R. S.; SAWNANI, S.; ADEOLU, M.; ALNAJAR, S.; OREN, A. Correction to: Phylogenetic framework for the phylum tenericutes based on genome sequence data: proposal for the creation of a new order mycoplasmoidales ord. nov., containing two new families mycoplasmoidaceae fam. nov. and metamycoplasmataceae fam. nov. harbouring eperythrozoon, ureaplasma and five novel genera. **Antonie van Leeuwenhoek**, 2018.

GUPTA, R. S.; SON, J.; OREN, A. A phylogenomic and molecular markers based taxonomic framework for members of the order entomoplasmatales: proposal for an emended order mycoplasmatales containing the family spiroplasmataceae and emended family mycoplasmataceae comprised of six genera. **Antonie van Leeuwenhoek**, Springer, p. 1–28, 2018.

HANNAN, P. Comparative susceptibilities of various aids-associated and human urogenital tract mycoplasmas and strains of mycoplasma pneumoniae to 10 classes of antimicrobial agent in vitro. **Journal of medical microbiology**, Microbiology Society, v. 47, n. 12, p. 1115–1122, 1998.

HASEGAWA, M.; KISHINO, H.; YANO, T.-a. Dating of the human-ape splitting by a molecular clock of mitochondrial dna. **Journal of molecular evolution**, Springer, v. 22, n. 2, p. 160–174, 1985.

HIMMELREICH, R.; HILBERT, H.; PLAGENS, H.; PIRKL, E.; LI, B.-C.; HERRMANN, R. Complete sequence analysis of the genome of the bacterium mycoplasma pneumoniae. **Nucleic acids research**, Oxford University Press, v. 24, n. 22, p. 4420–4449, 1996.

HODGE, T.; JAMIE, M.; COPE, T. A myosin family tree. **Journal of cell science**, The Company of Biologists Ltd, v. 113, n. 19, p. 3353–3354, 2000.

HUELSENBECK, J. P.; RONQUIST, F.; NIELSEN, R.; BOLLBACK, J. P. Bayesian inference of phylogeny and its impact on evolutionary biology. **science**, American Association for the Advancement of Science, v. 294, n. 5550, p. 2310–2314, 2001.

HUERTA-CEPAS, J.; DOPAZO, J.; GABALDÓN, T. Ete: a python environment for tree exploration. **BMC bioinformatics**, BioMed Central, v. 11, n. 1, p. 24, 2010.

INAMINE, J.; HO, K.-C.; LOECHEL, S.; HU, P.-C. Evidence that uga is read as a tryptophan codon rather than as a stop codon by mycoplasma pneumoniae, mycoplasma genitalium, and mycoplasma gallisepticum. **Journal of bacteriology**, Am Soc Microbiol, v. 172, n. 1, p. 504–506, 1990.

ITAN, Y.; GERBAULT, P.; PINES, G. **Evolutionary Genomics: Supplement Aims and Scope**. [S.l.]: SAGE Publications Sage UK: London, England, 2015.

JUKES, T. H.; CANTOR, C. R. et al. Evolution of protein molecules. **Mammalian protein metabolism**, New York, v. 3, n. 21, p. 132, 1969.

JUNIER, T.; ZDOBNOV, E. M. The newick utilities: high-throughput phylogenetic tree processing in the unix shell. **Bioinformatics**, Oxford University Press, v. 26, n. 13, p. 1669–1670, 2010.

KAMMINGA, T.; KOEHORST, J. J.; VERMEIJ, P.; SLAGMAN, S.-J.; SANTOS, V. A. Martins dos; BIJLSMA, J. J.; SCHAAP, P. J. Persistence of functional protein domains in mycoplasma species and their role in host specificity and synthetic minimal life. **Frontiers in cellular and infection microbiology**, Frontiers, v. 7, p. 31, 2017.

KANNAN, T.; BASEMAN, J. B. Adp-ribosylating and vacuolating cytotoxin of mycoplasma pneumoniae represents unique virulence determinant among bacterial pathogens. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 103, n. 17, p. 6724–6729, 2006.

KATZ, L. S.; GRISWOLD, T.; WILLIAMS-NEWKIRK, A. J.; WAGNER, D.; PETKAU, A.; SIEFFERT, C.; DOMSELAAR, G. V.; DENG, X.; CARLETON, H. A. A comparative analysis of the lyve-set phylogenomics pipeline for genomic epidemiology of foodborne pathogens. **Frontiers in microbiology**, Frontiers, v. 8, p. 375, 2017.

KIMURA, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. **Journal of Molecular Evolution**, v. 16, n. 2, p. 111–120, Jun 1980. ISSN 1432-1432. Disponível em: <https://doi.org/10.1007/BF01731581>.

KIMURA, M. Estimation of evolutionary distances between homologous nucleotide sequences. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 78, n. 1, p. 454–458, 1981.

KOBISCH, M.; FRIIS, N. Swine mycoplasmoses. **Revue Scientifique et Technique-Office International des Epizooties**, Paris: L'Office, 1982-, v. 15, n. 4, p. 1569–1614, 1996.

KUMAR, S.; STECHER, G.; PETERSON, D.; TAMURA, K. Mega-cc: computing core of molecular evolutionary genetics analysis program for automated and iterative data analysis. **Bioinformatics**, Oxford University Press, v. 28, n. 20, p. 2685–2686, 2012.

KUMAR, S.; STECHER, G.; TAMURA, K. Mega7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. **Molecular biology and evolution**, Society for Molecular Biology and Evolution, v. 33, n. 7, p. 1870–1874, 2016.

LAING, C. R.; WHITESIDE, M. D.; GANNON, V. P. Pan-genome analyses of the species salmonella enterica, and identification of genomic markers predictive for species, subspecies, and serovar. **Frontiers in microbiology**, Frontiers, v. 8, p. 1345, 2017.

LAKE, J. A.; RIVERA, M. C. Deriving the genomic tree of life in the presence of horizontal gene transfer: conditioned reconstruction. **Molecular Biology and Evolution**, Oxford University Press, v. 21, n. 4, p. 681–690, 2004.

LE, S. Q.; GASCUEL, O. An improved general amino acid replacement matrix. **Molecular biology and evolution**, Oxford University Press, v. 25, n. 7, p. 1307–1320, 2008.

LECLERCQ, S.; DITTMER, J.; BOUCHON, D.; CORDAUX, R. Phylogenomics of "candidatus hepatoplasma crinochetorum," a lineage of mollicutes associated with noninsect arthropods. **Genome biology and evolution**, Oxford University Press, v. 6, n. 2, p. 407–415, 2014.

LEI, N.; YU, X.; LI, S.; ZENG, C.; ZOU, L.; LIAO, W.; PENG, M. Phylogeny and expression pattern analysis of tcp transcription factors in cassava seedlings exposed to cold and/or drought stress. **Scientific Reports**, Nature Publishing Group, v. 7, n. 1, p. 10016, 2017.

LEIPZIG, J. A review of bioinformatic pipeline frameworks. **Briefings in bioinformatics**, Oxford University Press, v. 18, n. 3, p. 530–536, 2017.

LI, L.; STOECKERT, C. J.; ROOS, D. S. Orthomcl: identification of ortholog groups for eukaryotic genomes. **Genome research**, Cold Spring Harbor Lab, v. 13, n. 9, p. 2178–2189, 2003.

LIU, L.; DYBVIG, K.; PANANGALA, V. S.; SANTEN, V. L. van; FRENCH, C. T. Gaa trinucleotide repeat region regulates m9/pmga gene expression in mycoplasma gallisepticum. **Infection and immunity**, Am Soc Microbiol, v. 68, n. 2, p. 871–876, 2000.

LIU, W.; FANG, L.; LI, M.; LI, S.; GUO, S.; LUO, R.; FENG, Z.; LI, B.; ZHOU, Z.; SHAO, G. et al. Comparative genomics of mycoplasma: analysis of conserved essential genes and diversity of the pan-genome. **PLoS One**, Public Library of Science, v. 7, n. 4, p. e35698, 2012.

LJUBIN-STERNAK, S.; MEŠTROVIĆ, T. Chlamydia trachomatis and genital mycoplasmas: pathogens with an impact on human reproductive health. **Journal of pathogens**, Hindawi, v. 2014, 2014.

LO, W.-S.; GASPARICH, G. E.; KUO, C.-H. Convergent evolution among ruminant-pathogenic mycoplasma involved extensive gene content changes. **Genome biology and evolution**, Oxford University Press, v. 10, n. 8, p. 2130–2139, 2018.

LU, H.; GIORDANO, F.; NING, Z. Oxford nanopore minion sequencing and genome assembly. **Genomics, proteomics & bioinformatics**, Elsevier, v. 14, n. 5, p. 265–279, 2016.

LYSNYANSKY, I.; ROSENGARTEN, R.; YOGEV, D. Phenotypic switching of variable surface lipoproteins in mycoplasma bovis involves high-frequency chromosomal rearrangements. **Journal of bacteriology**, Am Soc Microbiol, v. 178, n. 18, p. 5395–5401, 1996.

MA, L.; JENSEN, J. S.; MANCUSO, M.; HAMASUNA, R.; JIA, Q.; MCGOWIN, C. L.; MARTIN, D. H. Genetic variation in the complete mgpa operon and its repetitive chromosomal elements in clinical strains of mycoplasma genitalium. **PLoS One**, Public Library of Science, v. 5, n. 12, p. e15660, 2010.

MADDISON, W. P. Mesquite: a modular system for evolutionary analysis. **Evolution**, v. 62, p. 1103–1118, 2008.

MAKIMURA, K.; TAMURA, Y.; MOCHIZUKI, T.; HASEGAWA, A.; TAJIRI, Y.; HANAZAWA, R.; UCHIDA, K.; SAITO, H.; YAMAGUCHI, H. Phylogenetic classification and species identification of dermatophyte strains based on dna sequences of nuclear ribosomal internal transcribed spacer 1 regions. **Journal of clinical microbiology**, Am Soc Microbiol, v. 37, n. 4, p. 920–924, 1999.

MANILOFF, J. The minimal cell genome:"on being the right size". **Proceedings of the National Academy of Sciences of the United States of America**, National Academy of Sciences, v. 93, n. 19, p. 10004, 1996.

MARE, C. Mycoplasma hyopneumoniae, a causative agent of virus pig pneumonia. **Vet. Med.**, v. 60, p. 841–845, 1965.

MARTINEZ-URTAZA, J.; AERLE, R. van; ABANTO, M.; HAENDIGES, J.; MYERS, R. A.; TRINANES, J.; BAKER-AUSTIN, C.; GONZALEZ-ESCALONA, N. Genomic variation and evolution of vibrio parahaemolyticus st36 over the course of a transcontinental epidemic expansion. **MBio**, Am Soc Microbiol, v. 8, n. 6, p. e01425–17, 2017.

MAVEDZENGE, S. N.; WEISS, H. A. Association of mycoplasma genitalium and hiv infection: a systematic review and meta-analysis. **Aids**, LWW, v. 23, n. 5, p. 611–620, 2009.

MAY, M.; BALISH, M. F.; BLANCHARD, A. The order mycoplasmatales. In: **The Prokaryotes**. [S.l.]: Springer, 2014. p. 515–550.

MEYLING, A.; FRIIS, N. Serological identification of a new porcine mycoplasma species, m. flocculare. **Acta Veterinaria Scandinavica**, v. 13, n. 2, p. 287, 1972.

MOROWITZ, H. J.; WALLACE, D. C. Genome size and life cycle of the mycoplasma. **Annals of the New York Academy of Sciences**, Wiley Online Library, v. 225, n. 1, p. 62–73, 1973.

MORRISON, D. A. Evolutionary genomics: Statistical and computational methods. volumes 1 and 2. – edited by maria anisimova. **Systematic Biology**, Oxford University Press, v. 62, 03 2013. Disponível em: <http://gen.lib.rus.ec/scimag/index.php?s=10.1093/sysbio/sys089>.

MUSATOVOVA, O.; KANNAN, T.; BASEMAN, J. B. Mycoplasma pneumoniae large dna repetitive elements repmp1 show type specific organization among strains. **PLoS One**, Public Library of Science, v. 7, n. 10, p. e47625, 2012.

MUTO, A.; OSAWA, S. The guanine and cytosine content of genomic dna and bacterial evolution. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 84, n. 1, p. 166–169, 1987.

NEIMARK, H. Origin and evolution of wall-less prokaryotes. **The bacterial L-forms**, Marcel Dekkar Inc, p. 21–42, 1986.

NIELSEN, R. **Statistical methods in molecular evolution**. [S.l.]: Springer, 2006.

NYE, T. M. Trees of trees: an approach to comparing multiple alternative phylogenies. **Systematic biology**, Taylor & Francis, v. 57, n. 5, p. 785–794, 2008.

OSHIMA, K.; NISHIDA, H. Phylogenetic relationships among mycoplasmas based on the whole genomic information. **Journal of Molecular Evolution**, Springer, v. 65, n. 3, p. 249–258, 2007.

PARADIS, E.; CLAUDE, J.; STRIMMER, K. Ape: analyses of phylogenetics and evolution in r language. **Bioinformatics**, Oxford University Press, v. 20, n. 2, p. 289–290, 2004.

PEER, Y. Van de; WACHTER, R. D. Treecon: a software package for the construction and drawing of evolutionary trees. **Computer Applications in the Biosciences**, Citeseer, v. 9, n. 2, p. 177–182, 1993.

PETERS, R. S.; MEYER, B.; KROGMANN, L.; BORNER, J.; MEUSEMANN, K.; SCHÜTTE, K.; NIEHUIS, O.; MISOF, B. The taming of an impossible child: a standardized all-in approach to the phylogeny of hymenoptera using public database sequences. **BMC biology**, BioMed Central, v. 9, n. 1, p. 55, 2011.

PEVSNER, J. **Bioinformatics and functional genomics**. [S.l.]: John Wiley & Sons, 2015.

PIEL, W. H.; VOS, R. A. Treebasedmp: A toolkit for phyloinformatic research. **bioRxiv**, Cold Spring Harbor Laboratory, p. 399030, 2018.

POSADA, D. Using modeltest and paup* to select a model of nucleotide substitution. **Current protocols in bioinformatics**, Wiley Online Library, n. 1, p. 6–5, 2003.

POSADA, D. jmodeltest: phylogenetic model averaging. **Molecular biology and evolution**, Oxford University Press, v. 25, n. 7, p. 1253–1256, 2008.

PRICE, M. N.; DEHAL, P. S.; ARKIN, A. P. Fasttree 2–approximately maximum-likelihood trees for large alignments. **PloS one**, Public Library of Science, v. 5, n. 3, p. e9490, 2010.

QUAST, C.; PRUESSE, E.; YILMAZ, P.; GERKEN, J.; SCHWEER, T.; YARZA, P.; PEPLIES, J.; GLÖCKNER, F. O. The silva ribosomal rna gene database project: improved data processing and web-based tools. **Nucleic acids research**, Oxford University Press, v. 41, n. D1, p. D590–D596, 2012.

RAGAN, M. A. Phylogenetic inference based on matrix representation of trees. **Molecular phylogenetics and evolution**, Elsevier, v. 1, n. 1, p. 53–58, 1992.

RAZIN, S. The genus mycoplasma and related genera (class mollicutes). In: **The Prokaryotes**. [S.l.]: Springer, 2006. p. 836–904.

RAZIN, S.; HAYFLICK, L. Highlights of mycoplasma research—an historical perspective. **Biologicals**, Elsevier, v. 38, n. 2, p. 183–190, 2010.

RAZIN, S.; YOGEV, D.; NAOT, Y. Molecular biology and pathogenicity of mycoplasmas. **Microbiology and molecular biology reviews**, Am Soc Microbiol, v. 62, n. 4, p. 1094–1156, 1998.

ROBBERTSE, B.; YODER, R. J.; BOYD, A.; REEVES, J.; SPATAFORA, J. W. Hal: an automated pipeline for phylogenetic analyses of genomic data. **PLoS currents**, Public Library of Science, v. 3, 2011.

ROBINSON, D. M.; JONES, D. T.; KISHINO, H.; GOLDMAN, N.; THORNE, J. L. Protein evolution with dependence among codons due to tertiary structure. **Molecular Biology and Evolution**, Oxford University Press, v. 20, n. 10, p. 1692–1704, 2003.

ROCHA, E. P.; BLANCHARD, A. Genomic repeats, genome plasticity and the dynamics of mycoplasma evolution. **Nucleic acids research**, Oxford University Press, v. 30, n. 9, p. 2031–2042, 2002.

ROKAS, A.; WILLIAMS, B. L.; KING, N.; CARROLL, S. B. Genome-scale approaches to resolving incongruence in molecular phylogenies. **Nature**, Nature Publishing Group, v. 425, n. 6960, p. 798, 2003.

RONQUIST, F.; HUELSENBECK, J. P. Mrbayes 3: Bayesian phylogenetic inference under mixed models. **Bioinformatics**, Oxford University Press, v. 19, n. 12, p. 1572–1574, 2003.

ROSALES, R. S.; PULEIO, R.; LORIA, G. R.; CATANIA, S.; NICHOLAS, R. A. Mycoplasmas: Brain invaders? **Research in veterinary science**, Elsevier, 2017.

ROSENGARTEN, R.; CITTI, C.; GLEW, M.; LISCHEWSKI, A.; DROESSE, M.; MUCH, P.; WINNER, F.; BRANK, M.; SPERGSER, J. Host-pathogen interactions in mycoplasma pathogenesis: virulence and survival strategies of minimalist prokaryotes. **International journal of medical microbiology**, Elsevier, v. 290, n. 1, p. 15–25, 2000.

ROSSUM, G. V. et al. Python programming language. In: **USENIX Annual Technical Conference**. [S.l.: s.n.], 2007. v. 41, p. 36.

ROTTEM, S. Interaction of mycoplasmas with host cells. **Physiological reviews**, American Physiological Society Bethesda, MD, v. 83, n. 2, p. 417–432, 2003.

SAITOU, N.; NEI, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. **Molecular biology and evolution**, v. 4, n. 4, p. 406–425, 1987.

SALEMI, M.; VANDAMME, A.-M.; LEMEY, P. **The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing**. [S.l.]: Cambridge University Press, 2009.

SANGER, F.; COULSON, A.; FRIEDMANN, T.; AIR, G.; BARRELL, B.; BROWN, N.; FIDDES, J.; III, C. H.; SLOCOMBE, P.; SMITH, M. The nucleotide sequence of bacteriophage φx174. **Journal of molecular biology**, Elsevier, v. 125, n. 2, p. 225–246, 1978.

SASAKI, Y.; ISHIKAWA, J.; YAMASHITA, A.; OSHIMA, K.; KENRI, T.; FURUYA, K.; YOSHINO, C.; HORINO, A.; SHIBA, T.; SASAKI, T. et al. The complete genomic sequence of mycoplasma penetrans, an intracellular bacterial pathogen in humans. **Nucleic acids research**, Oxford University Press, v. 30, n. 23, p. 5293–5300, 2002.

SELA, I.; ASHKENAZY, H.; KATOH, K.; PUPKO, T. Guidance2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. **Nucleic Acids Research**, Oxford University Press, v. 43, n. W1, p. W7–W14, 2015.

SHARP, P. M.; STENICO, M.; PEDEN, J. F.; LLOYD, A. T. **Codon usage: mutational bias, translational selection, or both?** [S.l.]: Portland Press Limited, 1993.

SIMMONS, W. L.; DENISON, A. M.; DYBVIG, K. Resistance of mycoplasma pulmonis to complement lysis is dependent on the number of vsa tandem repeats: shield hypothesis. **Infection and immunity**, Am Soc Microbiol, v. 72, n. 12, p. 6846–6851, 2004.

SIQUEIRA, F. M.; THOMPSON, C. E.; VIRGINIO, V. G.; GONCHOROSKI, T.; REOLON, L.; ALMEIDA, L. G.; FONSÊCA, M. M. da; SOUZA, R. de; PROSDOCIMI, F.; SCHRANK, I. S. et al. New insights on the biology of swine respiratory tract mycoplasmas from a comparative genome analysis. **BMC genomics**, BioMed Central, v. 14, n. 1, p. 175, 2013.

SIRAND-PUGNET, P.; CITTI, C.; BARRÉ, A.; BLANCHARD, A. Evolution of mollicutes: down a bumpy road with twists and turns. **Research in microbiology**, Elsevier, v. 158, n. 10, p. 754–766, 2007.

SMITH, S. A.; BEAULIEU, J. M.; DONOGHUE, M. J. Mega-phylogeny approach for comparative biology: an alternative to supertree and supermatrix approaches. **BMC evolutionary biology**, BioMed Central, v. 9, n. 1, p. 37, 2009.

SMITH, S. A.; DUNN, C. W. Phyutility: a phyloinformatics tool for trees, alignments and molecular data. **Bioinformatics**, Oxford University Press, v. 24, n. 5, p. 715–716, 2008.

SOKAL, R. R. A statistical method for evaluating systematic relationship. **University of Kansas science bulletin**, v. 28, p. 1409–1438, 1958.

STAJICH, J. E.; BLOCK, D.; BOULEZ, K.; BRENNER, S. E.; CHERVITZ, S. A.; DAGDIGIAN, C.; FUELLEN, G.; GILBERT, J. G.; KORF, I.; LAPP, H. et al. The bioperl toolkit: Perl modules for the life sciences. **Genome research**, Cold Spring Harbor Lab, v. 12, n. 10, p. 1611–1618, 2002.

STEEL, M.; LOCKHART, P.; PENNY, D. Confidence in evolutionary trees from biological sequence data. **Nature**, Nature Publishing Group, v. 364, n. 6436, p. 440, 1993.

SUKUMARAN, J.; HOLDER, M. T. Dendropy: a python library for phylogenetic computing. **Bioinformatics**, Oxford University Press, v. 26, n. 12, p. 1569–1571, 2010.

SWOFFORD, D. L. Phylogenetic analysis using parsimony. **Illinois Natural History Survey, Champaign, Illinois**, Citeseer, 1985.

TALEVICH, E.; INVERGO, B. M.; COCK, P. J.; CHAPMAN, B. A. Bio. phylo: a unified toolkit for processing, analyzing and visualizing phylogenetic trees in biopython. **BMC bioinformatics**, BioMed Central, v. 13, n. 1, p. 209, 2012.

TAMURA, K. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and g+ c-content biases. **Molecular biology and evolution**, v. 9, n. 4, p. 678–687, 1992.

TAVARÉ, S. Some probabilistic and statistical problems in the analysis of dna sequences. **Lectures on mathematics in the life sciences**, v. 17, n. 2, p. 57–86, 1986.

TENENBAUM, J. D. Translational bioinformatics: past, present, and future. **Genomics, proteomics & bioinformatics**, Elsevier, v. 14, n. 1, p. 31–41, 2016.

TOPRAK, E.; VERES, A.; MICHEL, J.-B.; CHAIT, R.; HARTL, D. L.; KISHONY, R. Evolutionary paths to antibiotic resistance under dynamically sustained drug selection. **Nature genetics**, Nature Publishing Group, v. 44, n. 1, p. 101, 2012.

TREANGEN, T. J.; ONDOV, B. D.; KOREN, S.; PHILLIPPY, A. M. The harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. **Genome biology**, BioMed Central, v. 15, n. 11, p. 524, 2014.

TSANG, A. K.; LEE, H. H.; YIU, S.-M.; LAU, S. K.; WOO, P. C. Failure of phylogeny inferred from multilocus sequence typing to represent bacterial phylogeny. **Scientific reports**, Nature Publishing Group, v. 7, n. 1, p. 4536, 2017.

TSIODRAS, S.; KELESIDIS, I.; KELESIDIS, T.; STAMBOULIS, E.; GIAMARELLOU, H. Central nervous system manifestations of mycoplasma pneumoniae infections. **Journal of Infection**, Elsevier, v. 51, n. 5, p. 343–354, 2005.

VASCONCELOS, A. T. R.; FERREIRA, H. B.; BIZARRO, C. V.; BONATTO, S. L.; CARVALHO, M. O.; PINTO, P. M.; ALMEIDA, D. F.; ALMEIDA, L. G.; ALMEIDA, R.; ALVES-FILHO, L. et al. Swine and poultry pathogens: the complete genome sequences of two strains of mycoplasma hyopneumoniae and a strain of mycoplasma synoviae. **Journal of bacteriology**, Am Soc Microbiol, v. 187, n. 16, p. 5568–5577, 2005.

VENTER, J. C.; ADAMS, M. D.; MYERS, E. W.; LI, P. W.; MURAL, R. J.; SUTTON, G. G.; SMITH, H. O.; YANDELL, M.; EVANS, C. A.; HOLT, R. A. et al. The sequence of the human genome. **science**, American Association for the Advancement of Science, v. 291, n. 5507, p. 1304–1351, 2001.

WANG, R.-H.; HAYES, M. M.; WEAR, D.; LO, S.; SHIH, J.-K.; ALTER, H.; GRANDINETTI, T.; PIERCE, P. High frequency of antibodies to mycoplasma penetrans in hiv-infected patients. **The Lancet**, Elsevier, v. 340, n. 8831, p. 1312–1316, 1992.

WANG, Z.; WU, M. A phylum-level bacterial phylogenetic marker database. **Molecular biology and evolution**, Oxford University Press, v. 30, n. 6, p. 1258–1262, 2013.

WHITAKER, J. W.; MCCONKEY, G. A.; WESTHEAD, D. R. The transferome of metabolic genes explored: analysis of the horizontal transfer of enzyme encoding genes in unicellular eukaryotes. **Genome biology**, BioMed Central, v. 10, n. 4, p. R36, 2009.

WOESE, C.; MANILOFF, J.; ZABLEN, L. Phylogenetic analysis of the mycoplasmas. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 77, n. 1, p. 494–498, 1980.

XIE, X.; YANG, M.; DING, Y.; CHEN, J. Microbial infection, inflammation and epithelial ovarian cancer. **Oncology letters**, Spandidos Publications, v. 14, n. 2, p. 1911–1919, 2017.

YOGEV, D.; WATSON-MCKOWN, R.; MCINTOSH, M. A.; WISE, K. Sequence and tnphoa analysis of a mycoplasma hyorhinis protein with membrane export function. **Journal of bacteriology**, Am Soc Microbiol, v. 173, n. 6, p. 2035–2044, 1991.

YOTOKO, K. S.; BONATTO, S. L. A phylogenomic appraisal of the evolutionary relationship of mycoplasmas. **Genetics and Molecular Biology**, SciELO Brasil, v. 30, n. 1, p. 270–276, 2007.

ZAHA, A.; FERREIRA, H. B.; PASSAGLIA, L. M. **Biologia Molecular Básica-5**. [S.l.]: Artmed Editora, 2014.

ZHANG, Q.; WISE, K. S. Molecular basis of size and antigenic variation of a mycoplasma hominis adhesin encoded by divergent vaa genes. **Infection and immunity**, Am Soc Microbiol, v. 64, n. 7, p. 2737–2744, 1996.

ZHANG, Y.-C.; LIN, K. Phylogeny inference of closely related bacterial genomes: Combining the features of both overlapping genes and collinear genomic regions. **Evolutionary Bioinformatics**, SAGE Publications Sage UK: London, England, v. 11, p. EBO–S33491, 2015.

ZHARKIKH, A. Estimation of evolutionary distances between nucleotide sequences. **Journal of molecular evolution**, Springer, v. 39, n. 3, p. 315–329, 1994.

ZUCKERKANDL, E.; PAULING, L. Molecular disease, evolution and genetic heterogeneity. Academic Press, 1962.

APÊNDICE A – Informações dos genomas de micoplasmas

Tabela 1 - Características gerais dos genomas de micoplasmas utilizados.

Espécie	Cepa	ID Montagem	Status	Tamanho (Mb)	%GC	Genes	Proteínas	Patogenicidade
Candidatus M. girerdii	VCU_M1	GCA_000770195.1	Completo	0,61898	28,6	611	572	*Vaginal Human Microbiome Project at VCU*, presente em infecções no sistema reprodutor feminino.
Candidatus M. haemobos	INIFAP01	GCA_001645765.1	Contig	0,935638	30,5	1166	1109	Patogênico em gados.
Candidatus M. haemolamae	Purdue	GCA_000281235.1	Completo	0,756845	39,3	961	925	Patógeno de Alpacas (*Vicugna pacos*) e Lhamas (*Lama glama*).
Candidatus M. haemominutum	Birmingham 1	GCA_000319365.1	Scaffold	0,51388	35,5	568	529	Comensal com baixo potencial patogênico.
M. agalactiae	PG2	GCA_000063605.1	Completo	0,877438	29,7	775	660	Agente etiológico de agalaxia contagiosa em ovinos e caprinos.
M. agassizii	PS6	GCA_002272945.1	Scaffold	1,274970	28,5	1025	965	Agente etiológico de infecções respiratórias em *Gopherus agassizii, Gopherus morafkai* e *Gopherus polyphemus.*
M. alkalescens	ATCC 29103	GCA_003208575.1	Scaffold	0,783239	25,7	671	606	Comensal em bezerros e gado mas associado a infecções como mastites e artrites.
M. alligatoris	A21JP2	GCA_000178375.1	Contig	0,97303	26,7	842	763	Patógeno em jacarés.
M. alvi	ATCC 29626	GCA_000701785.1	Scaffold	0,84064	25	730	677	Presente no trato intestinal de gados sem registro de atividade patogênica.
M. amphoriforme	A39	GCA_000723365.1_MAM001	Completo	1,029020	31,60	715	715	Ele foi encontrado em infecções respiratórias humanas e está associado à bronquite crônica em pacientes imunossuprimidos.
M. anatis	NCTC 10156	GCA_003285065.1	Completo	0,956094	26,7	832		Patógeno em patos domésticos e aves selvagens.
M. anseris	ATCC 49234	GCA_003285045.1	Completo	0,750009	26,4	658	611	Presente em aves aquáticas relacionado a infecção no sistema reprodutor de gansos mas sem comprovação de patogenicidade.
M. arginini	HAZ145_1	GCA_001547975.1	Completo	0,678592	26,4	615	569	Patogênico em uma série de mamíferos incluindo bovinos.
M. arthritidis	158L3-1	GCA_000020065.1	Completo	0,820453	30,7	665	619	Causa artrite em ratos e camundongos.
M. auris	ATCC 51348	GCA_003253435.1	Scaffold	0,750874	27,1	670	596	Comensal em bodes.
M. bovigenitalium	HAZ596	GCA_002356075.1	Completo	0,853553	29,1	746	692	Oportunista em gados.
M. bovirhinis	GS01	GCA_002688685.1	Completo	0,847985	27,6	705	645	Agente etiológico de pneumonia e mastites em bovinos.
M. bovis	PG45	GCA_000183385.1	Completo	1,003400	29,3	870	779	Causador de infecções em bovinos.
M. bovis	Hubei-1	GCA_000219375.1	Completo	0,948121	29,3	813	731	Causador da doença em bovinos e bezerros.
M. bovoculi	M165/69	GCA_000524555.1	Completo	0,76024	28,2	622	581	Agente etiológico de conjuntivite em bovinos.

M. buteonis	ATCC 51371	GCA_000733865.1	Scaffold	0,85003	27,6	706	643	Presente em urubus e patogênico em *Falco cherrug.*
M. californicum	ST-6	GCA_000695835.1	Completo	0,793841	30,8	682	630	Patogênico em gados.
M. canadense	HAZ360_1	GCA_000828855.1	Completo	0,693241	24,3	586	538	Envolvidas na mastite bovina micoplasmática.
M. canis	PG 14	GCA_001553195.1	Completo	0,897443	27,2	700	623	Patógeno oportunista em cães, humanos e bovinos.
M. capricolum subsp capricolum	ATCC 27343	GCA_000012765.1	Completo	1,010020	23,8	877	797	Patógeno em caprinos com registro de ocorrência em humanos (cepa 14DL0024).
M. cloacale	NCTC 10199	GCA_003269445.1	Completo	0,659552	27	579	539	Presente em perus. Estudos indicam potencial para causar morte embrionária.
M. collis	ATCC 35278	GCA_000701825.1	Scaffold	0,902877	22,4	823	749	Presente na conjuntiva de roedores sem registro de atividade patogênica.
M. columbinum	ATCC 29257	GCA_000712175.1	Scaffold	0,765157	26,9	650	598	Causa problemas respiratórios em pombos
M. columborale	ATCC 29258	GCA_000701845.1	Scaffold	0,910711	29,2	769	697	Presente em pombos causando infecções no trato respiratório inferior.
M. conjunctivae	HRC/581T	GCA_000026765.1	Cromossomo	0,846214	28,6	703	657	Causa conjuntivite em cabras e ovelhas
M. cricetuli	ATCC 35279	GCA_000526955.1	Contig	0,84775	24	745	667	Causador de conjuntivite em hamsters (Cricetulus griseus).
M. crocodyli	MP145	GCA_000025845.1	Completo	0,934379	27	792	746	Causa doença inflamatória multissistêmica letal aguda em hospedeiros suscetíveis
M. cynos	C142	GCA_000328725.1	Completo	0,998123	25,7	818	753	Importante agente etiológico de infecção respiratória canina.
M. dispar	ATCC 27140	GCA_000941075.1	Completo	1,08445	29	792	729	Oportunista causando pneumonia em bezerros
M. elephantis	ATCC 51980	GCA_000687815.1	Contig	0,769819	25,7	707	640	Presente nos órgãos reprodutivos de fêmeas de *Elephas maximus* e *Loxodonta aficana* com artrite, não é certa a relação entre as ocorrências.
M. felifaucium	ATCC 43428	GCA_000687775.1	Contig	0,764934	28,1	696	621	Causador de gastroenterites em pumas e chitas
M. felis	ATCC 23391	GCA_000701865.1	Scaffold	0,811556	24,3	735	660	Oportunista em gatos causando conjuntivite e encefalite.
M. feriruminatoris	G5847	GCA_000327395.1	Scaffold	1,01751	24,1	884	797	Próximo ao grupo mycoides e portanto potencial patógeno em ruminantes domesticados.
M. fermentans	M64	GCA_000186005.1	Completo	1,118750	26,9	1015	938	Implicado em doenças respiratórias e artrites em humanos imunossuprimidos.
M. fermentans	JER	GCA_000148625.1	Completo	0,977524	26,9	872	782	Implicado em doenças respiratórias e artrites em humanos.

M. flocculare	Ms42	GCA_000815065.1	Completo	0,778866	29	639	581	Comensal com baixa patogenicidade em suínos.
M. gallinaceum	B2096 8B	GCA_000965765.1	Completo	0,845307	28,4	631	571	Patogênico em galinhas.
M. gallinarum	DSM 19816	GCA_000621085.1	Scaffold	0,834674	26,4	703	652	Comensal em pombos e alguns mamíferos.
M. gallisepticum	R	GCA_000092585.1	Completo	1,012800	31,5	823	733	Causa doença respiratória em aves domésticas.
M. genitalium	G-37	GCA_000027325.1	Completo	0,580076	31,7	566	515	Agente causador de uma ampla gama de infecções urogenitais e do trato respiratório de humanos.
M. glycophilum	ATCC 35277	GCA_000687855.1	Contig	0,89383	28,5	702	626	Presente em aves domésticas, apresenta potencial patogenicidade em análises in vitro.
M. haemocanis	Illinois	GCA_000238995.1	Completo	0,919992	35,3	1182	1130	Causa doença aguda em cães imunossuprimidos ou esplenectomizados.
M. haemofelis	Langford 1	GCA_000200735.1	Completo	1,14726	38,9	1559	1487	Causa anemia hemolítica grave em gatos
M. hominis	ATCC 23114	GCA_000085865.1	Completo	0,665445	27,1	599	549	Oportunistas comensais que residem no trato urogenital inferior de humanos.
M. hyopneumoniae	7448	GCA_000008225.1	Completo	0,920079	28,5	745	683	Causa pneumonia enzoótica em suínos.
M. hyopneumoniae	7422	GCA_000427215.1	Completo	0,898495	28,5	733	653	Causa pneumonia enzoótica em suínos.
M. hyopneumoniae	J	GCA_000008205.1	Completo	0,897405	28,5	728	666	Cepa não patogênica em suínos.
M. hyopneumoniae	232	GCA_000008405.1	Completo	0,892758	28,6	726	662	Causa pneumonia enzoótica em suínos.
M. hyorhinis	HUB-1	GCA_000145705.1	Completo	0,839615	25,9	766	656	Patogênico em suínos.
M. hyosynoviae	NPL4	GCA_000691345.1	Contig	0,892195	27	721	647	Patogênico em suínos.
M. imitans	ATCC 51306	GCA_000518305.1	Scaffold	0,922439	30,6	762	671	Presente em aves de rapina mas sem registro de patogenicidade.
M. iners	ATCC 19705	GCA_000701805.1	Scaffold	0,767866	28,3	645	580	Presente em galináceos com registro de patogenicidade.
M. iowae	695	GCA_000227355.2	Contig	1,19515	24,5	977	906	Patogênica em aves como pombos e perus.
M. leachii	PG50	GCA_000183365.1	Completo	1,008950	23,8	903	820	Patogênico em bovinos.
M. leonicaptivi	ATCC 49890	GCA_000622205.1	Scaffold	0,901123	24,9	806	684	Presente em leões, sem registro na literatura sobre patogenicidade.
M. lipofaciens	ATCC 35015	GCA_000686585.1	Scaffold	0,778716	25,4	692	639	Patogênico em galináceos e aves de rapina causando morte embrionária. A cepa ML64 causa uma gama de sintomas como nanismo, malformações etc.
M. meleagridis	IZSVE/2944/9/2011	GCA_001484825.1	Scaffold	0,647259	25,9	572	510	Patógeno de peru associado com airsacculitis e desordens reprodutivas.
M. mobile	163K	GCA_000008365.1	Completo	0,777079	25	689	652	Patogênico em peixes como *Tinca tinca.*

M. molare	ATCC 27746	GCA_000622165.1	Scaffold	0,842941	24,9	743	697	Presente em cães sem atividade patogênica registrada.
M. mycoides subsp capri	95010	GCA_000253075.1	Completo	1,155840	23,81	1004	903	Agente causador de pleuropneumonia contagiosa em gado.
M. opalescens	ATCC 27921	GCA_000712185.1	Scaffold	0,779012	29,1	676	616	Presente no sistema respiratório de cães sem registro de patogenicidade.
M. orale	ATCC 23714	GCA_000420105.1	Contig	0,710549	25,4	639	559	Bactéria oportunista do trato respiratório humano.
M. ovipneumoniae	NM2010	GCA_000753815.1	Scaffold	1,084160	29,3	781	693	Infecta ovinos e caprinos, causando pleuropneumonia.
M. ovis	Michigan	GCA_000508245.1	Completo	0,702511	31,7	810	766	Agente etiológico de infecções em ovelhas.
M. parvum	Indiana	GCA_000477415.1	Completo	0,564395	27	572	530	Presente em suínos sem registro de quadros de infecção.
M. penetrans	HF-2	GCA_000011225.1	Completo	1,35863	25,7	1068	1016	Causa doença urogenital e respiratória em humanos.
M. pirum	ATCC 25960	GCA_000685905.1	Contig	0,839829	24,2	718	668	Há registros de causar problemas respiratórios em humanos diagnosticados com HIV.
M. pneumoniae	M129; ATCC 29342	GCA_000027345.1	Completo	0,816394	40	1061	691	Patogênico em humanos.
M. primatum	ATCC 25948	GCA_000702785.1	Contig	0,902769	27,6	811	718	Estão presentes no trato urogenital de humanos e possuem potencial patogênico.
M. pullorum	B359_6	GCA_001900245.1	Completo	1,00717	29,1	825	758	Infecção no sistema reprodutor de galinhas, mas também foi encontrado em perus, faisões e perdizes.
M. pulmonis	UAB CTIP	GCA_000195875.1	Completo	0,963879	26,6	790	745	Causa doença semelhante à pneumonia em roedores.
M. putrefaciens	KS1	GCA_000224105.1	Completo	0,832603	26,9	722	654	Agente etiológico de agalaxia contagiosa em caprinos.
M. salivarium	ATCC 23064	GCA_000485555.1	Contig	0,713526	26,4	642	587	Patógeno oportunista em humanos, normal da microbiota oral.
M. simbae	ATCC 49888	GCA_000702705.1	Contig	0,852572	31,8	713	661	Presente em leões, sem registro na literatura sobre patogenicidade.
M. spumans	ATCC 19526	GCA_000620005.1	Scaffold	0,835846	27,3	698	632	Associado a problemas respiratórios em cães mas sem registros como agente etiológico principal.
M. sturni	DSM 22021	GCA_000701485.1	Scaffold	0,813344	29	654	607	Causa conjuntivite em *Sturnus vulgaris* selvagens.
M. suis	KI3806	GCA_000203215.1	Completo	0,70927	31,1	817	771	Patogênico em suínos.
M. suis	Illinois	GCA_000179035.2	Completo	0,742431	31,1	874	817	Patogênico em suínos.

M. synoviae	53	GCA_000008245.1	Completo	0,799476	28,5	725	651	Responsável pela doença do trato respiratório e sinovite em frangos e perus.
M. testudineum	BH29	GCA_002245785.1	Scaffold	0,960895	27,6	825	782	Causa rinite crônica e conjuntivite em tartarugas (*Gopherus agassizii*).
M. testudinis	ATCC 43263	GCA_000687795.1	Contig	1,32591	31,5	1126	1040	Presente em *Testudo graeca* (tartarugas) sem registro de patogenicidade.
M. verecundum	ATCC 27862	GCA_900167035.1	Contig	0,872997	26,9	714	625	Causa conjuntivite em gados e bezerros
M. wenyonii	Massachusetts	GCA_000277795.1	Completo	0,650228	33,9	720	675	Causa infecções agudas e crônicas em bovinos.
M. yeatsii	GM274B	GCA_000875755.1	Completo	0,895051	25,7	788	735	Comensal em caprinos.

Fonte: https://www.ncbi.nlm.nih.gov/genome/browse/#!/overview/mycoplasma. Acesso em: Junho de 2018.