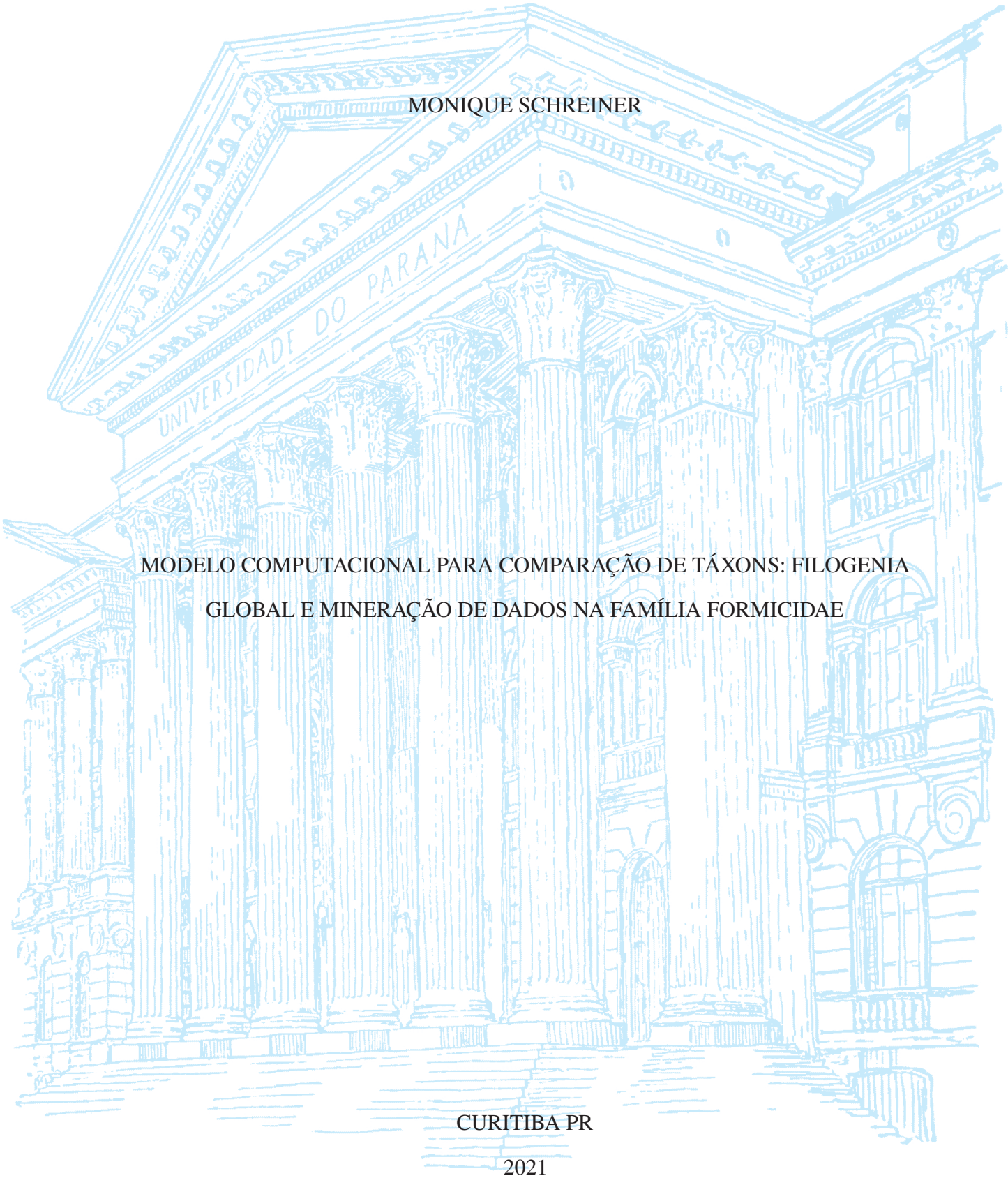
UNIVERSIDADE FEDERAL DO PARANÁ

MONIQUE SCHREINER



MODELO COMPUTACIONAL PARA COMPARAÇÃO DE TÁXONS: FILOGENIA GLOBAL E MINERAÇÃO DE DADOS NA FAMÍLIA FORMICIDAE

CURITIBA PR

2021

MONIQUE SCHREINER

MODELO COMPUTACIONAL PARA COMPARAÇÃO DE TÁXONS: FILOGENIA GLOBAL E MINERAÇÃO DE DADOS NA FAMÍLIA FORMICIDAE

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre em Bioinformática no Programa de Pós-Graduação em Bioinformática, Setor de Educação Profissional e Tecnológica, da Universidade Federal do Paraná.

Área de concentração: *Inteligência Artificial*.

Orientador: Roberto Tadeu Raittz.

CURITIBA PR

2021

Catalogação na publicação
Sistema de Bibliotecas UFPR
Biblioteca de Educação Profissional e Tecnológica

Schreiner, Monique
Modelo computacional para comparação de táxons: filogenia global e mineração de dados na família *Formicidae* [recurso eletrônico] / Monique Schreiner. – Curitiba, 2020.

Dissertação (Mestrado) – Universidade Federal do Paraná, Setor de Educação Profissional e Tecnológica, Programa de Pós-Graduação em Bioinformática, 2021.
Orientador: Roberto Tadeu Raittz.

1. Método SWeeP. 2. Formigas. 3. Inteligência artificial. 4. Bioinformática. I. Raittz, Tadeu. II. Título. III. Universidade Federal do Paraná.

Elaboração: Ângela Pereira de Farias Mengatto - CRB 9/1002

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em BIOINFORMÁTICA da Universidade Federal do Paraná foram convocados para realizar a arguição da dissertação de Mestrado de **MONIQUE SCHREINER** intitulada: **Modelo computacional para comparação de táxons: Filogenia global e mineração de dados na família *formicidae***, sob orientação do Prof. Dr. ROBERTO TADEU RAITTZ, que após terem inquirido a aluna e realizada a avaliação do trabalho, são de parecer pela sua APROVAÇÃO no rito de defesa.

A outorga do título de mestre está sujeita à homologação pelo colegiado, ao atendimento de todas as indicações e correções solicitadas pela banca e ao pleno atendimento das demandas regimentais do Programa de Pós-Graduação.

CURITIBA, 18 de Fevereiro de 2021.

Assinatura Eletrônica
16/03/2021 11:35:00.0
ROBERTO TADEU RAITTZ
Presidente da Banca Examinadora

Assinatura Eletrônica
16/03/2021 13:47:07.0
MARCELO ALBANO MORET SIMÕES GONÇALVES
Avaliador Externo (UNIVERSIDADE DO ESTADO DA BAHIA)

Assinatura Eletrônica
16/03/2021 14:09:22.0
DIEVAL GUIZELINI
Avaliador Interno (UNIVERSIDADE FEDERAL DO PARANÁ)

Assinatura Eletrônica
16/03/2021 11:33:34.0
RODRIGO DOS SANTOS MACHADO FEITOSA
Avaliador Externo (UNIVERSIDADE FEDERAL DO PARANÁ)

À minha Família.

AGRADECIMENTOS

Primeiramente, agradeço aos meus parceiros de projeto. Meu orientador, Roberto, pela paciência e dedicação ao nosso trabalho, sua criatividade e genialidade foram essenciais a esse trabalho. Meu colega, Giuseppe, que trouxe uma energia incrível ao projeto e sempre esteve disposto a me ajudar em tudo o que eu precisei.

Agradeço também aos meus colegas da bioinformática. Mari, por ter me guiado e me ensinado muito sobre informática. Camilinha, por ter sido uma grande parceira que me ajudou muito durante esses dois anos e me ajudou a concretizar muitas ideias. Su, que além de ser uma excelente secretária, é uma excelente amiga. E a todos os colegas que dividiram bons momentos e risadas, seja no café ou nas chamadas de vídeo (Luan, Guilherme, Charlie, Bruno, Camilla, Selma e todos os outros que fizeram parte dessa jornada).

Agradeço aos professores do departamento de Zoologia, Pie e Barbeitos, pelas ideias, aulas, materiais e discussões iniciais.

Agradeço aos meus amigos Lucas, Muci, Celso e Renan, que mesmo nesse período longe continuaram sendo um grande apoio pra mim.

Agradeço aos integrantes da banca, professores Dieval, Feitosa e Marcelo, que se dispuseram a avaliar e contribuir com o trabalho.

Agradeço ao meu melhor amigo e parceiro, Luiz, pela ótima convivência e por todo o carinho. Obrigada por ter cuidado de mim, ter me ajudado tanto a crescer e por trazer tanta alegria para a minha vida.

Por fim, agradeço à minha família, mãe, pai, vó, madrinha e Ricardo. Obrigada por acreditarem e apoiarem todos os meus sonhos e por estarem sempre zelando por mim. Vocês são a base de todas as minhas conquistas.

"We but mirror the world. All the tendencies present in the outer world are to be found in the world of our body. If we could change ourselves, the tendencies in the world would also change. As a man changes his own nature, so does the attitude of the world change towards him. This is the divine mystery supreme. A wonderful thing it is and the source of our happiness. We need not wait to see what others do."
– Mahatma Gandhi

RESUMO

Formigas são consideradas “engenheiras do ecossistema” pois oferecem inúmeros serviços ecológicos e têm impacto na produção de sistemas agrícolas. A interação delas com o ambiente pode afetar desde a composição do solo até o controle de pragas. Estudos filogenéticos acerca do grupo são importantes pois contribuem para a compreensão do funcionamento do ecossistema em que esses animais estão inseridos, além de permitirem a predição de como mudanças nesse funcionamento se comportarão no futuro. Apesar da importância, ainda não há uma filogenia que contemple todas as espécies de formigas. A grande diversidade de espécies (mais de 15 mil espécies e subespécies), técnicas de montagem de árvore computacionalmente custosas, a heterogeneidade na distribuição dos táxons e falta de dados moleculares são fatores que contribuem para a ausência de uma filogenia que contemple todas as espécies. O objetivo desse trabalho é propor uma metodologia para a construção de filogenias de grupos taxonômicos grandes, tendo como resultado final uma filogenia completa de formigas. O método proposto explora o modelo SWeeP e aprendizado de máquina para a vetorização e diminuição da dimensionalidade das sequências, inferência de dados faltantes e integração com informações taxonômicas já existentes. Como resultado, criou-se uma matriz (MAM) que sumarizou a informação molecular disponível. Os testes realizados mostraram que, apesar de dados incompletos, desbalanceados e heterogêneos, a MAM conseguiu representar os padrões taxonômicos e fenotípicos. Em um segundo momento, com a integração da informação taxonômica já existente, foi possível construir uma filogenia com 2.981 espécies congruente com a literatura e, por fim, integrar as espécies sem informação molecular, alcançando a filogenia global com 13.812 espécies de formiga.

Palavras-chave: SWeeP. Dados Faltantes. Formigas. Inteligência Artificial.

ABSTRACT

Ants are considered “ecosystem engineers” as they offer numerous ecological services and have impact on the production of agricultural systems. Their interaction with the environment can affect from soil composition to pest control. Phylogenetic studies about the group are important because they contribute to the understanding of the ecosystem functioning in which these animals are inserted, in addition to allowing the prediction of how changes in this functioning will behave in the future. Despite its importance, there is still no phylogeny that includes all species of ants. The great diversity of species (more than 15 thousand species and subspecies), computationally expensive tree assembly techniques, heterogeneity in the distribution of taxa and lack of molecular data are factors that contribute to the absence of a phylogeny that includes all species. The objective of this work is to propose a methodology for the construction of phylogenies of large taxonomic groups, resulting in a complete phylogeny of ants. The proposed method explores the SWeeP model and machine learning for the vectorization and reduction of the dimensionality of the sequences, inference of missing data and integration with existing taxonomic information. As a result, a matrix (MAM) that summarized the available molecular information was created. The tests performed showed that, despite incomplete, unbalanced and heterogeneous data, MAM was able to represent the taxonomic and phenotypic patterns. In a second step, with the integration of the existing taxonomic information, it was possible to build a phylogeny with 2,981 species congruent with the literature. Finally, species without molecular information were integrated and the global phylogeny with 13,812 ant species was reached.

Keywords: SWeeP. Missing Data. Ants. Artificial Intelligence.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ACRÔNIMOS

CSV	*Comma-separated-values*
EMBL-EBI	*European Molecular Biology Laboratory - European Bioinformatics Institute*
GLAD	*Global Ants Database*
HL	Comprimento da Cabeça (*head length*)
LM	Levenberg-Marquardt
MAF	Matriz Final
MAM	Matriz Alvo Molecular
MSE	Erro Quadrático Médio (*Mean-squared Error*)
NCBI	*National Center for Biotechnology Information*
PCA	Análise de Componentes Principais (*Principal Component Analysis*)
SCG	Gradiente Conjugado Escalonado (*Scaled Conjugate Gradient*)
SWeeP	Spaced Words Projection
UCE	Elementos Ultraconservados (*ultraconserved elements*)
UPGMA	*Un-weighted Pair Group Method with Arithmetic mean*

LISTA DE SÍMBOLOS

®	Marca registrada
ρ	Coeficiente de Correlação de Spearman

SUMÁRIO

1 INTRODUÇÃO

Todo organismo vivo interage com outros organismos e com o ambiente físico em que estão inseridos, afetando os mesmos direta ou indiretamente (ODUM, 2007). Alguns grupos de organismos, como as formigas, possuem maior impacto nos ecossistemas que habitam. Formigas são animais abundantes e podem ser encontradas em praticamente todos os ambientes terrestres, com exceção somente da Groelândia, Antártica e algumas ilhas (FOLGARAIT, 1998). Compreendem cerca de um terço de toda a biomassa de insetos, equivalendo-se à biomassa total de humanos (FOLGARAIT, 1998; OFFENBERG, 2015). Possuem a maior riqueza de espécies de todos os insetos sociais, além de serem o grupo de organismos terrestres mais dominante em climas tropicais, tanto ecologicamente quanto numericamente (MOREAU; BELL, 2013; Ward P., 2007). Pela abundância e diversidade do grupo, interação com o ambiente e impacto nos ecossistemas, as formigas são consideradas "engenheiras do ecossistema", oferecendo inúmeros serviços ecológicos.

Em sistemas agrícolas, formigas impactam a produção de diversas maneiras, podendo reduzir a produção das plantações em até 80% (CHAN; GUÉNARD, 2020). Primeiramente, as formigas afetam diretamente a dispersão de sementes, a decomposição de matérias, o ciclo de nutrientes, as propriedades físicas e químicas dos solos, a biomanta e outros organismos presentes no solo (ANGOTTI *et al.*, 2018; FOLGARAIT, 1998), além de que espécies cortadeiras podem danificar plantações (SILVA; ROSA, 2017). Por serem muito responsivas à atividades humanas, as formigas têm grande valor como indicadores ambientais, sendo organismos que fornecem informações sobre a qualidade do solo, propensão a alagamento, propensão a queimadas e mudanças causadas por espécies invasoras (ANGOTTI *et al.*, 2018; ESTRADA, 2017; FOLGARAIT, 1998). Por fim, por serem animais sociais com polimorfismo intraespecífico, as formigas conseguem atuar em diferentes níveis tróficos e podem ser usadas no controle biológico de pestes agrícolas (ANGOTTI *et al.*, 2018; FOLGARAIT, 1998; OFFENBERG, 2015).

Visto o impacto do grupo, fica clara a importância de estudos evolutivos, como os abordados pela filogenética. A filogenética possibilita não somente o entendimento de como espécies evoluíram para o que são hoje, mas também permite a predição de como as mudanças poderão se comportar no futuro. Entre diversas aplicações, este conhecimento pode ser utilizado para classificação de organismos, identificação da origem de características e estudos de macroevolução, além de poder ser empregado em inúmeras áreas como conservação, biogeografia, bioinformática e computação (EMERY, 2019; HILL; DAVIS, 2014; THOMAS *et al.*, 2013).

Apesar da importância, ainda não há uma filogenia que contemple todas as espécies de formigas e diversos fatores contribuem para essa ausência. Primeiramente, são necessários a coleta, armazenamento e processamento de uma quantidade muito grande de dados, visto que o táxon possui mais de 15 mil espécies e subespécies (HILL; DAVIS, 2014). As técnicas atuais

são custosas computacionalmente e/ou não conseguem sintetizar e relacionar a informação total do táxon (HAESELER, 2012). Por segundo, há uma grande heterogeneidade na distribuição taxonômica do grupo e a maioria dos gêneros e espécies válidas estão contidos em poucas subfamílias (ANTWIKI, 2020). Por fim, há a falta de dados moleculares. A grande maioria das espécies possui pouco ou nenhum dado molecular disponível em bancos de dados públicos (SCHOCH *et al.*, 2020).

Este trabalho tem como proposta desenvolver uma metodologia para a construção de filogenias de grupos taxonômicos grandes, tendo como resultado final uma filogenia completa de formigas. Essa filogenia fornecerá o conhecimento global acerca das relações filogenéticas do grupo, guiando estudos de diversificação, filogenética de comunidades, predição de estrutura e função de proteínas, detecção de seleção positiva, coevolução, isolamento evolutivo, localização de potenciais agentes de controle biológico, melhoramento genético e identificação de espécies de interesse (BURGIO *et al.*, 2019; DONOGHUE, 2009; THOMAS *et al.*, 2013; WARNOW, 2018).

O método terá como base dados moleculares disponíveis em bancos de dados públicos e informação taxonômica já existente. Serão exploradas técnicas vetoriais que serão ampliadas e adaptadas para que seja possível o processamento dos táxons simultaneamente, diferente da metodologia de super-árvore (HAESELER, 2012; WARNOW, 2018). Mais especificamente, o método SWeeP será implementado para a criação dos vetores que representam as sequências biológicas (DE PIERRE *et al.*, 2020). Metodologias de aprendizado de máquina serão utilizadas para a predição de dados faltantes e desenvolvimento de modelos complementares que possam gerar a filogenia mais coerente e contemplar todas as espécies viventes de formigas.

1.1 OBJETIVOS

O objetivo geral deste trabalho é desenvolver uma metodologia para a construção de filogenias de táxons grandes utilizando técnicas de aprendizado de máquina e de vetorização livre de alinhamento. A metodologia permitirá o processamento de uma grande quantidade de dados moleculares e a predição de dados faltantes. O método desenvolvido será aplicado na construção de uma filogenia completa da família Formicidae.

1.1.1 Objetivos Específicos

- Fazer levantamento dos dados moleculares de proteínas da família Formicidae através de pesquisas em bancos de dados públicos disponíveis *online*.
- Analisar os dados coletados, reunindo informações como quais espécies e quais proteínas são contempladas pelos dados moleculares coletados e qual a qualidade desses dados.
- Utilizar métodos livres de alinhamento para a vetorização das sequências, diminuindo a dimensionalidade dos dados, porém preservando a informação necessária para o estudo.

- Criar modelos alvos a partir dos dados moleculares disponíveis, de modo que as análises possam ser feitas apenas com parte das informações moleculares.
- Utilizar aprendizado de máquina para a predição de dados moleculares faltantes, completando lacunas que dificultariam a construção da filogenia.
- Adicionar informação taxonômica já existente para enriquecer a informação do modelo.
- Criar uma filogenia final, possibilitando a integração dos modelos criados a partir de diferentes fontes de informação.
- Analisar a qualidade da informação do modelo e da filogenia gerada.

1.2 JUSTIFICATIVA

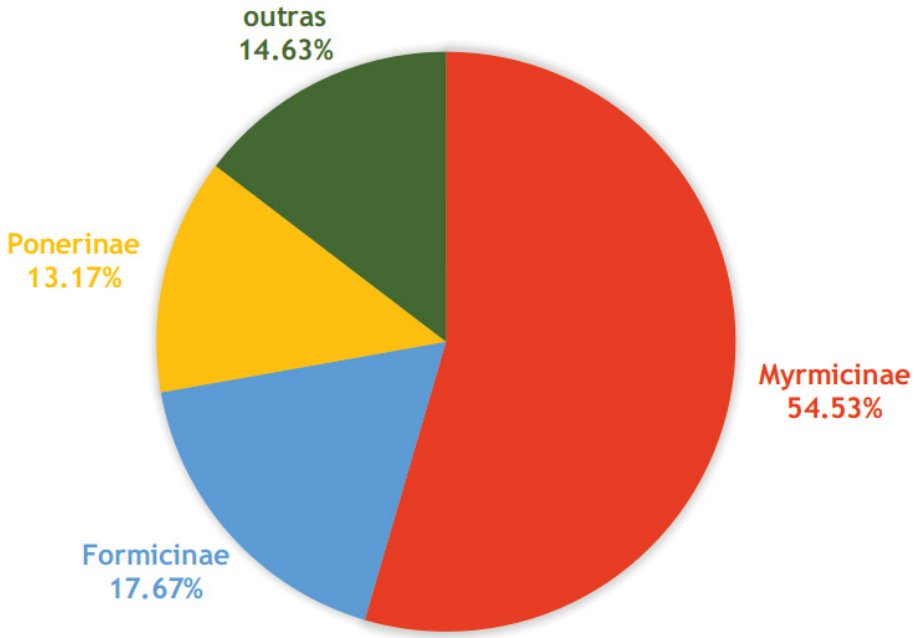
Formigas são insetos pertencentes à família Formicidae (ordem Hymenoptera) e atualmente são classificadas em 17 subfamílias, 338 gêneros e 15.695 espécies e subespécies (ANTWIKI, 2020). A distribuição dos táxons é bastante heterogênea e apenas uma subfamília, a subfamília Myrmicinae, contém 43,19% de todos os gêneros válidos e 48,19% das espécies e subespécies válidas. Esta heterogeneidade também se reflete na distribuição de dados moleculares disponíveis nos bancos de dados públicos (FIGURA 1.1), sendo que apenas três subfamílias (Myrmicinae, Formicinae e Ponerinae) detêm 85,37% dos dados moleculares de proteínas disponíveis (SCHOCH *et al.*, 2020). Ou seja, poucas espécies aparecem com milhares de sequências cadastradas nos bancos de dados, enquanto a maioria das espécies apresenta-se com poucas ou nenhuma sequência cadastrada, dificultando comparações filogenéticas moleculares.

Atualmente, limitações computacionais das técnicas utilizadas também dificultam o processamento de uma grande quantidade de dados simultaneamente (WARNOW, 2018). As filogenias mais recentes da ordem Hymenoptera, incluindo formigas, vespas e abelhas, utilizam de técnicas de alinhamento de transcriptomas (PETERS *et al.*, 2017), mitogenomas (TANG *et al.*, 2019) e elementos ultraconservados (UCE, do inglês *ultraconserved elements*) (FAIRCLOTH *et al.*, 2015; BRANSTETTER *et al.*, 2017). Essas filogenias exploram as relações evolutivas entre os grupos de Hymenoptera porém as análises são feitas com apenas alguns representantes de cada grupo.

Quando o objetivo é construir uma filogenia de táxons extensos, com todas as espécies representadas, é comum que se use a metodologia de super-árvore que consiste basicamente na construção de árvores filogenéticas pequenas que são combinadas em uma única filogenia (vide seção 2) (WARNOW, 2018). Essas árvores menores geralmente são obtidas através do alinhamento múltiplo de sequências que é computacionalmente bastante custoso. A combinação das árvores menores também requer grande capacidade computacional.

Métodos livres de alinhamento (*Alignment-free*) são uma alternativa para diminuir o custo computacional na comparação de sequências, pois conseguem quantificar a similaridade

FIGURA 1.1: DISTRIBUIÇÃO DOS DADOS DE SEQUÊNCIAS DE PROTEÍNAS CADASTRADOS NO NCBI PARA AS SUBFAMÍLIAS DE FORMICIDAE



FONTE: (SCHOCH *et al.*, 2020)
LEGENDA: Três subfamílias (Myrmicinae, Ponerinae e Formicinae) se destacam pela quantidade de sequências depositadas no NCBI. Em verde, concentra-se a soma de sequências depositadas para 14 subfamílias.

ou dissimilaridade de sequências sem usar ou produzir alinhamento no algoritmo da aplicação (ZIELEZINSKI *et al.*, 2017). Estes métodos geralmente representam sequências de nucleotídeos ou aminoácidos como vetores numéricos, deste modo as análises podem ser realizadas de forma mais eficiente (VINGA, 2014). São menos custosos computacionalmente pois seus algoritmos possuem complexidade linear, dependendo somente do tamanho da sequência de entrada. Além disso, a vetorização das sequências permite que as mesmas sejam usadas como dados de entrada para técnicas de aprendizado de máquina, como no algoritmo de vetorização SWeeP (DE PIERRE *et al.*, 2020).

Aprendizado de Máquina (*Machine Learning*) é uma forma de inteligência artificial que possibilita que máquinas aprendam a partir de um conjunto de dados sem serem especificamente programadas para cada situação possível (CHAKRABORTY; CHOUDHURY, 2017). Algoritmos de aprendizado de máquina implementam conjuntos de regras baseados em métodos estatísticos que extraem padrões de grandes quantidades de dados, criando modelos de forma automática, os quais podem ser usados para o entendimento dos dados e para a inferência de dados latentes ou faltantes (GHAHRAMANI, 2015).

O método proposto nesse trabalho permitirá a comparação entre sequências moleculares com pouco custo computacional, a predição de informações faltantes a partir da criação de um

modelo base, a integração de dados moleculares com dados taxonômicos já existentes e, por fim, a construção de uma filogenia completa da família Formicidae.

1.3 ESTRUTURA DA DISSERTAÇÃO

Essa dissertação encontra-se dividida em nove capítulos, sendo o presente capítulo o primeiro, a introdução.

A revisão teórica se encontra do capítulo 2 ao capítulo 4, sendo que os conteúdos revisados são, respectivamente, sistemática filogenética, métodos livres de alinhamento e análise de *big data*.

O desenvolvimento do projeto encontra-se dividido nos capítulos 5, 6, 7 e 8. O capítulo 5 visa relatar os passos para a obtenção do modelo e filogenia molecular. O capítulo 6 tem como objetivo realizar comparações filogenéticas entre a filogenia gerada e filogenias feitas em estudos anteriores. O capítulo 7 visa testar a qualidade da informação do modelo através de técnicas de aprendizado de máquina. E, por fim, o capítulo 8 mostra a integração de dados novos e sem informação molecular ao modelo e a geração da filogenia global. Cada capítulo mostra o material e métodos utilizados, os resultados alcançados e a discussão acerca do capítulo.

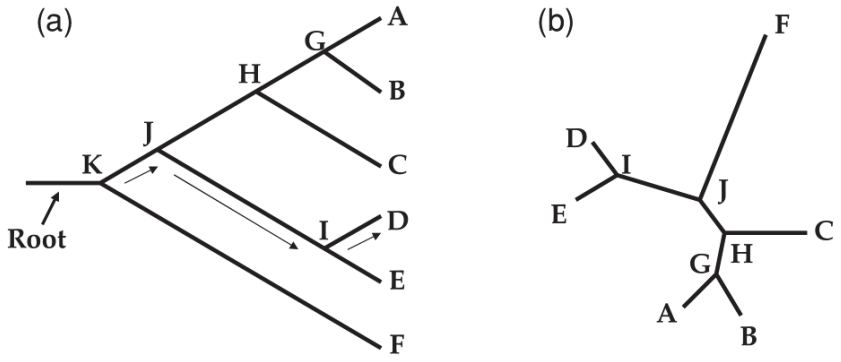
Por último, as conclusões alcançadas durante todo o processo e as perspectivas futuras encontram-se no capítulo 9.

2 SISTEMÁTICA FILOGENÉTICA

A sistemática é o campo de pesquisa da biologia dedicado a descrever e organizar a biodiversidade entre os seres vivos, encontrar que tipo de ordem existe na diversidade (se existir), compreender os processos que são responsáveis pela geração de diversidade e apresentar um sistema geral de referência sobre a diversidade biológica (AMORIM, 2002). A sistemática filogenética, ou somente filogenética, é um ramo da sistemática que baseia sua metodologia de classificação na história evolutiva dos organismos, ou seja, no grau de parentesco filogenético entre eles (OLIVEIRA, 2010). A filogenética possibilita não somente o entendimento de como espécies evoluíram para o que são hoje, mas também permite a predição de como as mudanças poderão se comportar no futuro. Entre diversas aplicações, este conhecimento pode ser utilizado para classificação de organismos, identificação da origem de patógenos e ser empregado nas áreas de ciências forenses, conservação, bioinformática e computação (EMERY, 2019).

As relações evolutivas são representadas através de árvores filogenéticas (FIGURA 2.1). As unidades taxonômicas dos organismos viventes são representadas nos nós terminais. Cada nó interno é uma unidade taxonômica hipotética que representa um antepassado comum ao ramos que descendem deste nó (LEMEY; SALEMI; VANDAMME, 2009). A raiz da árvore representa o ancestral de todas as unidades taxonômicas representadas nela. Quando a direção do processo evolutivo e o ancestral comum a todas as unidades taxonômicas não são indicados, a árvore é denominada não-enraizada. Em algumas árvores, o tamanho de ramo pode representar a quantidade de mudanças ou o tempo ocorrido entre os nós.

FIGURA 2.1: EXEMPLOS DE ÁRVORES FILOGENÉTICAS



FONTE: Lemey, Salemi e Vandamme (2009)

LEGENDA: em ambas as árvores A, B, C, D, E e F são as unidades taxonômicas viventes. G, H, I, J K são as unidades taxonômicas hipotéticas. (a) Árvore filogenética enraizada. *Root* é a raiz, ou seja, o ancestral comum a todos. As setas indicam o sentido do processo evolutivo. (b) Árvore não enraizada. O ancestral comum a todos não é indicado.

A construção de árvores filogenéticas moleculares tende a ser composta de duas etapas: alinhamento de sequências (vide seção 2.1) e processamento da árvore a partir do alinhamento (WARNOW, 2018). Os principais métodos usados para o processamento de árvores filogenéticas são (HAUBOLD, 2014; LEMEY; SALEMI; VANDAMME, 2009):

- **Métodos de parcimônia:** encontram a topologia que requer a menor quantidade de mudanças de estado para produzir as características das unidades taxonômicas observadas (SOBER, 2004). Exemplo: máxima parcimônia (*Maximum Parsimony*).
- **Métodos de probabilidade:** através de modelos evolutivos, encontram a topologia com a maior probabilidade de gerar as características das unidades taxonômicas observadas (SOBER, 2004). Exemplo: máxima verossimilhança (*Maximum Likelihood*) e inferência Bayesiana.
- **Métodos de distância:** ajustam a topologia a uma matriz de distância estimada através das verdadeiras distâncias genéticas (NOVELLO; VILLAR; URIOSTE, 2009). Exemplos: UPGMA (*Un-weighted Pair Group Method with Arithmetic mean*) e agrupamento de vizinhos (*Neighbor-joining*).

Para a construção de filogenias muito extensas, é comum que seja utilizado o método de super-árvore em que árvores filogenéticas de grupos menores (sub-táxons) são combinadas para evitar o processamento simultâneo de uma quantidade grande de dados. Isso se deve ao fato de que o alinhamento de sequências é um processo computacionalmente custoso (vide seção 2.1) e os métodos de construção de árvores de maior acurácia também requerem grande capacidade computacional (WARNOW, 2018). O método de super-árvore consiste basicamente de três etapas (HAESELER, 2012; WARNOW, 2018): alinhamento múltiplo de sequências entre as espécies compreendidas nos sub-táxons, construção da topologia das árvores dos sub-táxons e, por fim, construção da filogenia do táxon completo através da combinação das árvores geradas para os sub-táxons. A filogenia completa obtida pelo método de super-árvore acaba refletindo somente a relação entre as espécies dos táxons menores e não do táxon como um todo (HAESELER, 2012).

Alguns trabalhos recentes que construíram grandes filogenias utilizando o método de super-árvore contemplam a diversidade global de aves viventes com 9.993 espécies, a diversidade de camarões Caridae com 765 táxons e a diversidade de aves Psittaciformes viventes com 413 espécies (JETZ *et al.*, 2012; DAVIS *et al.*, 2018; BURGIO *et al.*, 2019).

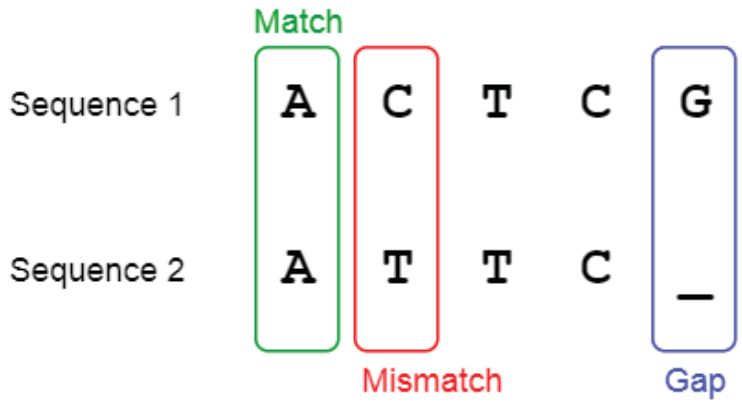
2.1 EVOLUÇÃO MOLECULAR E FILOGENÉTICA

Inicialmente, técnicas de reconstrução da história evolutiva de organismos eram baseadas em poucos critérios, principalmente em características morfológicas e limitadas a pequenos grupos taxonômicos (HILLIS; MORITZ; MABLE, 1996). A partir dos anos 60, com o avanço de técnicas para o estudo de estruturas moleculares de proteínas e ácidos nucleicos, sistematas

passaram a utilizar dados moleculares em análises filogenéticas (HILLIS; MORITZ; MABLE, 1996).

Segundo Hillis, Moritz e Mable (1996), alinhamento é a etapa mais difícil e menos compreendida nas análises genéticas. É comum que análises filogenéticas com dados moleculares utilizem de técnicas de alinhamento de sequências para inferir a história evolutiva de organismos. De uma perspectiva biológica, um alinhamento de sequências é uma hipótese de homologia, ou seja, assume-se que as sequências derivaram de uma sequência ancestral em comum. Neste tipo de técnica, algoritmos buscam definir a similaridade entre diferentes sequências através da correspondência individual de cada base ou aminoácido (ou grupo de bases ou aminoácidos) (LEMEY; SALEMI; VANDAMME, 2009; ZIELEZINSKI *et al.*, 2017). Cada posição no alinhamento será categorizada em dois estados: similar/conservada (match), quando as sequências apresentam a mesma base ou aminoácido na posição; ou não-conservada (mismatch), quando as sequências apresentam bases ou aminoácidos diferentes na mesma posição (ZIELEZINSKI *et al.*, 2017) (FIGURA 2.2). A maioria dos algoritmos de alinhamento ainda modelam um terceiro estado, inserção/deleção (*gap*) que representa aminoácidos que possam ter sido inseridos ou deletados da sequência durante o processo evolutivo.

FIGURA 2.2: EXEMPLO DE ALINHAMENTO ENTRE DUAS SEQUÊNCIAS



FONTE: Towards Data Science (2019)

LEGENDA: alinhamento entre duas sequências (*Sequence 1* e *Sequence 2*). Na primeira posição a base aparece conservada (*match*), pois é a mesma nas duas sequências. Na segunda posição a base aparece não-conservada (*mismatch*), pois difere nas sequências. Na quinta posição, há uma inserção/deleção (*gap*).

Técnicas de alinhamentos tendem a perder acurácia em algumas situações como, por exemplo, quando os genomas homólogos de organismos possuem um número e ordem de elementos gênicos extremamente variável devido à altas taxas de mutação, recombinação e duplicação gênica, perda e ganho de genes e transferências horizontais, como é o caso dos vírus (ZIELEZINSKI *et al.*, 2017). Outro ponto a ser discutido, é que técnicas de alinhamento dependem de suposições acerca da história evolutiva das sequências que estão sendo comparadas. Estas suposições se dão através de parâmetros como matrizes de substituição, penalidade para

gaps, e estabelecimento de valores limites para parâmetros estatísticos. Estes parâmetros possuem valores arbitrários, não são consenso entre diferentes algoritmos e a escolha de parâmetros errados podem afetar o alinhamento. Por fim, estas técnicas também são complexas e custosas computacionalmente tanto no quesito memória quanto no quesito tempo, o que pode acabar limitando a quantidade de dados que pode ser analisada (VINGA, 2014; ZIELEZINSKI *et al.*, 2017).

3 MÉTODOS LIVRES DE ALINHAMENTO

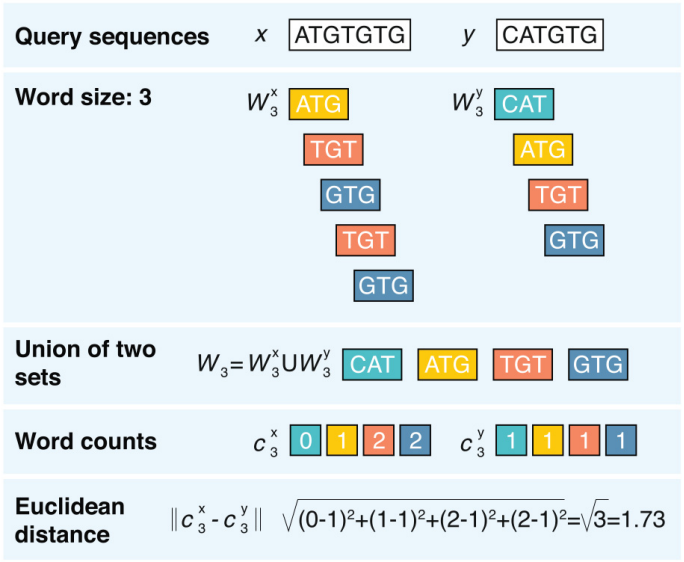
Métodos livres de alinhamento (*Alignment-free*) são técnicas que quantificam a similaridade ou dissimilaridade de sequências sem usar ou produzir alinhamento no algoritmo da aplicação (ZIELEZINSKI *et al.*, 2017). Esses métodos geralmente representam sequências de nucleotídeos ou aminoácidos como vetores numéricos, deste modo as análises podem ser realizadas de forma mais eficiente (VINGA, 2014). Esse tipo de método é menos custoso computacionalmente pois seus algoritmos possuem complexidade linear, dependendo somente do tamanho da sequência de entrada. Consequentemente, são adequados para quantidade de dados grandes, como análises de genomas completos. Métodos livres de alinhamento são adequados também para sequências pouco conservadas, pois são resistentes a eventos de recombinação e embaralhamento genético. Por fim, diferente dos métodos de alinhamento, estes métodos não dependem de suposições prévias acerca da história evolutiva dos organismos (ZIELEZINSKI *et al.*, 2017).

De modo geral, métodos livres de alinhamento podem ser classificados em dois grupos. O primeiro grupo é composto dos métodos baseados nas frequências de subsequências de tamanho definido, também chamados de métodos *word-based* (FIGURA 3.1). Este tipo de método parte do princípio que sequências similares possuem subsequências de tamanho k (também chamadas de *k-mers*) similares. De forma resumida, o método é composto de três passos: quebra da sequência em subsequências de tamanho definido (*k-mers*), geralmente utilizando janelas móveis, transformação das subsequências em vetores numéricos e quantificação da dissimilaridade através de uma função para cálculo de distância (distância Euclidiana, por exemplo).

As subsequências são definidas através de máscaras que podem ser exatas ou inexatas. As máscaras são compostas de 1's que implicam em correspondência exata e 0's que significam que a correspondência não precisa ser necessariamente exata (*do not care*) (HAUBOLD, 2014). Por exemplo, para uma máscara 101 (exata, *do not care*, exata), a subsequência ATA seria correspondente com a subsequência ATA, mas também com a subsequência AAA, visto que a posição central não precisa ser exata.

O segundo grupo contempla os métodos que avaliam o conteúdo de sequências completas, denominados métodos *information-theory-based* (FIGURA 3.2). Esses métodos computam a quantidade de informação que é compartilhada entre sequências biológicas. Normalmente a quantidade de informação é medida através da complexidade por meio de compressão (o tamanho das sequências comprimidas tende a aumentar com o aumento da complexidade das sequências originais) ou através da entropia (quantificação de subsequências mais significativas).

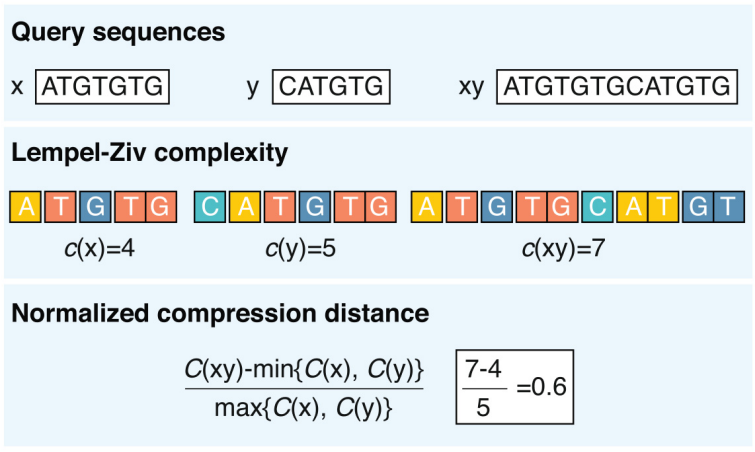
FIGURA 3.1: EXEMPLO DE MÉTODO *WORD-BASED*



FONTE: Zielezinski *et al.* (2017)

LEGENDA: duas sequências (x e y) são quebradas utilizando subsequências de tamanho 3 (3-mers). As subsequências são transformadas em vetores numéricos a partir da ocorrência das subsequências em cada sequência. Por fim, a dissimilaridade é calculada através da Distância Euclidiana.

FIGURA 3.2: EXEMPLO DE MÉTODO *INFORMATION-THEORY-BASED*



FONTE: Zielezinski *et al.* (2017)

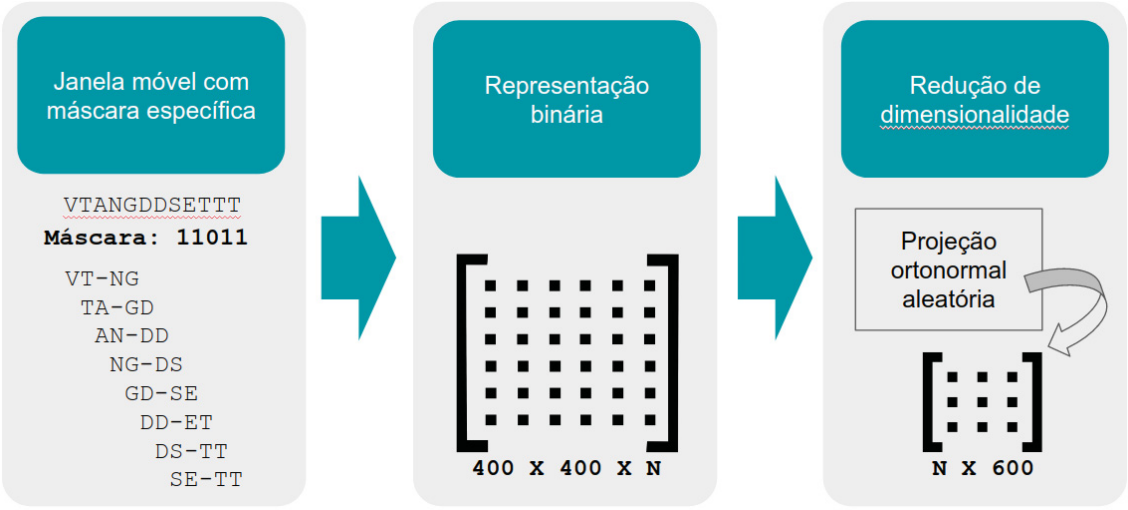
LEGENDA: Uso do algoritmo Lempel-Ziv. Duas sequências (x e y) são concatenadas em uma terceira sequência (xy). As três sequências são varridas da esquerda para direita e o número de subsequências diferentes é contabilizado para cada uma delas. Após isso a distância entre as sequências X e Y é calculada utilizando uma função de Distância de Compressão Normalizada.

3.1 MÉTODO *SWEEP*

O método SWeeP (*Spaced Words Projection*)[1] é uma técnica livre de alinhamento que representa sequências de proteínas na forma de vetores numéricos compactos (DE PIERRE *et al.*, 2020). O método é baseado na projeção de *k-mers* em uma base ortonormal orientada aleatoriamente que possui um número de coordenadas suficientes para manter a comparabilidade das sequências. O método SWeeP utiliza o conceito *word-based* com máscaras inexatas para varrer as sequências e gerar vetores binários que representam a informação contida na sequência. Esse vetor é então projetado em um vetor de menor dimensão que mantém a maioria das informações de comparação (FIGURA 3.3). A dimensionalidade do vetor de menor dimensão pode ser alterada de acordo com a necessidade do projeto.

SWeeP é um método novo que consegue lidar eficientemente com volumes grandes de dados e pode ser usado tanto para o propósito de mineração de dados, quanto para comparação de sequências. Os vetores gerados pelo SWeeP podem ser usados para o treinamento de modelos de aprendizado de máquina. A partir dos vetores, também é possível a construção de árvores filogenéticas consistentes como já mostrado em alguns estudos como o da filogenia de todos os proteomas mitocondriais disponíveis no RefSeQ (8.426 proteomas), filogenia de todos os genomas bacterianos completos disponíveis no NCBI (10.324 micro-organismos), filogenia das linhagens *Azoarcus- Aromatoleum* (30 genomas) e filogenia de todos os proteomas virais disponíveis no NCBI (31,386 proteomas) (DE PIERRE *et al.*, 2020; RAITTZ *et al.*, 2021; FERNANDES *et al.*, 2020).

FIGURA 3.3: FLUXOGRAMA DO MÉTODO SWEEP



FONTE: Kulik (2020)

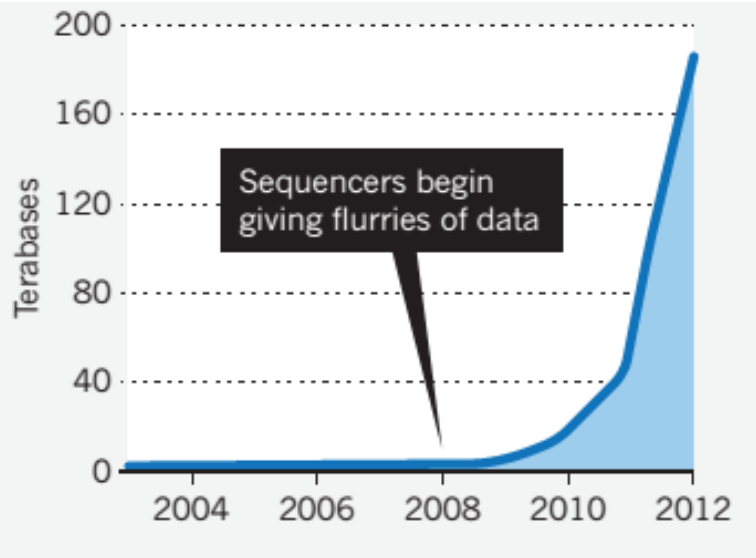
LEGENDA: a sequência é varrida por uma janela móvel (máscara 11011), gerando uma representação binária de alta dimensão da sequência que é então projetada num vetor de menor dimensão. No caso do exemplo, ao fim do processo, cada sequência é representada por um vetor numérico de 600 coordenadas.

[1]Disponível no endereço <https://sourceforge.net/projects/spacedwordsprojection/>.

4 ANÁLISE DE *BIG DATA*

Com o avanço tecnológico, principalmente as novas tecnologias de sequenciamento que possibilitaram que até pequenos laboratórios sequenciassem em grande escala, a geração de dados biológicos sofreu um grande aumento e discussões acerca de volumes muito grande de dados, conhecidos como *big data*, começaram a surgir (MARX, 2013; CHICCO, 2017). Calcula-se que no período de 2004 a 2012, a quantidade de dados biológicos cresceu exponencialmente (FIGURA 4.1), sendo que dados de sequências genômicas dobraram de quantidade num período menor que um ano (SHAKIL; ALAM, 2018). Em 2013, somente o repositório do EMBL-EBI (*European Molecular Biology Laboratory - European Bioinformatics Institute*) já armazenava 20 petabytes de dados biológicos (MARX, 2013). Em razão disso, alguns dos desafios que vêm surgindo aos cientistas são maneiras de analisar, processar, comparar, compartilhar e armazenar essas grandes quantidades de dados disponíveis nos repositórios, visto que estas tarefas não são possíveis de serem feitas manualmente devido ao grande volume, além de exigirem muito poder computacional (SHAKIL; ALAM, 2018; MARX, 2013).

FIGURA 4.1: EXPLOSÃO DE DADOS



FONTE: Marx (2013)

LEGENDA: crescimento em terabases da quantidade de dados provenientes de sequenciamento armazenados no EMBL. 1 terabase = 10^{12} pares de bases

O processo de coletar, selecionar, processar, analisar e extrair informações de dados é denominado mineração de dados (*data mining*) (AGGARWAL, 2015). No contexto do *big data*, a mineração de dados tem por objetivo automatizar a extração de informações concisas de dados desestruturados, heterogêneos e de fontes diversas (AGGARWAL, 2015). Algumas das técnicas

comumente usadas na mineração de grandes volumes de dados são o aprendizado de máquina e a redução de dimensionalidade (KHAN *et al.*, 2010).

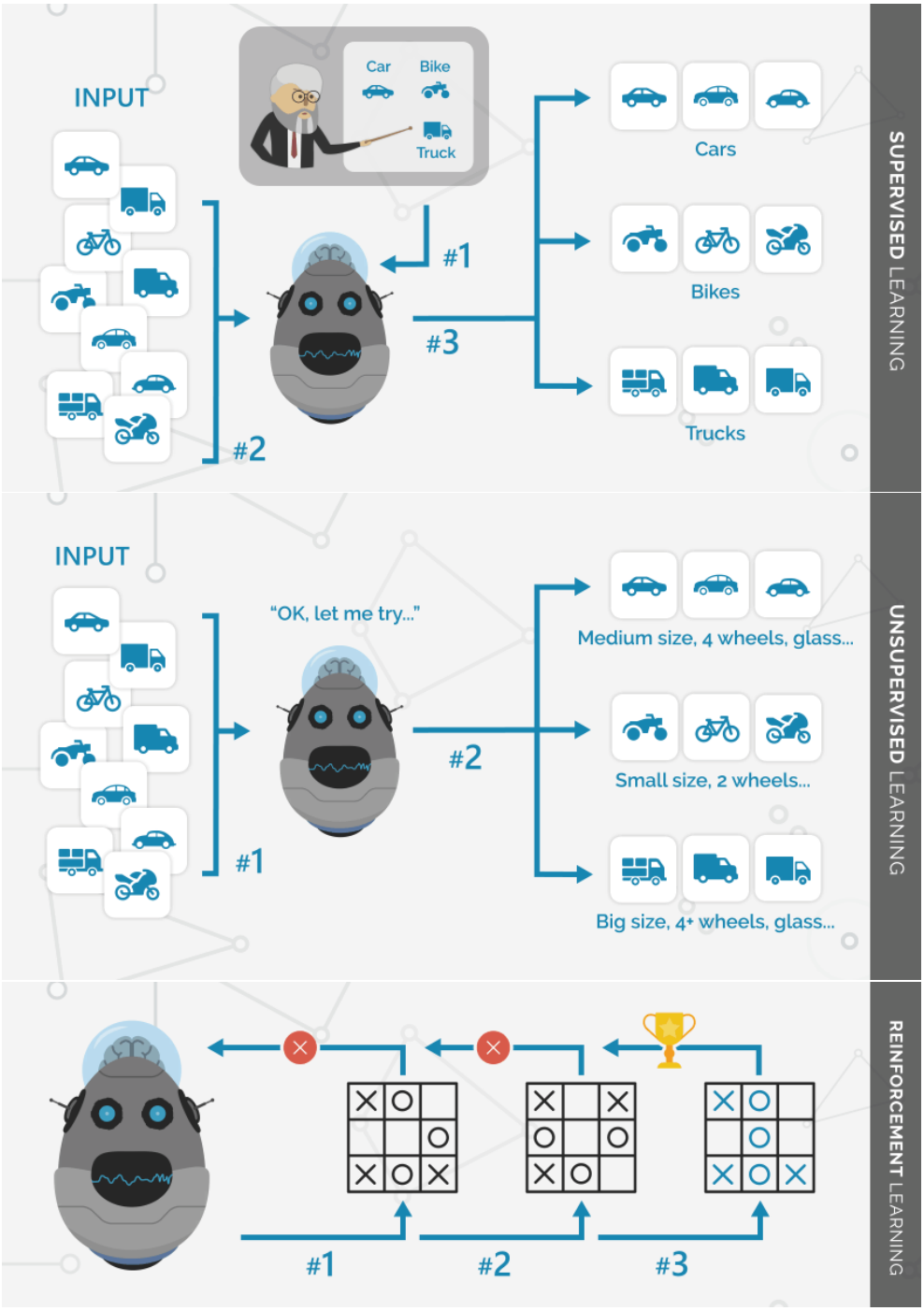
4.1 APRENDIZADO DE MÁQUINA

Aprendizado de Máquina (*Machine Learning*) é uma forma de inteligência artificial que possibilita que máquinas aprendam a partir de um conjunto de dados sem serem especificamente programadas para cada situação possível (CHAKRABORTY; CHOUDHURY, 2017). Algoritmos de aprendizado de máquina implementam conjuntos de regras baseados em métodos estatísticos que extraem padrões de grandes quantidades de dados, criando modelos de forma automática que poderão ser usados para o entendimento dos dados e para a inferência de dados latentes ou faltantes (GHAHRAMANI, 2015).

Os métodos de aprendizado podem ser classificados em três categorias principais (FIGURA 4.2): aprendizado supervisionado, aprendizado não-supervisionado e aprendizado por reforço (RUSSELL; NORVIG, 2013). Modelos de aprendizado supervisionados fazem suas predições mapeando dados já classificados (JORDAN; MITCHELL, 2015). De forma geral, o algoritmo realiza ciclos em que infere a classificação dos dados, comparando com a classificação real (CHAKRABORTY; CHOUDHURY, 2017). Os erros são usados para controlar o processo de aprendizagem, repetindo os ciclos até que o erro seja minimizado. Modelos de aprendizado não-supervisionados envolvem a análise de dados não classificados e o padrão para a classificação é aprendido pelo algoritmo durante o processo de aprendizado (CHAKRABORTY; CHOUDHURY, 2017; JORDAN; MITCHELL, 2015). Várias hipóteses de classificação são testadas pelo algoritmo e a mais otimizada é usada. Por fim, no aprendizado de reforço, o agente interage com o ambiente através de observações, ações e recompensas (MNIH *et al.*, 2015). Ou seja, no lugar de usar dados para o aprendizado que mostram a classificação real correta, os dados são usados apenas como uma indicação sobre se uma ação é correta ou não (JORDAN; MITCHELL, 2015). Neste tipo de aprendizagem, a recompensa é relacionada ao conjunto das ações tomadas e não a ações individualizadas.

Os dados utilizados no processo de aprendizado podem ser caracterizados em quantitativos ou qualitativos (também denominados categóricos) (JAMES *et al.*, 2013). Dados quantitativos assumem valores dentro de um conjunto numérico e podem ser representados por números inteiros (variáveis discretas) como idade e número de filhos, ou números decimais (variáveis contínuas) como altura, peso e preço. Dados qualitativos representam categorias e assumem o valor de n diferentes classes. As classes podem ser representadas por variáveis nominais em que não existe ordenação entre as categorias (gênero e marca de um produto, por exemplo) ou variáveis ordinais em que existe ordenação (escolaridade e estágio da doença, por exemplo). De forma geral, problemas que exigem uma resposta quantitativa são chamados de problemas de regressão, enquanto problemas que exigem resposta qualitativa são chamados de problemas de classificação (FIGURA 4.3) (JAMES *et al.*, 2013).

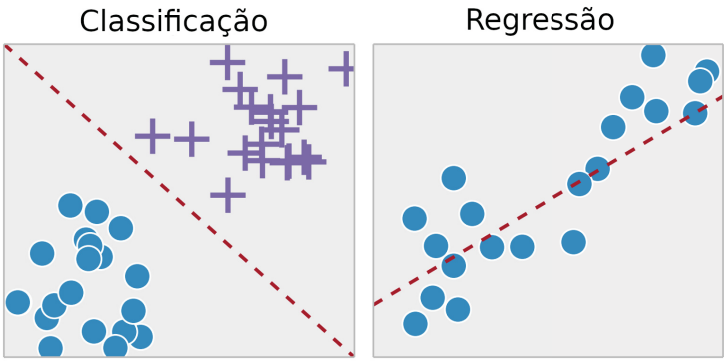
FIGURA 4.2: MÉTODOS DE APRENDIZADO DE MÁQUINA



FONTE: Business Process Incubator (2019)

LEGENDA: *Supervised learning:* aprendizado supervisionado em que o computador aprende a partir de dados previamente rotulados. *Unsupervised learning:* aprendizado não-supervisionado em que o computador agrupa os dados de acordo com o padrão encontrado durante o treinamento. *Reinforcement learning*: aprendizado por reforço em que o computador aprende através de recompensas e punições. Os *INPUT* representam as variáveis preditoras das variáveis respostas.

FIGURA 4.3: EXEMPLO DE PROBLEMAS DE CLASSIFICAÇÃO E REGRESSÃO

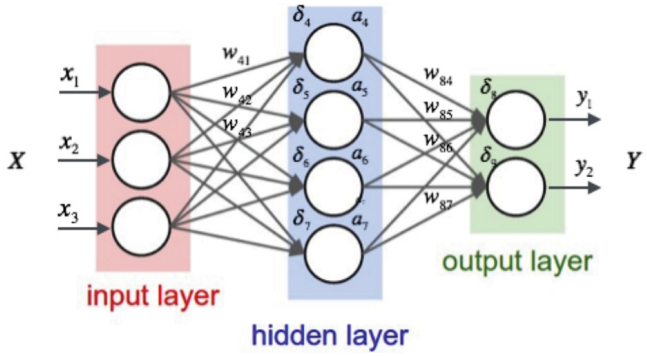


FONTE: tradução livre de Towards Data Science (2018)
LEGENDA: na classificação os dados são separados em categorias bem definidas. Na regressão, é encontrada a função que melhor se adapta aos dados.

4.1.1 Redes Neurais Artificiais

Redes neurais artificiais, também chamadas de redes neuronais, são modelos de aprendizado de máquina inspirados em redes neurais biológicas, o sistema nervoso central de animais. A redes neurais artificiais típicas são organizadas na forma de três camadas (entrada, intermediária ou oculta e saída) e compostas de neurônios (unidades de processamento) e sinapses (conexões entre os neurônios) (VALAFAR, 2003) (FIGURA 4.4). Os neurônios da camada de entrada recebem um determinado tipo de sinal (variáveis preditoras), os neurônios da camada intermediária são responsáveis pelo aprendizado e os da camada de saída apresentam o resultado do processamento (variáveis resposta). Durante o processo de aprendizado, a rede ajusta a intensidade (pesos) de suas sinapses de acordo com o algoritmo de aprendizado e os dados observados.

FIGURA 4.4: ARQUITETURA DE UMA REDE NEURAL ARTIFICIAL TÍPICA



FONTE: Dang (2013)
LEGENDA: os círculos representam os neurônios na camada de entrada (*input layer*), oculta (*hidden layer*) e saída (*output layer*). As sinapses são representadas pelas setas e seus respectivos pesos por w. O sinal de entrada é x e o resultado é y.

4.1.2 Ajuste de modelo e Amostragem de Dados

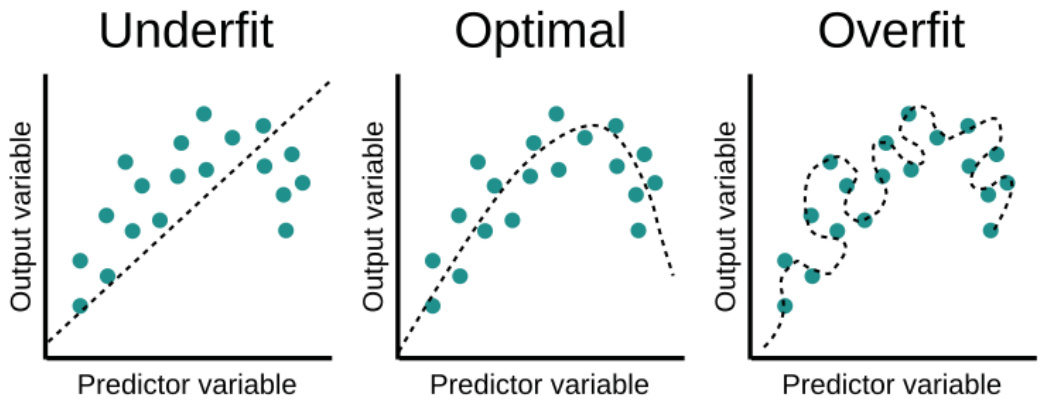
Modelos de aprendizado de máquina podem acabar se ajustando excessivamente aos dados, tornando-se ineficientes para inferir dados novos que não fizeram parte do conjunto utilizado no processo de aprendizado (JAMES *et al.*, 2013). Este fenômeno é chamado de *overfitting* ou sobreajuste (FIGURA 4.5) e acontece quando o modelo "decora"o padrão dos dados durante o processo de aprendizagem e não consegue extrapolar esse padrão para dados novos. O fenômeno contrário também pode ser observado, ou seja, o modelo de aprendizado não consegue encontrar as relações entre os dados durante a aprendizagem. Nesse caso, o fenômeno é chamado de *underfitting* ou subajuste. Para evitar tais fenômenos, é importante que o conjunto total de dados disponíveis seja dividido em três subconjuntos (FIGURA 4.6):

1. **Conjunto de treino:** dados utilizados para treinar o modelo, ou seja, definir as estimativas dos parâmetros para que ocorra o ajuste do modelo de aprendizado.

2. **Conjunto de validação:** dados utilizados para avaliar a performance de modelos com diferentes parâmetros.

3. **Conjunto de teste:** dados utilizados para avaliar a performance final do modelo já ajustado.

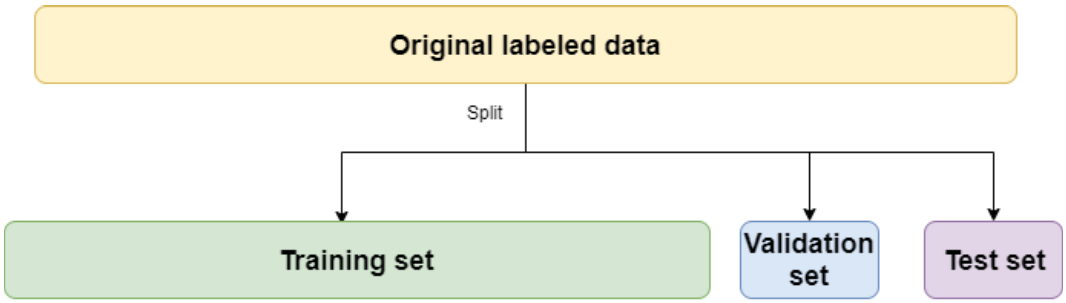
FIGURA 4.5: AJUSTE DO MODELO AOS DADOS



FONTE: Educative (2020)

LEGENDA: o eixo Y representa a variável resposta (*Output variable*). O eixo X representa a variável preditora (*Predictor variable*). No gráfico à esquerda, o modelo de aprendizado não encontrou o padrão correto dos dados, causando *underfitting*. No gráfico à direita, o modelo se adaptou excessivamente aos dados, causando *overfitting*. O modelo ideal é representado pelo gráfico do meio em que o padrão dos dados foi aprendido de maneira satisfatória.

FIGURA 4.6: CONJUNTOS DE DADOS



FONTE: Towards Data Science (2017)

LEGENDA: o conjunto original de dados (*Original labeled data*) é divido em três subconjuntos: conjunto de treino (*Train set*), validação (*Validation set*) e teste (*Test set*).

Também é possível combinar modelos de aprendizado diferentes, criando um conjunto de modelos que é chamado de *ensemble* (VALENTINI; MASULLI, 2002). Estudos mostram que tanto para problemas de regressão, quanto para problemas de classificação, *ensembles* são comumente mais acurados que os modelos individuais que o compõe (VALENTINI; MASULLI, 2002; BAUER; KOHAVI, 1999; DIETTERICH, 2000; FREUND; SCHAPIRE, 1996).

4.1.3 Aprendizado de Máquina e Aplicações na Biologia

Com o aumento da quantidade de dados disponíveis em bancos de dados biológicos, surgiu a necessidade do uso de técnicas que permitissem analisar grandes quantidades de dados de forma rápida e eficiente (CHICCO, 2017). Métodos de aprendizado de máquina, além de permitirem o processamento e análise de uma quantidade de dados extensa demais para ser analisada manualmente, são capazes de identificar padrões não-óbvios nos dados (principalmente quando o conhecimento humano é incompleto ou inacurado) e ainda realizar inferências e predições sobre dados novos (CHICCO, 2017; YIP; CHENG; GERSTEIN, 2013).

Técnicas de aprendizado de máquina vêm sendo bastante aplicadas para resolver demandas das áreas de biológicas e de saúde. Algumas dessas aplicações são elencadas abaixo:

- Discriminação de íntros e éxons no DNA (VALAFAR, 2003).
- Anotação de genomas (YIP; CHENG; GERSTEIN, 2013).
- Classificação de genes e famílias de proteínas (VALAFAR, 2003).
- Identificação de espécies através de imagens (VALAN *et al.*, 2019)
- Predição de estrutura secundária e terciária de proteínas (VALAFAR, 2003).
- Identificação morfológica de células-tronco (KUSUMOTO; YUASA, 2019).

Nas análises filogenéticas, o uso de aprendizado de máquina permite a classificação dos dados não somente provenientes de métodos clássicos, como, por exemplo, alinhamento múltiplo de sequências (HALGASWATHTHA *et al.*, 2012). Dados de alinhamento podem ser usados para o treinamento de modelos de aprendizado de máquina (HALGASWATHTHA *et al.*, 2012; FIORAVANTI *et al.*, 2018), porém métodos livres de alinhamento, como o SWeeP (DE PIERRE *et al.*, 2020), permitem que a classificação seja alcançada sem a etapa do alinhamento.

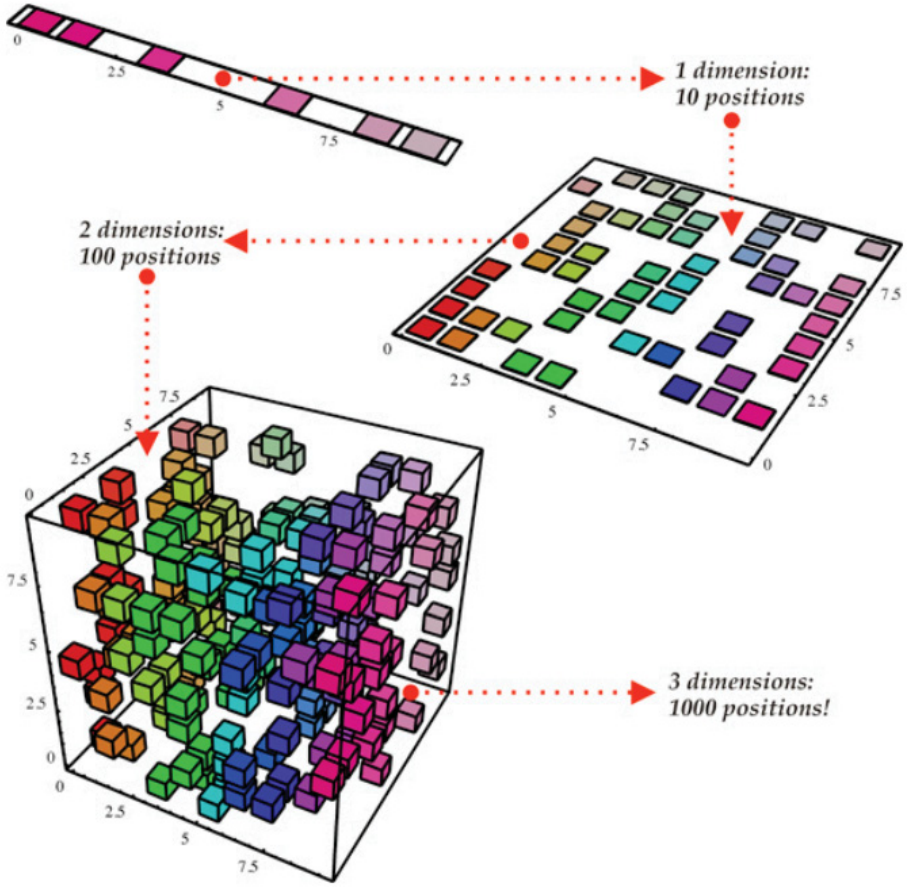
4.2 REDUÇÃO DE DIMENSIONALIDADE

A redução de dimensionalidade tem como objetivo representar dados de alta dimensão em uma representação significativa desses dados com a dimensionalidade reduzida (VAN DER MAATEN; POSTMA; Van Den Herik, 2009). Idealmente, a redução de dimensão deve chegar à dimensionalidade intrínseca que corresponde ao número mínimo de parâmetros necessários para representar as propriedades dos dados observados (VAN DER MAATEN; POSTMA; Van Den Herik, 2009). Com esta redução, é possível representar os dados de uma forma mais compacta de modo que a aplicação de algoritmos, principalmente os mais extensos e complexos, seja facilitada (AGGARWAL, 2015). Algumas técnicas utilizadas para redução de dimensionalidade seguem elencadas abaixo:

- **Amostragem de dados:** indivíduos são amostrados da população, criando uma base de dados menor, mas que represente a população total.
- **Seleção de atributos (*feature selection*):** apenas alguns atributos são utilizados no processo analítico.
- **Redução com rotação de eixo:** a correlação dos dados é aproveitada para representá-los em um número menor de dimensões.
- **Redução com transformação de tipo:** acontece a portabilidade do tipo do dado, convertendo-os em tipos de menor complexidade.

Nesse trabalho, serão utilizadas técnicas de redução com rotação de eixo, mais especificamente o método SWeeP (vide seção 3.1) e a análise de componentes principais (PCA - *Principal Component Analysis*). A PCA é um técnica que constrói uma representação de baixa dimensão dos dados através de combinações lineares que descreve a maior variância possível das variáveis originais (FIGURA 4.7) (VAN DER MAATEN; POSTMA; Van Den Herik, 2009)

FIGURA 4.7: PCA - ANÁLISE DE COMPONENTES PRINCIPAIS



FONTE: Free Code Camp (2019)

LEGENDA: redução de dimensionalidade. Os dados originais possuíam 3 dimensões e 1000 posições. A primeira redução diminui para 2 dimensões e 100 posições. A segunda redução diminui para 1 dimensão e apenas 10 posições.

5 OBTENÇÃO DO MODELO

O objetivo desse capítulo é apresentar o modelo para a construção da filogenia molecular da família Formicidae. A metodologia utiliza fonte de dados de diferentes proteínas, unificando toda a informação num único modelo que consegue inferir dados faltantes.

5.1 MATERIAL E MÉTODOS

O método para a obtenção do modelo foi constituído de seis partes: levantamento e coleta de dados, análise e curadoria dos dados, vetorização das sequências, treinamento e validação das redes neurais, predição dos dados moleculares e construção da filogenia.

5.1.1 Levantamento e Coleta de Dados

Primeiramente foi realizado o levantamento e coleta dos dados moleculares das formigas disponíveis no banco de dados de proteínas do NCBI. Para isto, usou-se um robô desenvolvido em *Python*[1] que realizou o *download* de todas as sequências de proteínas classificadas como pertencentes à família Formicidae do banco de dados de proteínas do NCBI. As sequências foram baixadas em formato FASTA e, a partir das anotações dos arquivos, pode-se levantar a quantidade de sequências completas e de fragmentos.

5.1.2 Análise e Curadoria dos Dados

A classificação mais recente de subfamílias, gêneros e espécies de formigas foi obtida do banco de dados AntWiki (ANTWIKI, 2020) que disponibiliza uma planilha da classificação taxonômica mais recente no formato de aquivo XLSX[2].

Utilizando um robô próprio desenvolvido em MATLAB®[3], foram selecionadas as sequências que possuem a proteína identificada e classificação a nível de espécie. As sequências selecionadas foram agrupadas e então separadas em arquivos Multi-FASTA por proteína. Dessa forma, cada arquivo Multi-FASTA continha todas as sequências de uma proteína. Todas as análises seguintes foram feitas utilizando as sequências das proteínas que apresentaram mais de 1.000 sequências (seis proteínas que são descritas na seção 5.2.1), para que houvesse informação suficiente para o treinamento das redes neurais. Cada arquivo Multi-FASTA foi então separado por espécie, gerando vários arquivos Multi-FASTA cada qual contendo todas as sequências de uma determinada espécie para determinada proteína.

[1]Disponível no endereço <https://www.bioinfodiscentes.com.br/python>. O robô foi executado no dia 26 de junho de 2020.

[2]Disponível no endereço <https://www.antwiki.org/wiki/images/0/0c/AntWiki_Regional_Taxon_List.txt>. A planilha utilizada nesse trabalho contém atualizações de até dia 01 de maio de 2020.

[3]Disponível no endereço <https://github.com/moniquesch/manage-fasta-files>.

Paralelamente, as espécies que possuíam sequências das seis proteínas trabalhadas foram selecionadas, totalizando 375 espécies (vide seção 5.2.1). Essas sequências foram separadas por espécies, mas não por proteína, ou seja, foram gerados 375 arquivos Multi-FASTA com sequências de seis proteínas diferentes.

5.1.3 Vetorização das Sequências

Para a vetorização das sequências, foi utilizado o algoritmo SWeeP (DE PIERRE *et al.*, 2020) que permite a representação vetorial das sequências reais de aminoácidos sem a perda da comparabilidade entre elas. Nesse trabalho, foi utilizado o SWeeP com máscara padrão e projeção de 200, ou seja, as sequências são representadas por um vetor numérico com 200 coordenadas.

Dois tipos de matrizes SWeeP foram geradas pelo algoritmo: **matriz de proteínas** e **matriz alvo**. A partir dos Multi-FASTAS separados por espécie e por proteínas, foram geradas seis matrizes de proteínas, uma para cada proteína, em que cada linha da matriz representa a vetorização de todas as sequências da proteína para uma determinada espécie (FIGURA 5.1). Em resumo, as matrizes de proteínas contêm a informação de todas as sequências com identificação a nível de espécie da proteína. A matriz alvo é gerada a partir de todas as sequências, independente da proteína, das 375 espécies citadas anteriormente (FIGURAS 5.2 e 5.7), sendo cada linha da matriz uma representação vetorial de todas as sequências de uma determinada espécie. Ou seja, a matriz alvo irá conter toda a informação molecular das espécies que possuem sequências de todas as proteínas abordadas.

5.1.4 Treinamento e Validação das Redes Neurais

Os dados utilizados para o treinamento das redes neurais foram as matrizes de proteínas e a matriz alvo. Para a preparação dos dados, primeiramente, somente as espécies contidas na matriz alvo (375 espécies) foram selecionadas nas matrizes de proteínas. Por segundo, as matrizes foram divididas randomicamente em três grupos: grupo de treino (70% dos dados), grupo de validação (15% dos dados) e grupo de teste (15% dos dados).

Utilizando a função *Fitting Neural Network* do MATLAB®versão 2018b[4], um *ensemble* de redes neurais foi criado para inferir a matriz alvo a partir das matrizes de proteína. Cada rede individual tinha como entrada uma matriz de proteína e, como saída, uma coluna da matriz alvo (FIGURA 5.3). Ou seja, o *ensemble* de redes neurais foi composto por 200 redes individuais para prever a matriz alvo por completo (FIGURA 5.4).

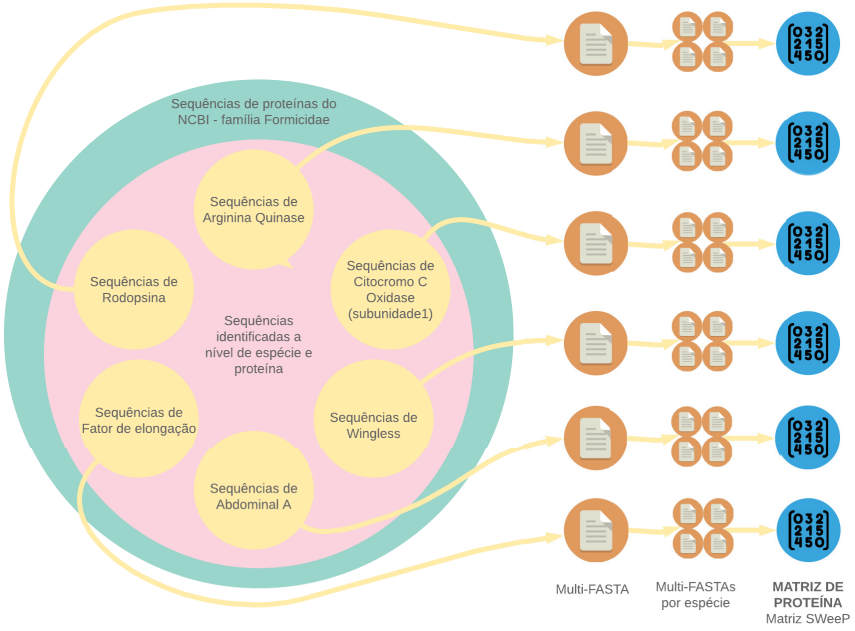
Cada rede individual foi estruturada com uma camada oculta de 15 neurônios. Para cada coluna da matriz alvo, 30 redes foram treinadas e a com melhor performance foi escolhida para as análises seguintes. Esse processo foi repetido com duas funções de treino diferentes, sendo elas: Levenberg-Marquardt (LM), e *Scaled Conjugate Gradient* (SCG). A função com

[4]MATLAB versão 9.4.0.813654 (R2018a) The MathWorks, Inc., Natick, Massachusetts, United States.

melhor performance, ou seja, menor MSE (erro quadrático médio), foi utilizada para a predição dos dados. Ao final do processo de treinamento, foram obtidos seis *ensembles*, um para cada matriz de proteína.

Para a validação das redes, como a predição das matrizes alvos é feita por um *ensemble* de 200 redes independentes, foi calculada o MSE de cada rede individual e valor de erro do *ensemble* se deu pela média dos erros das redes individuais. A validação foi feita nas seis matrizes alvos que foram preditas durante o processo.

FIGURA 5.1: FLUXOGRAMA PARA A CRIAÇÃO DAS MATRIZES DE PROTEÍNA



FONTE: a Autora (2020)

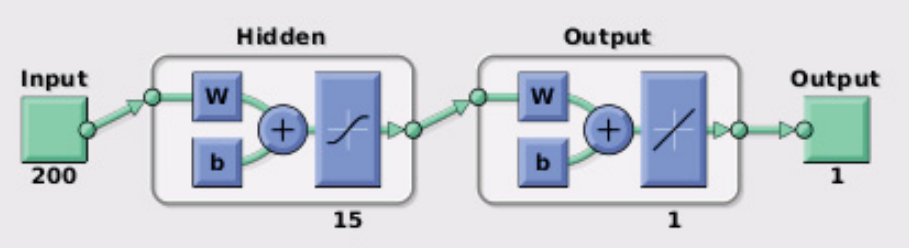
LEGENDA: Seis subconjuntos de sequências foram utilizadas para a criação dos arquivos Multi-FASTA. Todas as sequências selecionadas possuíam classificações a nível de espécie. Na matriz de proteínas, cada linha da matriz passou a representar a vetorização das sequências da proteína para uma espécie.

FIGURA 5.2: FLUXOGRAMA PARA A CRIAÇÃO DA MATRIZ ALVO

FONTE: a Autora (2020)

LEGENDA: Todas as sequências das 375 espécies foram agrupadas em um único arquivo, para depois serem separadas por espécies para a construção da matriz alvo. Cada linha da matriz alvo contém informação das sequências das seis proteínas para uma espécie.

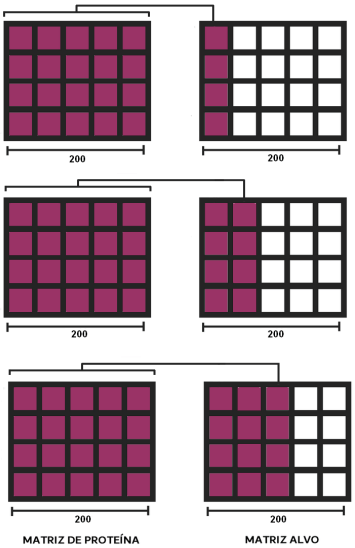
FIGURA 5.3: ESTRUTURA DAS REDES NEURAIS INDIVIDUAIS



FONTE: a Autora (2020)

LEGENDA: Cada rede individual tem como entrada (*input*) uma matriz 200xn, em que n é o número de espécies contidas no grupo de treino e, como saída (*output*), um vetor 200x1 que corresponde a uma coluna da matriz alvo. A camada oculta (*hidden*) foi estruturada com 15 neurônios

FIGURA 5.4: ESQUEMA DO *ENSEMBLE* DE REDES NEURAIS



FONTE: a Autora (2020)

LEGENDA: Cada rede individual utiliza a matriz de proteína completa para inferir uma coluna da matriz alvo. O *ensemble* composto por 200 redes infere as 200 colunas da matriz alvo

5.1.5 Predição das Matrizes Alvo

Os seis *ensembles* finais foram então utilizados para fazer a predição das demais espécies, ou seja, das espécies que possuíam alguma informação molecular acerca de pelo menos uma das seis proteínas abordadas, mas não de todas as seis proteínas. Para tal, para cada proteína, utilizou-se a matriz de proteína completa (com todas as espécies) como entrada do ensemble e, como saída a matriz alvo final para todas as espécies da proteína. Como resultado, seis matrizes alvo foram geradas, uma para cada proteína. A informação das seis matrizes alvo foi sumarizada através do cálculo da média dos valores das matrizes, gerando uma matriz final contendo um

vetor de tamanho 200 para cada espécie que possui algum dado acerca das proteínas abordadas, totalizando 2.918 espécies (vide seção 5.2.1). Essa matriz gerada será chamada de **Matriz Alvo Molecular (MAM)** nesse trabalho.

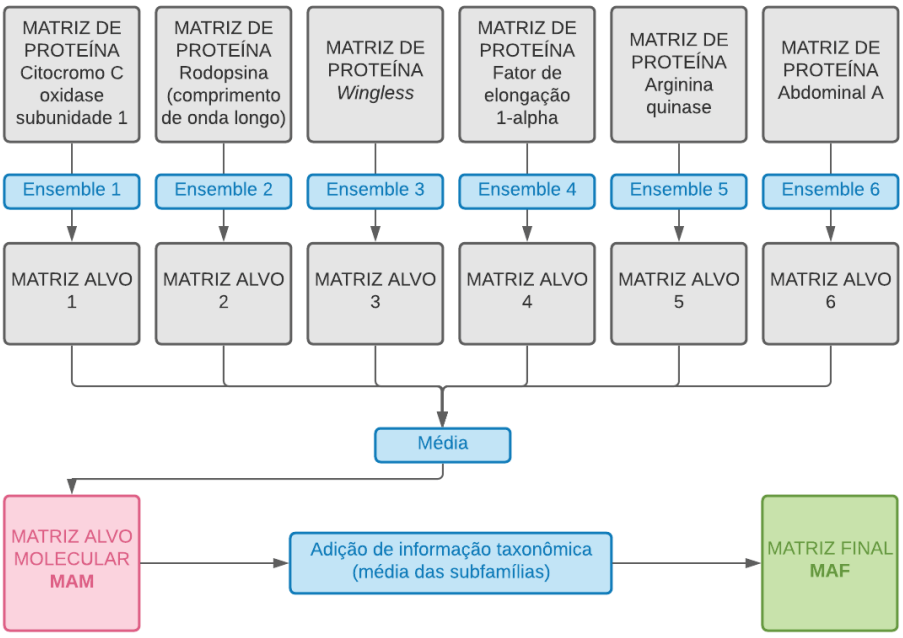
5.1.6 Construção da Filogenia Molecular

Com o objetivo de adicionar mais informação taxonômica ao modelo, a partir da MAM, primeiramente foi calculada a média dos vetores para cada subfamília. Depois, os vetores da MAM e os vetores das médias das subfamílias foram normalizados separadamente através da fórmula de Min-Max. Então, a MAM normalizada foi concatenada aos vetores normalizados das médias das subfamílias, gerando uma uma matriz de tamanho 2981x400 que será chamada nesse trabalho de **Matriz Final (MAF)**. A partir da MAF, foi calculada a matriz de distância pelo método de distância par-a-par Euclidiana (*pairwise Euclidean distance*) e a árvore filogenética foi construída pelo método de agrupamento de vizinhos (*neighbor-joining*). A validação da árvore foi feita através de comparação com as filogenias de DE PIERRE *et al.* (2020), Branstetter *et al.* (2017) e Economo *et al.* (2015). Paralelamente, foi feita uma análise de componentes principais da MAM para observar o agrupamento entre as subfamílias.

Em resumo, as análises feitas nesse trabalho utilizarão duas matrizes diferentes (FIGURA 5.5):

- **Matriz Alvo Molecular (MAM)**: sumarização das seis matrizes alvo preditas.
- **Matriz Final (MAF)**: concatenação da MAM com as médias das subfamílias.

FIGURA 5.5: FLUXOGRAMA DA PREDIÇÃO DOS DADOS MOLECULARES



FONTE: a Autora (2020)

5.2 RESULTADOS

5.2.1 Levantamento, Análise e Curadoria dos Dados

Foram encontradas e baixadas 768.891 sequências de proteínas de organismos pertencentes à família Formicidae. Dentre essas sequências, grande parte possuía somente a classificação a nível de família ou a nível de gênero e somente 3.066 sequências estavam classificadas a nível de espécie.

Inúmeras proteínas foram identificadas, porém somente seis proteínas possuíam ao menos 1.000 sequências identificadas à nível de espécie, sendo elas: Citocromo C oxidase subunidade 1 (gene COI), Rodopsina (comprimento de onda longo) (gene W_Rh), *Wingless* (gene wg), Fator de elongação 1-alpha (gene EF-1 alpha), Arginina quinase (gene ArgK) e Abdominal A (gene abd-A).

Praticamente todas as sequências são parciais (fragmentos) e não contemplam a proteína por completo (FIGURA 5.6 e TABELA 5.1). A proteína com o menor número de sequências completas foi a rodopsina, pois 100% das sequências eram fragmentos (TABELA 5.1). As proteínas que possuíam o maior número de sequências completas foram, respectivamente, Fator de Elongação 1-alpha (3,26%), Arginina Quinase (4,25%) e Abdominal A (5,79%). A proteína Citocromo C Oxidase subunidade 1 foi a proteína que mais apresentou sequências (35.212 sequências), além de possuir o maior número de espécies representadas (2.505 espécies). Por outro lado, a proteína *Wingless* apresentou o maior número de gêneros distintos (272 gêneros). Citocromo C Oxidase subunidade 1 também foi a proteína que apresentou a maior quantidade de espécies únicas (espécies que não aparecem em nenhuma outra proteína), com 1.085 espécies (FIGURA 5.7). Ao todo, 375 espécies foram encontradas em comum nas seis proteínas e foram utilizadas para a construção das matrizes alvo.

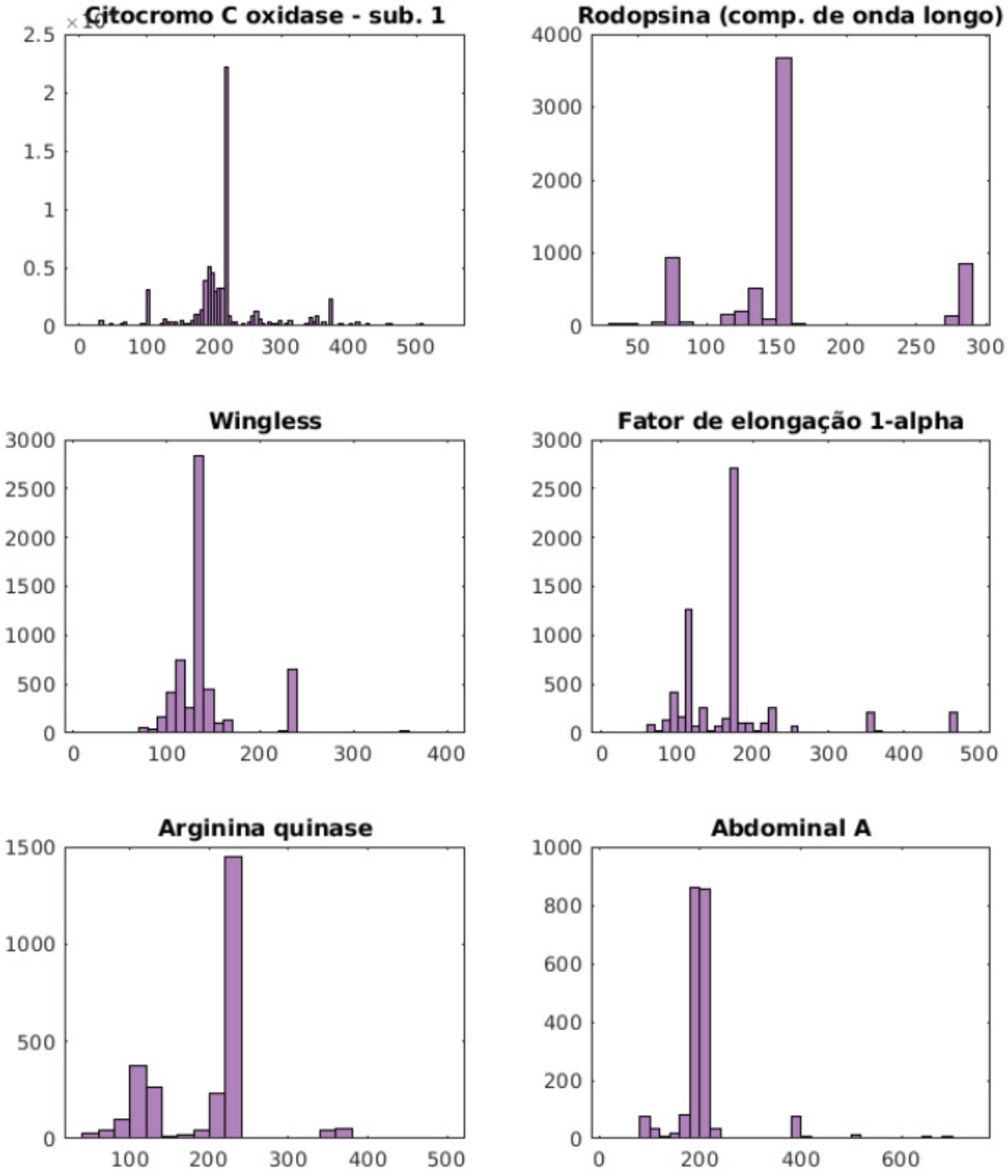
TABELA 5.1: Informações acerca das sequências

Proteína	Gêneros abrangidos	Espécies abrangidas	Total de sequências	Total de seq. parciais	Total de seq. completas
Citocromo C oxidase subunidade 1	198	**2505**	**35212**	**35090**	**122 (0,35%)**
Rodopsina (comprimento de onda longo)	268	1682	3392	3392	0
Wingless	**272**	1463	2972	2957	15 (0,5%)
Fator de elongação 1-alpha	260	1233	3250	3144	106 (3,26%)
Arginina quinase	249	993	1341	1284	57 (4,25%)
Abdominal A	256	844	1053	992	61 (5,79%)

FONTE: a Autora (2020)

NOTA: Proteínas que possuíam mais de 1000 sequências cadastradas no NCBI. Na coluna "Total de seq. completas", entre parênteses se encontram as porcentagens de sequências completas em relação ao número total de sequências. Em negrito, destacam-se os maiores valores da coluna.

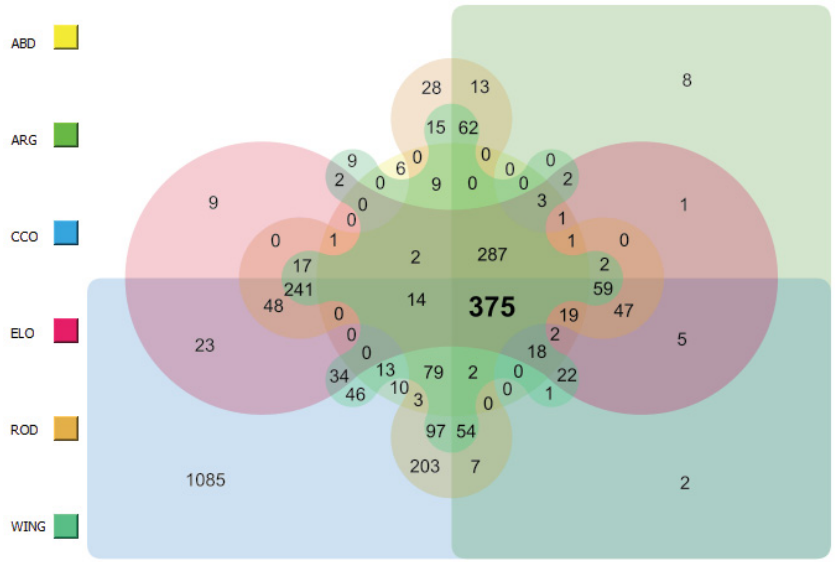
FIGURA 5.6: DISTRIBUIÇÃO DA QUANTIDADE DE SEQUÊNCIAS POR TAMANHO



FONTE: a Autora (2020)

LEGENDA: O eixo x corresponde ao tamanho da sequência medida pela quantidade de aminoácidos. O eixo y corresponde à quantidade de sequências.

FIGURA 5.7: DIAGRAMA DE SOBREPOSIÇÃO DE ESPÉCIES



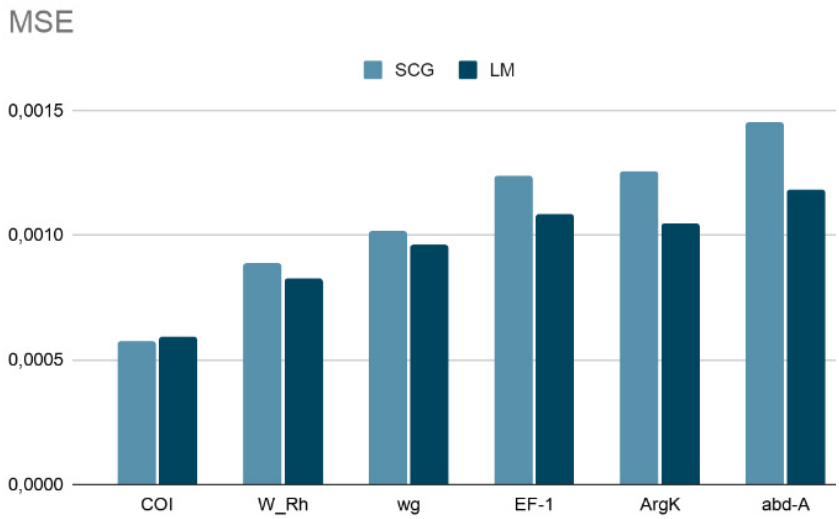
FONTE: a Autora (2020)

LEGENDA: Diagrama de Edwards mostrando a quantidade de espécies abrangidas pelas proteínas, bem como a interseção dessas espécies. Em destaque, as 375 espécies que são abrangidas pelas seis proteínas. ABD: abdominal A; ARG: arginina quinase; CCO: citocromo C oxidase subunidade 1; ELO: fator de elongação 1-alpha; ROD: rodopsina (comprimento de onda longo); WING: *Wingless*.

5.2.2 Redes Neurais

A função de treino com a melhor performance de inferência foi a função LM (Levenberg-Marquardt) (FIGURA 5.8), pois obteve os menores valores de MSE durante a validação dos *ensembles* das seis proteínas.

FIGURA 5.8: PERFORMANCE DAS FUNÇÕES DE TREINO



FONTE: a Autora (2020)

LEGENDA: MSE para diferentes funções de treino utilizadas nos *ensembles* de cada proteína.

5.2.3 Predição das Matrizes Alvo

Seis matrizes alvo foram inferidas a partir das seis matrizes de proteínas completas, cada uma delas contendo todas as espécies de sua respectiva proteína. Seguem abaixo as matrizes alvo geradas e a respectiva quantidade de espécies abrangidas:

1. Matriz alvo de citocromo C oxidase subunidade 1 abrangendo 2.505 espécies;
2. Matriz alvo de rodopsina (comprimento de onda longo) abrangendo 1.682 espécies;
3. Matriz alvo de *wingless* abrangendo 1.463 espécies;
4. Matriz alvo de fator de elongação 1-alpha abrangendo 1.233 espécies;
5. Matriz alvo de arginina quinase abrangendo 993 espécies;
6. Matriz alvo de abdominal A abrangendo 844 espécies.

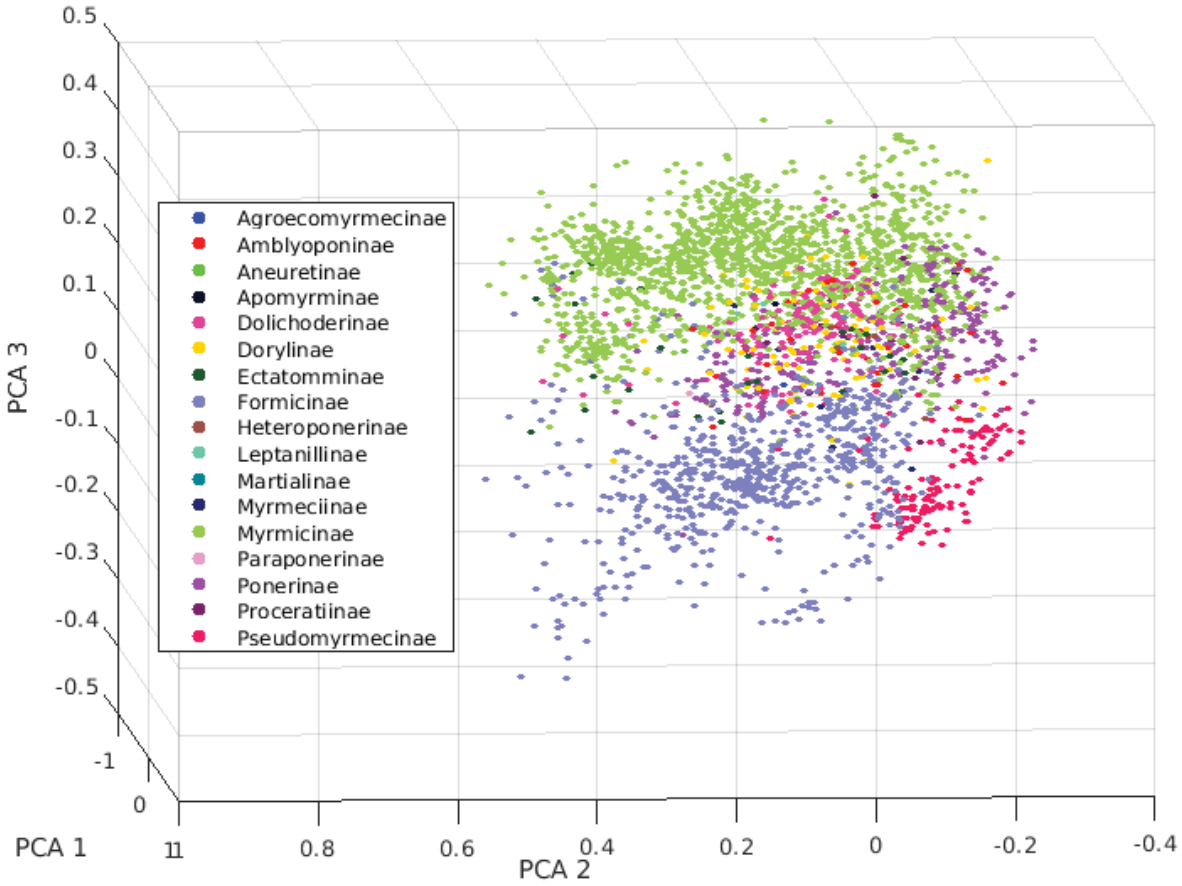
A partir da sumarização da informação dessas seis matrizes alvo, foi gerada a MAM. Com a adição da informação taxonômica, gerou-se a MAF que contemplou 2.981 espécies.

5.2.4 Filogenia Molecular (2.981 espécies)

A filogenia alcançada no trabalho engloba todas as espécies que possuem alguma informação, parcial ou completa, das seis proteínas abordadas, totalizando 2.981 espécies de formigas de 281 gêneros e das 17 subfamílias, correspondendo a 21,58% do total de espécies (FIGURA 5.10). Na base, a filogenia dividiu-se em dois ramos basais. O primeiro ramo basal é composto pelas subfamília Leptanillinae e pelas subfamílias que compõe o clado Poneroide (Agroecomyrmecinae, Amblyoponinae, Ponerinae e Proceratiinae) (FIGURA 5.11). O segundo ramo basal compreende a subfamília Martialinae, Apomyrminae, Paraponerinae e as subfamílias pertencentes ao clado Formicoide, sendo elas: Aneuretinae, Dolichoderinae, Dorylinae, Ectatomminae, Formicinae, Heteroponerinae, Myrmeciinae, Myrmicinae e Pseudomyrmecinae (FIGURA 5.12).

Na análise de componentes principais feita a partir da MAM (FIGURA 5.9) é possível observar a formação de clusters entre as subfamílias com mais espécies. Os dois maiores clusters são formados pelas subfamílias Myrmicinae e Formicinae.

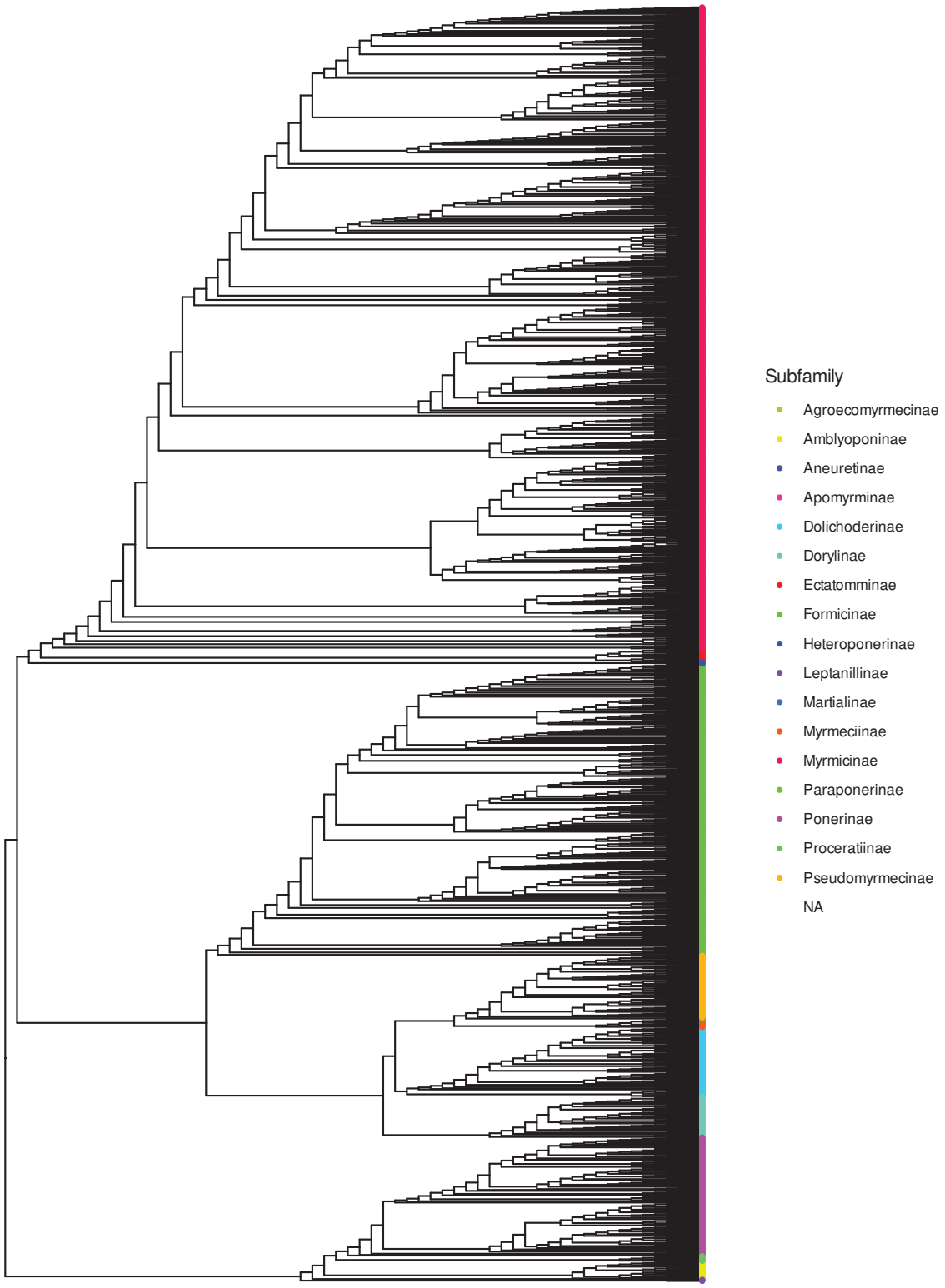
FIGURA 5.9: PCA - ANÁLISE DE COMPONENTES PRINCIPAIS



FONTE: a Autora (2020)

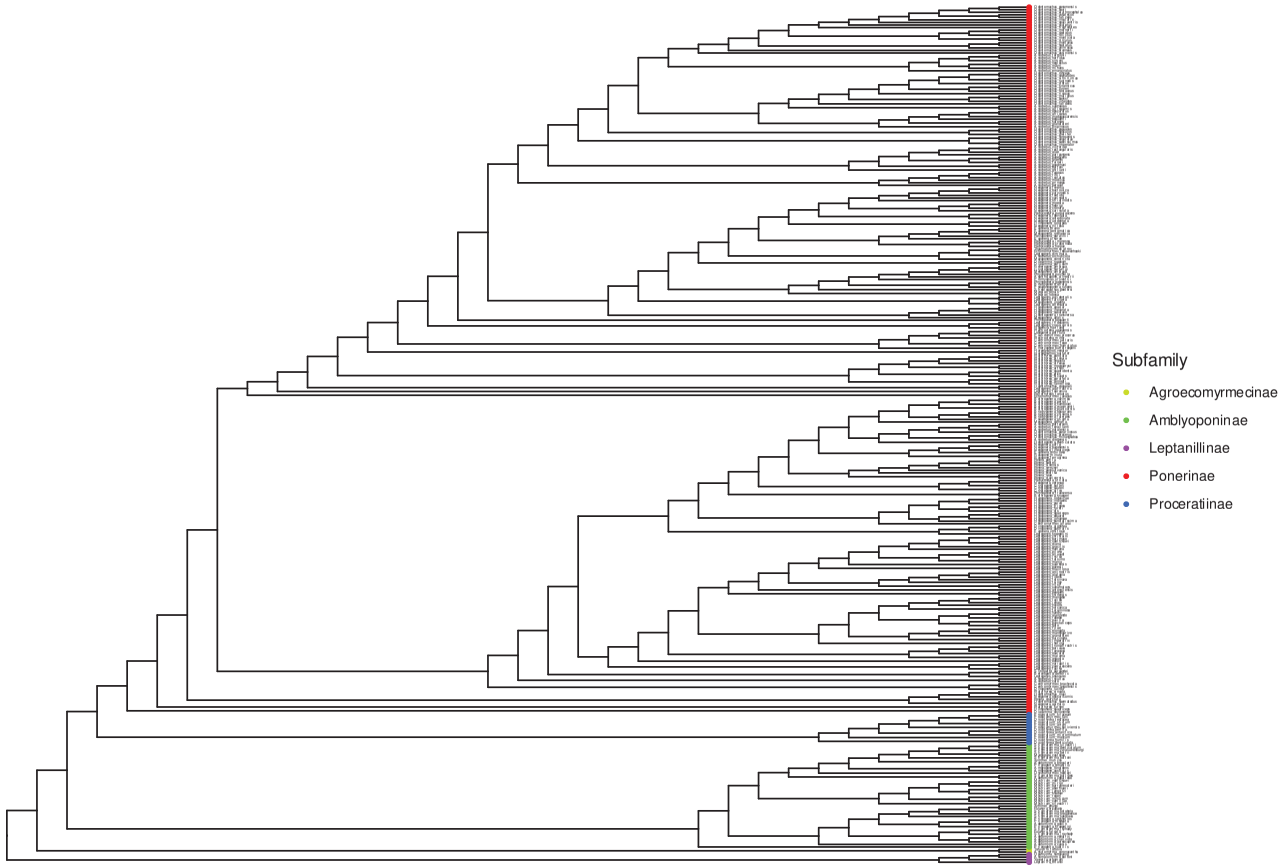
LEGENDA: os eixos correspondem aos três principais componentes da MAM

FIGURA 5.10: FILOGENIA DE DADOS MOLECULARES - COMPLETA



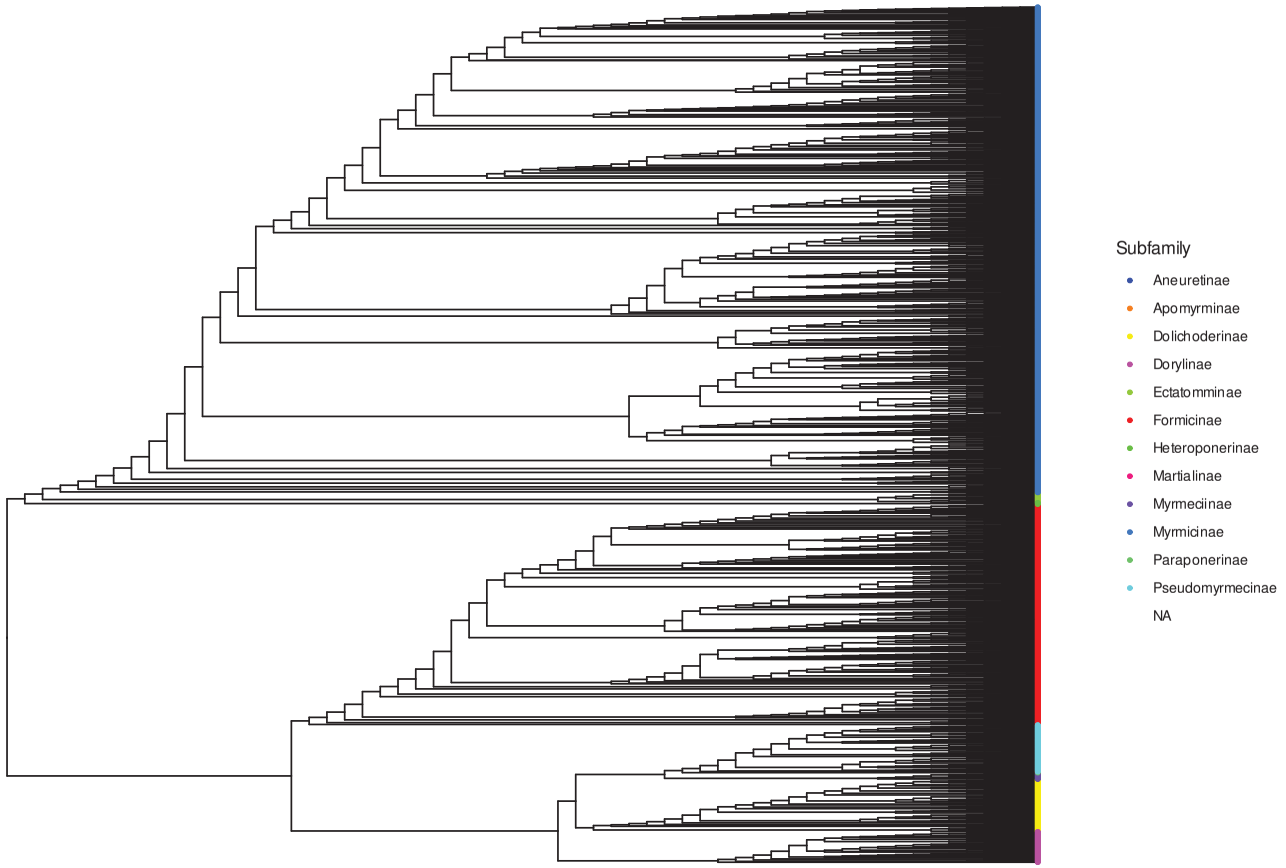
FONTE: a Autora (2021)

FIGURA 5.11: CLADO PONEROIDE



FONTE: a Autora (2021)

FIGURA 5.12: CLADO FORMICOIDE



FONTE: a Autora (2021)

5.3 DISCUSSÃO

A primeira matriz gerada, a MAM, conseguiu agrupar as principais subfamílias, como pode ser observado no gráfico de dispersão de PCA (FIGURA 5.9), porém não teve resolução suficiente para resolver a filogenia, gerando uma árvore que não agrupou os clados de maneira correta (APÊNDICE A). Para melhorar o modelo adicionando mais informação ao mesmo, foi utilizada o método de concatenação da MAM com a matriz de vetores das médias das subfamílias, como descrito na seção 5.1.6, criando a MAF. Dessa forma, criou-se uma tendência para que espécies da mesma subfamília se agrupassem, porém não interferindo nas relações entre as subfamílias. A filogenia construída a partir da MAF, além de agrupar corretamente as subfamílias, conseguiu recuperar as relações das mesmas de forma coerente.

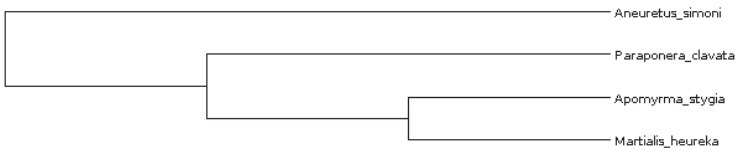
O modelo final obtido através de técnicas de redução de dimensionalidade e aprendizado de máquina obteve sucesso processando uma grande quantidade de dados simultaneamente e conseguiu construir uma filogenia muito próxima às filogenias propostas na literatura. Assim como em propostas anteriores (MOREAU *et al.*, 2006; FERNÁNDEZ; LATTKE, 2015; BRANSTETTER *et al.*, 2017; KÜCK *et al.*, 2011), a filogenia alcançada nesse trabalho dividiu a família Formicidae em dois clados principais: Poneroide e Formicoide.

O grupo Poneroide (FIGURA 5.11) ficou monofilético, corroborando a filogenia apresentada por Branstetter *et al.* (2017). Porém, na literatura, a família Leptanilinae aparece como um grupo basal irmão de todas as formigas Poneroides e Formicoides. Já na filogenia gerada neste trabalho, a família aparece como grupo irmão somente do clado Poneroide, mas mantendo-se como grupo mais basal. Agroecomyrmicinae aparece como grupo mais basal entre as Poneroides, seguida por Amblyoponinae e Proceratinae. A subfamília Ponerinae aparece como a mais apical. Na literatura, ainda há discussões acerca das relações entre as subfamílias do clado Poneroide e suas posições não estão bem estabelecidas (FERNÁNDEZ; LATTKE, 2015; BRANSTETTER *et al.*, 2017; BOROWIEC *et al.*, 2019).

O clado Formicoide divide-se em dois ramos: um com espécies da subfamília Myrmicinae, Heteroponerinae e Ectatomminae e outro com as demais subfamílias (FIGURA 5.12). Na literatura, Myrmicinae é grupo irmão Heteroponerinae e Ectatomminae (BRANSTETTER *et al.*, 2017). Na filogenia apresentada neste trabalho, as três subfamílias continuam formando um grupo monofilético com a única diferença que Heteroponerinae aparece como grupo irmão de Myrmicinae e Ectatomminae. Nos ramos com as demais subfamílias, o grupo formado pela subfamília Formicinae está como irmão do grupo formado pelas subfamílias Pseudomyrmecinae, Myrmeciinae, Dolichoderinae e Dorylinae. Na literatura, Aneuretinae e Dolichoderinae formam um grupo irmão de Myrmeciinae e Pseudomyrmecinae e Dorylinae aparece como grupo mais basal de Formicoide, irmão de todas as outras subfamílias (BRANSTETTER *et al.*, 2017).

As subfamílias Martialinae, Apomyrminae, Paraponerinae e Aneuretinae aparecem todas em um só ramo (FIGURA 5.13) e no clado das formigas Formicoides. Segundo a literatura (BRANSTETTER *et al.*, 2017; ANTWIKI, 2020; FERNÁNDEZ; LATTKE, 2015; KÜCK *et al.*,

FIGURA 5.13: MARTIALINAE + APOMYRMINAE + PARAPONERINAE + ANEREURETUS



FONTE: a Autora (2020)
LEGENDA: ramo extraído da Filogenia da MAF. O ramo é completamente formado por subfamílias monotípicas.

2011), a única subfamília que realmente pertence ao clado Formicoide é a Anereutinae, pois as subfamílias Paraponerinae e Apomyrminae pertencem ao clado Poneroide e a subfamília Martialinae, juntamente com Leptanilinae, estaria mais basal na filogenia, fora desses dois grupos. Segundo o AntWiki (2020), nessas quatro subfamílias há apenas um gênero e uma espécie representando toda a subfamília, sendo elas:

- *Martialis heureka* (Rabeling e Verhaagh, 2008) para Martialinae;
- *Apomyrma stygia* (Brown, Gotwald e Levieux, 1971) para Apomyrminae;
- *Paraponera clavata* (Fabricius, 1775) para Paraponerinae;
- *Aneuretus simoni* (Emery, 1893) para Aneuretinae;

O fato das subfamílias possuírem apenas uma espécie, faz com que a mesma seja representada somente por um vetor na MAM, empobrecendo os dados de entrada das redes de predição e, consequentemente, alterando a posição dessas subfamílias na filogenia. As divergências encontradas nas outras subfamílias provavelmente são decorrentes da qualidade dos dados depositados no NCBI.

Como mostrado nos resultados (seção 5.2.1), as sequências de proteínas disponíveis no NCBI são bastante numerosas, porém com pouca qualidade de informação. Menos de 0,4% das sequências possuíam informação taxonômica a nível de espécie e informação sobre a proteína. Em diversas sequências a anotação dava como classificação somente a família Formicidae e a proteína era descrita como "desconhecida"ou "proteína de baixa qualidade". Caso essas sequências fossem depositadas com melhores anotações, o modelo poderia ser enriquecido com mais informações. Outro fator que pode ter afetado as relações entre as subfamílias é a fragmentação das sequências, visto que sequências parciais não possuem toda a informação da proteína. A falta da informação completa da proteína e a qualidade das sequências são fatores que influenciam as análises e podem enviesar a filogenia (KEMENA; DOHMEN; BORNBERG-BAUER, 2019; XIA, 2014).

Apesar da pouca qualidade dos dados e da heterogeneidade na distribuição dos dados (vide seção 1.2), as técnicas livres de alinhamento, de redução de dimensionalidade de e aprendizado de máquina mostraram-se eficazes em processar simultaneamente uma grande quantidade de dados, sintetizar a informação das sequências e entender os padrões de cada grupo.

6 COMPARAÇÕES FILOGENÉTICAS

O objetivo desse capítulo é avaliar mais detalhadamente a filogenia final gerada nesse trabalho. Para isso, foi feita comparação com quatro filogenias: duas filogenias construídas a partir do método SWeeP (uma de proteomas completos e outra de proteomas mitocondriais), uma filogenia feita usando técnicas de alinhamento de UCEs e uma filogenia somente do gênero *Pheidole*.

6.1 MATERIAL E MÉTODOS

As filogenias a serem usadas nas comparações foram obtidas das seguintes maneiras:

- **Filogenia de proteomas completos:** os proteomas disponíveis de organismos da família Formicidae foram baixados do banco de dados UniProt[1] (BATEMAN, 2019). Como grupo externo foi utilizado o proteoma de referência da abelha *Apis mellifera*. Os proteomas foram vetorizados utilizando o método SWeeP [2] com máscara 11011101 e projeção vetorial de 14.000 dimensões.
- **Filogenia de proteomas mitocondriais:** o ramo que continha as espécies de formigas foi extraído da árvore de proteomas mitocondriais apresentada no trabalho de DE PIERRE *et al.* (2020).
- **Filogenia de UCE:** filogenia apresentada no trabalho de Branstetter *et al.* (2017). Os dados provém do sequenciamento de 2590 loci de UCE de Hymenoptera.
- **Filogenia de *Pheidole*:** filogenia apresentada no trabalho de Economo *et al.* (2015). Os dados provém do sequenciamento dos loci His3.3B, Lop1, GRIK2, unc_4, LOC15, CAD1, EF-1α F2, Top1 e CO1.

Nas quatro comparações, as espécies correspondentes na MAF foram selecionadas e quatro árvores foram geradas a partir dos vetores dessas espécies (Filogenia 1, 2, 3 e 4). A construção das árvores seguiu a metodologia explicada na seção 5.1.6.

6.2 RESULTADOS

6.2.1 Filogenia 1 X Filogenia de Proteomas completos

O UniProt possui 18 proteomas de formigas cadastrados, sendo 13 deles proteomas de referência (*Atta cephalotes* (Linnaeus, 1758), *Nylanderia fulva* (Mayr, 1862), *Acromyrmex*

[1] Acesso em 11/01/2021

[2] Algoritmo adaptado para suportar máscaras de maior tamanho e com projeção virtual.

echinatior (Forel, 1899), *Trachymyrmex septentrionalis* (McCook, 1881), *Atta colombica* (Guérin-Méneville, 1844), *Trachymyrmex zeteki* (Weber, 1940), *Camponotus floridanus* (Buckley, 1866), *Trachymyrmex cornetzi* (Forel, 1912), *Harpegnathos saltator* (Jerdon, 1851), *Cyphomyrmex costatus* (Mann, 1922), *Lasius niger* (Linnaeus, 1758), *Ooceraea biroi* (Forel, 1907) e *Temnothorax longispinosus*) e outros 5 proteomas (*Ooceraea biroi*, *Temnothorax curvispinosus* (Roger, 1863), *Solenopsis invicta* (Buren, 1972), *Pogonomyrmex barbatus* (Smith, F., 1858) e *Dinoponera quadriceps* (Kempf, 1971)). Maiores informações sobre os proteomas, como contagem de proteínas, BUSCO Score e CPD, se encontram na FIGURA 6.1.

FIGURA 6.1: INFORMAÇÕES SOBRE OS PROTEOMAS

Proteome ID	Organism	Protein count	BUSCO Single Duplicated Fragmented Missing	CPD
UP000005205	Atta cephalotes (Leafcutter ant)	10634	n:5991 C:95.3% (S:93.9% D:1.5%) F:2.5% M:2.2% hymenoptera_odb10	Close to Standard
UP000479987	Nylanderia fulva	426	n:5991 C:0% (S:0% D:0%) F:0% M:100% hymenoptera_odb10	Outlier
UP000007755	Acromyrmex echinatior (Panamanian leafcutter ant) (Acromyrmex octospinosus echinatior) (Strain: colony Ae372)	13962	n:5991 C:89.5% (S:89.1% D:0.4%) F:3.8% M:6.7% hymenoptera_odb10	Standard
UP000078541	Trachymyrmex septentrionalis	15167	n:5991 C:94% (S:93.8% D:0.3%) F:1.9% M:4.1% hymenoptera_odb10	Standard
UP000078540	Atta colombica	14143	n:5991 C:91.3% (S:91.2% D:0.1%) F:2.5% M:6.2% hymenoptera_odb10	Standard
UP000075809	Trachymyrmex zeteki	14302	n:5991 C:89.5% (S:89.4% D:0.1%) F:2.5% M:8% hymenoptera_odb10	Standard
UP000000311	Camponotus floridanus (Florida carpenter ant) (Strain: C129)	14787	n:5991 C:89.8% (S:89.2% D:0.5%) F:4.9% M:5.4% hymenoptera_odb10	Standard
UP000078492	Trachymyrmex cornetzi	18657	n:5991 C:92.1% (S:91.8% D:0.2%) F:2.1% M:5.8% hymenoptera_odb10	Close to Standard
UP000008237	Harpegnathos saltator (Jerdon's jumping ant) (Strain: R22 G/1)	15029	n:5991 C:87.2% (S:86.8% D:0.4%) F:6% M:6.8% hymenoptera_odb10	Standard
UP000078542	Cyphomyrmex costatus	15716	n:5991 C:91.9% (S:91.3% D:0.6%) F:2.3% M:5.8% hymenoptera_odb10	Standard
UP000036403	Lasius niger (Black garden ant)	18044	n:5991 C:64.1% (S:63.5% D:0.6%) F:18.1% M:17.8% hymenoptera_odb10	Standard
UP000053097	Ooceraea biroi (Clonal raider ant) (Cerapachys biroi)	16497	n:5991 C:94.1% (S:93.6% D:0.6%) F:3.1% M:2.8% hymenoptera_odb10	Standard
UP000310200	Temnothorax longispinosus	13035	n:5991 C:83.4% (S:82.9% D:0.5%) F:3.4% M:13.2% hymenoptera_odb10	Standard
UP000504618	Temnothorax curvispinosus	21925	n:5991 C:94.7% (S:50.2% D:44.5%) F:2.4% M:3% hymenoptera_odb10	Outlier
UP000006539	Solenopsis invicta (Red imported fire ant) (Solenopsis wagneri)	13	n:5991 C:0% (S:0% D:0%) F:0% M:100% hymenoptera_odb10	Outlier
UP000504615	Pogonomyrmex barbatus (red harvester ant)	13753	n:5991 C:96.4% (S:62.9% D:33.5%) F:2.5% M:1.1% hymenoptera_odb10	Standard
UP000515204	Dinoponera quadriceps (South American ant)	17661	n:5991 C:98.6% (S:52.1% D:46.5%) F:0.8% M:0.6% hymenoptera_odb10	Close to Standard
UP000279307	Ooceraea biroi (Clonal raider ant) (Cerapachys biroi)	12855	n:5991 C:93.7% (S:92.4% D:1.2%) F:2.6% M:3.8% hymenoptera_odb10	Standard

FONTE: Bateman (2019)

Dois proteomas foram desconsiderados pois eram compostos em sua totalidade de genes ausentes (100% *missing genes*) (FIGURA 6.1). Esses proteomas pertenciam às espécies *Solenopsis invicta* e *Nylanderia fulva*. Como haviam dois proteomas de *Ooceraea biroi*, o de referência foi utilizado. Para a construção da filogenia, por fim, foram utilizados 16 proteomas (15 proteomas de formigas e 1 proteoma de abelha).

A comparação entre a filogenia gerada a partir dos proteomas totais e a filogenia gerada a partir da MAF (Filogenia 1) pode ser observada na FIGURA 6.2. Tirando o fato de duas espécies presentes na filogenia de proteomas não estarem na MAF (*Trachymyrmex zeteki* e *Trachymyrmex cornetzi*), as duas filogenias mostraram-se iguais: a subfamília mais basal sendo Ponerinae, seguida da subfamília Dorylinae e, mais apicalmente, Formicinae e Myrmicinae como grupos irmãos. A distribuição das espécies dentro das subfamílias também foi a mesma.

6.2.2 Filogenia 2 X Filogenia de Proteomas Mitocondriais

Na filogenia de proteomas mitocondriais haviam 11 espécies (*Pristomyrmex punctatus* (Smith, F., 1860), *Solenopsis geminata* (Fabricius, 1804), *Solenopsis richteri* (Forel, 1909), *Solenopsis invicta*, *Myrmica scabrinodis* (Nylander, 1846), *Vollenhovia emeryi* (Wheeler, W.M., 1906), *Camponotus atrox* (Emery, 1925), *Formica fusca* (Linnaeus, 1758), *Formica selysi* (Bondroit, 1918), *Linepithema humile* (Mayr, 1868) e *Leptomyrmex pallens* (Emery, 1883)). Todas as espécies possuíam correspondência na MAF e a comparação entre as duas filogenias se encontra na FIGURA 6.3. Na filogenia de proteoma mitocondrial, a subfamília Dolichoderinae constitui um grupo parafilético mais basal enquanto na filogenia da MAF (Filogenia 2) a subfamília encontra-se num ramo menos basal que a subfamília Formicinae. A subfamília Myrmicinae também constitui um grupo parafilético na filogenia de proteomas mitocondriais, com *Pristomyrmex puctatus* como grupo irmão de Formicinae e das demais espécies de Myrmicinae. Enquanto a Filogenia 2 conseguiu agrupar Myrmicinae em um ramo apical. A distribuição das espécies dentro das subfamílias também difere entre as duas filogenias.

6.2.3 Filogenia 3 X Filogenia de UCE

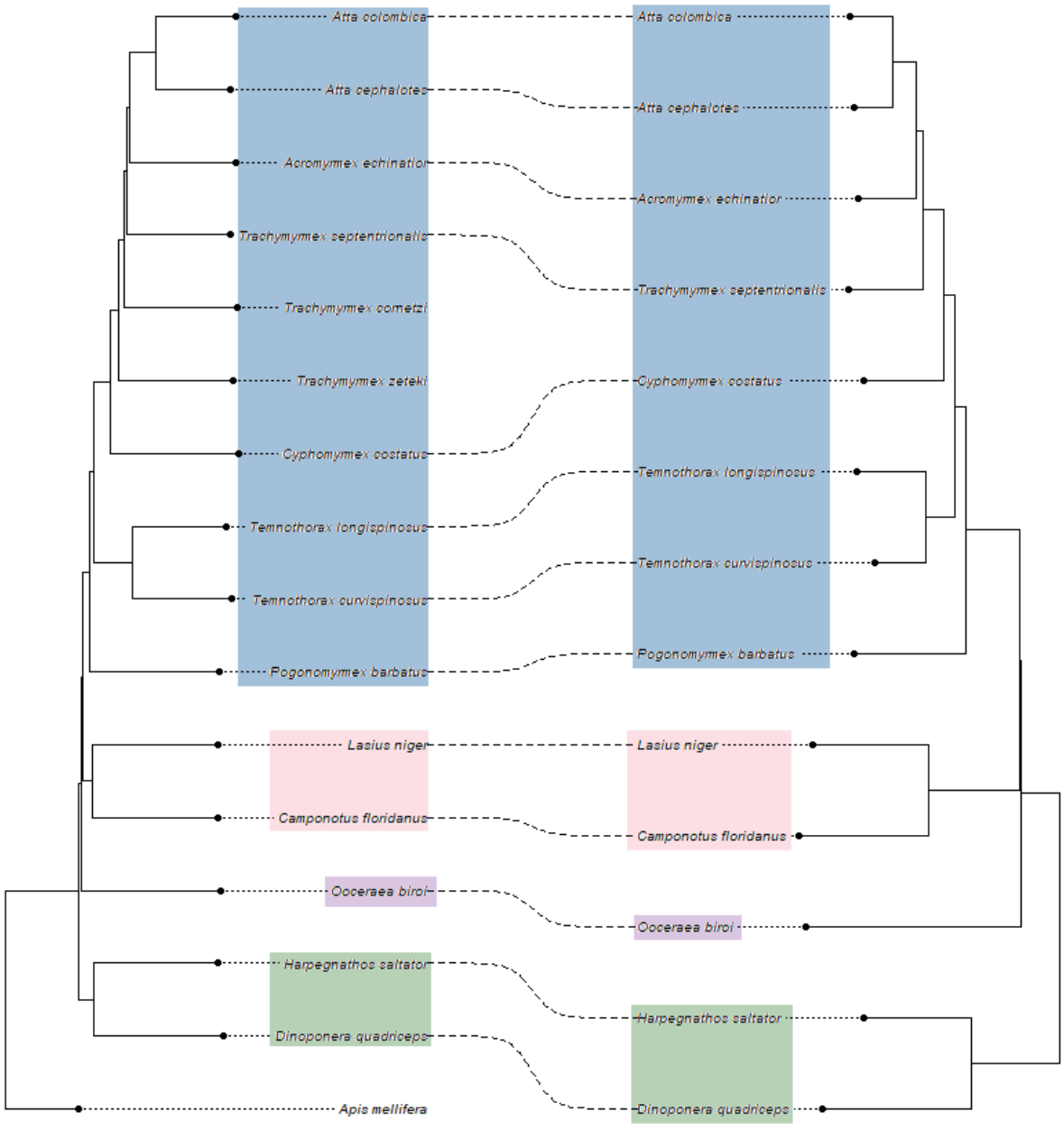
Na filogenia de UCE, havia 73 espécies com correspondência na MAF que podem ser visualizadas na FIGURA 6.4. De modo geral, as filogenias equivaleram-se, pois as subfamílias se agruparam de forma similar em grupos monofiléticos. Em ambas as filogenias há uma clara divisão entre os clados Formicoide e Poneroide que constituem grupos irmãos e são separados na base da filogenia. A única subfamília que encontra-se deslocada na filogenia feita a partir da MAF (Filogenia 3) é a Paraponerinae, representada pela espécie *Paraponera clavata* que, como discutido na seção 5.3, é representada por uma única espécie e gênero, o que empobreceu sua representação vetorial na MAF. As outras diferenças na filogenia acontecem a nível das relações entre os gêneros dentro da subfamília.

6.2.4 Filogenia 4 X Filogenia de *Pheidole*

Na filogenia de *Pheidole*, haviam 140 espécies com correspondência na MAF que podem ser visualizadas na FIGURA 6.5. Há várias divergências entre as duas filogenias. A mais evidente é com relação à base da árvore, já que a filogenia da literatura coloca *Pheidole fimbriata* (Roger, 1863) e *Pheidole rhea* (Wheeler, W.M., 1908) como as espécies mais basais do gênero e a filogenia gerada a partir da MAF (Filogenia 4) separa o gênero em dois grandes grupos, com *P. fimbriata* e *P. rhea* aparecendo como espécies irmãs, porém num ramo mais apical. Em relação às demais espécies, a topologia e relações entre as mesmas difere bastante.

Analisando as espécies de *Pheidole* na filogenia da MAF (com 2.981 espécies) (FIGURA 6.6), o gênero acaba aparecendo separado em três grupos. O grupo maior é um ramo que contempla 251 espécies do gênero. Os outros grupos menores são parafiléticos e contemplam 29 e 8 espécies, respectivamente.

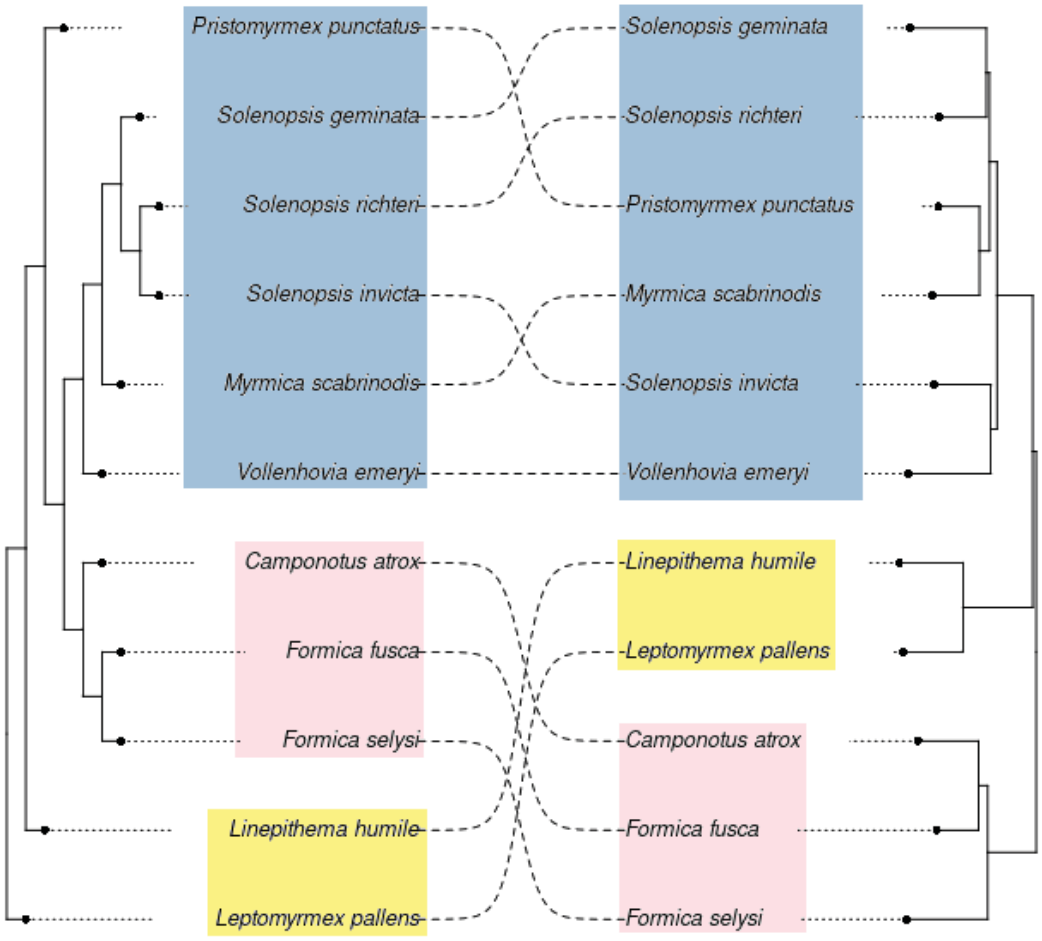
FIGURA 6.2: FILOGENIA DE PROTEOMA TOTAL X FILOGENIA 1



FONTE: a Autora (2021)

LEGENDA: à esquerda encontra-se a filogenia de proteomas totais e, à direita, a filogenia produzida neste trabalho através da MAF. As cores referem-se às subfamílias. Azul: Myrmicinae; Rosa: Formicinae; Lilás: Dorylinae; Verde: Ponerinae.

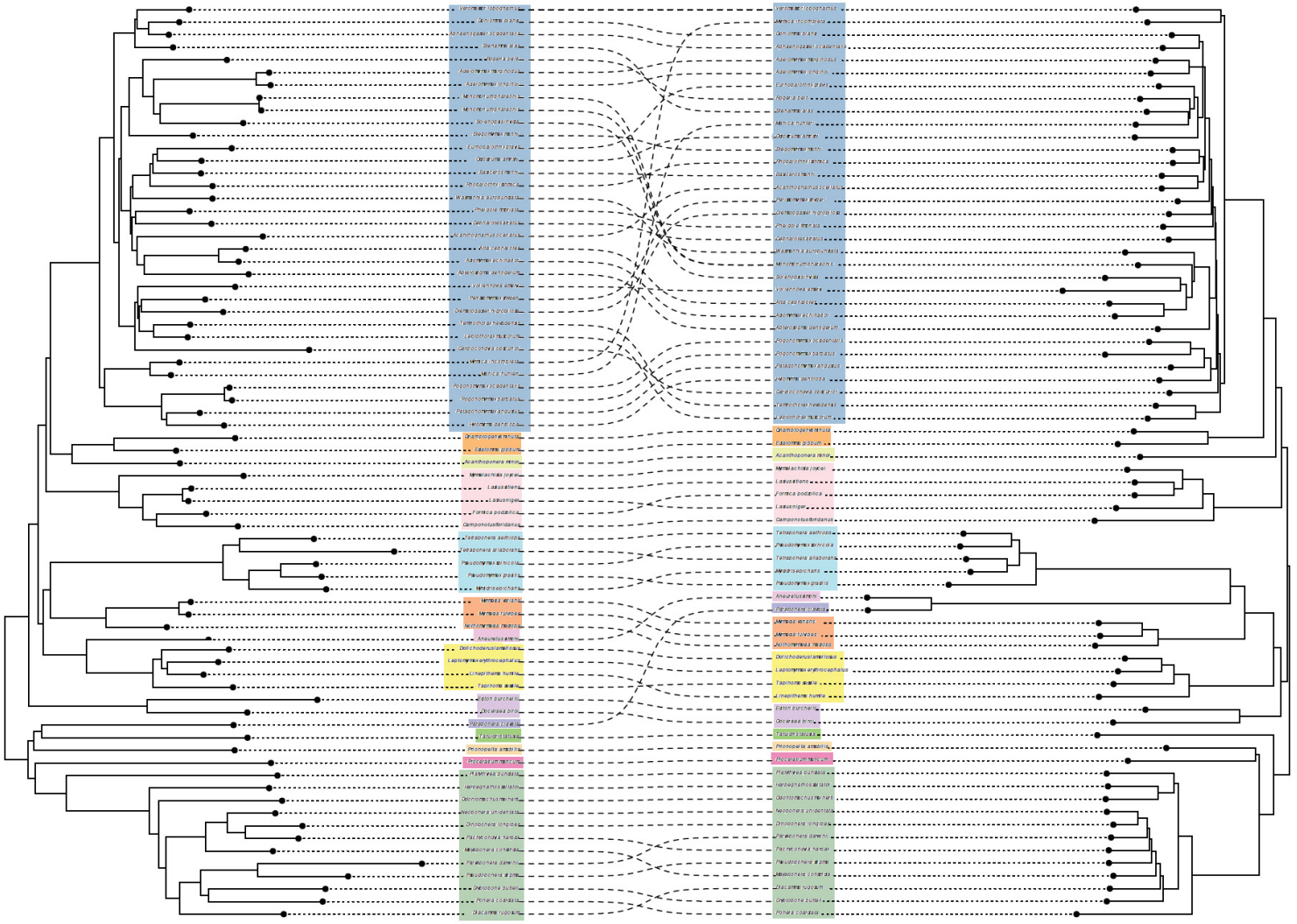
FIGURA 6.3: FILOGENIA DE PROTEOMA MITOCONDRIAL X FILOGENIA 2



FONTE: a Autora (2021)

LEGENDA: à esquerda encontra-se a filogenia de proteomas mitocondriais e, à direita, a filogenia produzida neste trabalho através da MAF. As cores referem-se às subfamílias. Azul: Myrmicinae; Rosa: Formicinae; Amarelo: Dolichoderinae.

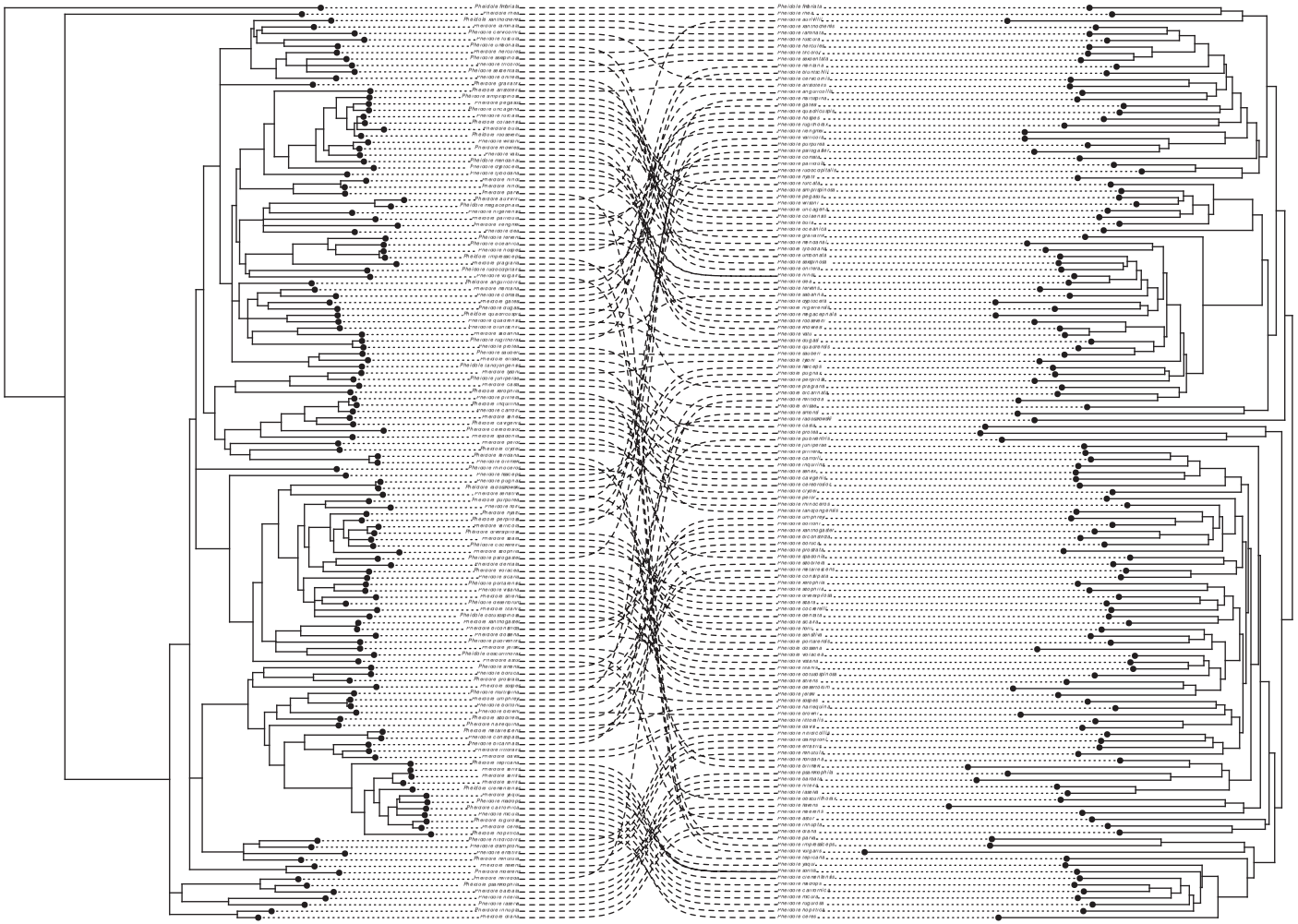
FIGURA 6.4: FILOGENIA DE UCE X FILOGENIA 3



FONTE: a Autora (2021)

LEGENDA: à esquerda encontra-se a filogenia de UCE e, à direita, a filogenia produzida neste trabalho através da MAF. Cada cor representa uma subfamília.

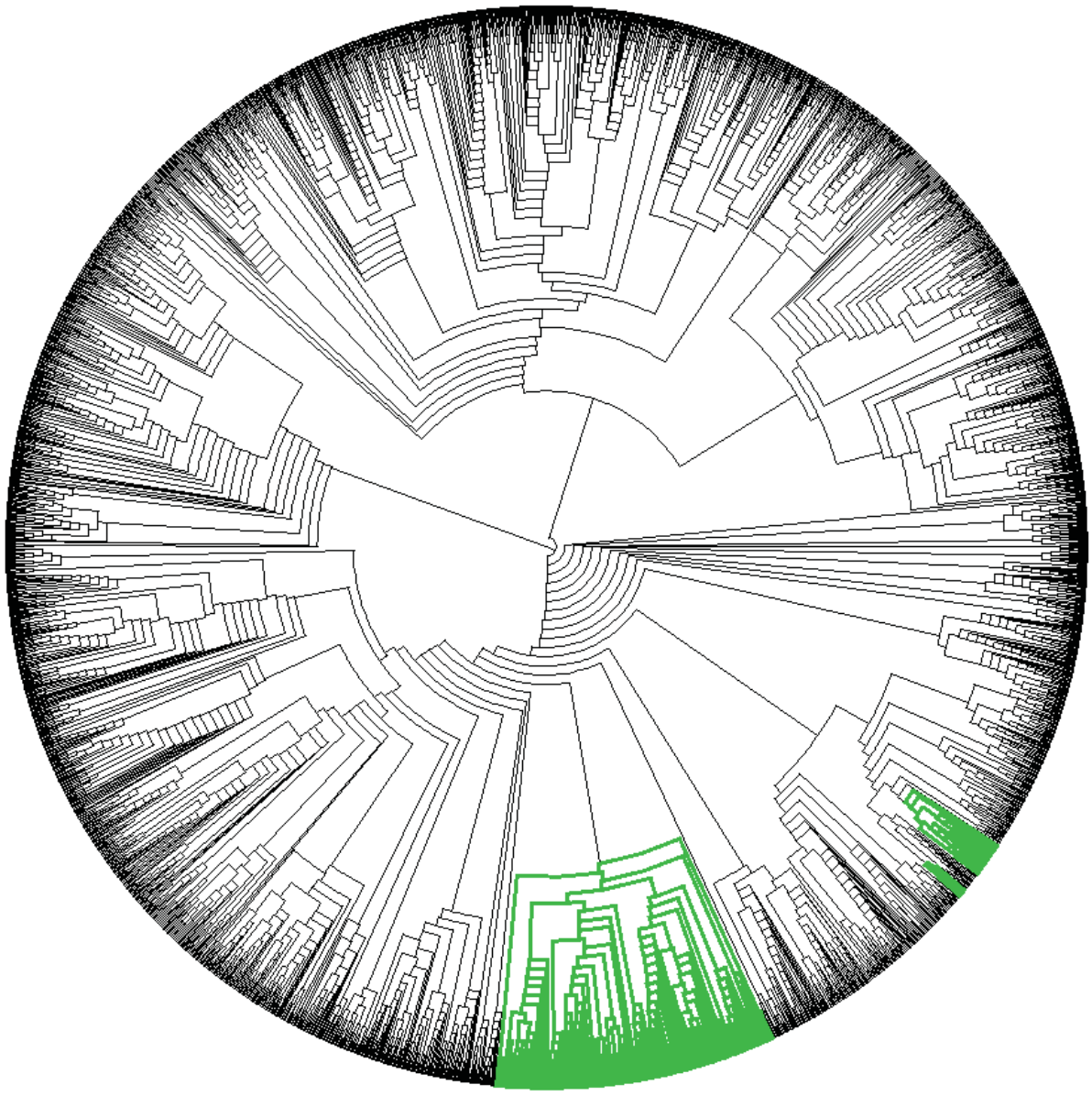
FIGURA 6.5: FILOGENIA DE *PHEIDOLE* X FILOGENIA 4



FONTE: a Autora (2021)

LEGENDA: à esquerda encontra-se a filogenia de *Pheidole* da literatura e, à direita, a filogenia produzida neste trabalho através da MAF.

FIGURA 6.6: *PHEIDOLE* NA FILOGENIA FINAL



FONTE: a Autora (2021)
LEGENDA: em verde, os ramos em que se encontram formigas do gênero *Pheidole* na filogenia feita através da MAF.

6.3 DISCUSSÃO

As filogenias construídas através dos vetores da MAF (Filogenias 1, 2, 3 e 4) mostraram-se coerentes com filogenias feitas por estudos anteriores e por técnicas de vetorização similares com dados mais robustos (BRANSTETTER *et al.*, 2017; ECONOMO *et al.*, 2015; DE PIERRE *et al.*, 2020). Nas duas primeiras comparações, com proteoma total e com proteoma mitocondrial, as filogenias utilizadas usaram o método SWeeP para a vetorização das sequências. Na primeira comparação, as espécies *Trachymyrmex zeteki* e *Trachymyrmex cornetzi* não possuíam

correspondência na MAF, pois, de acordo com AntWiki (2020), as duas espécies sofreram uma mudança de nomenclatura, passando a serem nomeadas *Paratrachymyrmex cornetzi* e *Mycetomoellerius zeteki*, respectivamente. A lista de espécies utilizada para a curadoria das sequências estava atualizada com as novas nomenclaturas, enquanto as sequências do NCBI estavam com a nomenclatura antiga e, consequentemente, não foram selecionadas para as etapas de vetorização. Apesar da falta dessas duas espécies, as filogenias mostraram-se idênticas e com a distribuição das subfamílias correspondente com a literatura. A filogenia de proteomas mitocondriais diferiu da literatura pela parafilia de Myrmicinae. Já a Filogenia 2 recuperou Myrmicinae como um grupo monofilético, porém colocou Formicinae como um grupo mais basal que Dolichoderinae, divergindo da literatura.

Ao comparar a acurácia das Filogenias 1 e 2, pode-se notar que a Filogenia 1 mostrou-se mais próxima dos resultados da literatura. Um motivo que explica esse fato é que todas as espécies contidas na Filogenia 1 possuíam o proteoma completo, grande parte sendo proteomas de referência, o que garantiu uma maior qualidade dos dados. Na Filogenia 1, 13 espécies possuíam sequências de todas as 6 proteínas usadas na construção da MAM e as outras 3 espécies possuíam sequências de pelo menos 3 dessas proteínas. Na Filogenia 2, apenas 3 espécies possuíam informação das 6 proteínas, ou seja, a informação das espécies da Filogenia 1 era mais rica que a das espécies da Filogenia 2. As maiores contradições com as filogenias apresentadas em outros estudos só aparecem a nível de espécie.

A filogenia 3 e 4 mostram que o modelo tem resolução para agrupar subfamílias e gêneros de forma satisfatória, porém a resolução diminui nos ramos mais apicais. As formigas do gênero *Pheidole* possuem uma das maiores riquezas de espécies e tem representantes nos mais diversos biomas (florestas tropicais, desertos, savanas e regiões frias) (ECONOMO *et al.*, 2015; ECONOMO *et al.*, 2019). Por isso, as formigas desse gênero possuem várias adaptações evolutivas de acordo com o ambiente em que vivem e seus gêneros costumam agrupar por regiões geográficas (ECONOMO *et al.*, 2015). Tendo em vista essa diversidade, é notável que a filogenia completa feita através da MAF (FIGURA 6.6) tenha agrupado mais de 250 espécies num grupo monofilético.

7 TESTES DE QUALIDADE DO MODELO

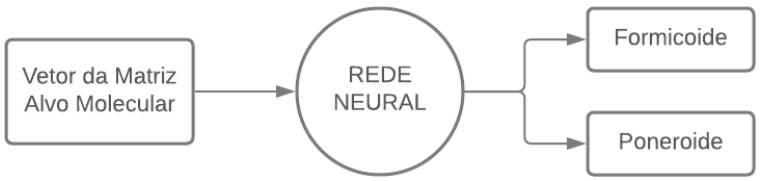
O objetivo do presente capítulo é avaliar a qualidade da informação contida no modelo de predição através do aprendizado de máquina, mais especificamente, redes neurais artificiais. Foi avaliado se a MAM gerada pelo modelo foi capaz de representar os padrões de dois diferentes complexos da família Formicidae (Poneroide e Formicoide), das subfamílias e de características fenotípicas.

7.1 MATERIAL E MÉTODOS

7.1.1 Complexos Poneroide e Formicoide

As espécies presentes na MAM foram classificadas em Poneroide ou Formicoide seguindo a classificação proposta no estudo de Branstetter *et al.* (2017). Os vetores MAM constituíram a entrada da rede neural e, a classificação, a saída (FIGURA 7.1). A rede foi estruturada em duas camadas de 13 e 7 neurônios, respectivamente, com função de treino LM e validação cruzada (5-*fold*). Os dados foram divididos aleatoriamente em um conjunto de treinamento com 70% dos dados e um conjunto de teste com 30% dos dados.

FIGURA 7.1: REDE DE CLASSIFICAÇÃO - PONEROIDE E FORMICOIDE



FONTE: a Autora (2020)

Para a avaliação da rede, foram gerados quatro parâmetros (considere VP como verdadeiros positivos, VN como verdadeiros negativos, FP como falsos positivos e FN falsos negativos):

- **Acurácia:** proporção entre classificações corretas e o número total de observações. Varia de 0 (nenhuma classificação correta) a 1 (todas as classificações corretas).

$$Acuracia = \frac{(VP + VN)}{(FP + FN + VN)}$$

- **Precisão:** proporção entre verdadeiros positivos e o total de classificados positivos.

$$Precisao = \frac{VP}{(VP + FP)}$$

- **Sensibilidade (*recall*):** proporção entre verdadeiros positivos e todas as observações realmente positivas.

$$Sensibilidade = \frac{VP}{(VP + FN)}$$

- **F1-Score:** média ponderada entre a precisão e a sensibilidade.

$$F1 - Score = \frac{2 * (sensibilidade * precisao)}{(sensibilidade * precisao)}$$

7.1.2 Subfamílias

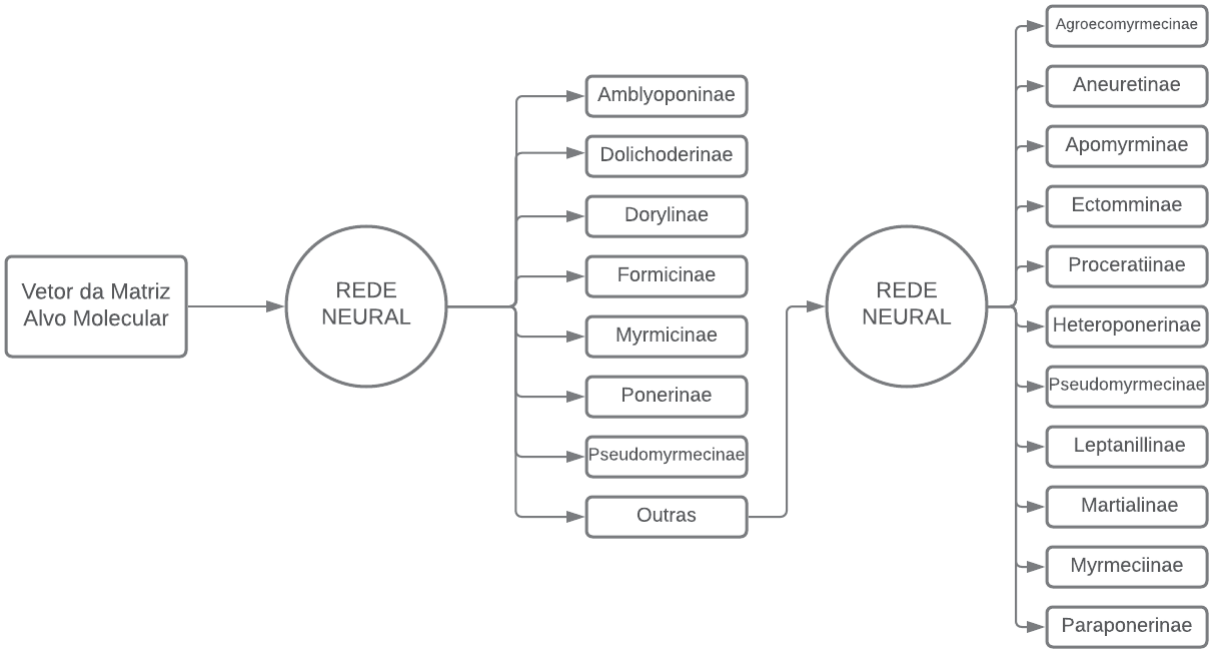
A partir da classificação de subfamília do AntWiki (2020), as espécies da MAM foram classificadas em uma das 17 subfamílias. Por conta da heterogeneidade da quantidade de dados em cada subfamília, a matriz foi dividida em dois grupos e para cada grupo treinou-se uma rede de classificação em que a entrada da rede era a MAM e a saída a subfamília correspondente (FIGURA 7.2). O primeiro grupo englobou as sete subfamílias com maior quantidade de sequências na MAM (Amblyoponinae, Dolichoderinae, Dorylinae, Formicinae, Myrmicinae, Ponerinae e Pseudomyrmecinae) e as subfamílias restantes foram colocadas numa oitava categoria de subfamília denominada "Outras". O segundo grupo continha todas as subfamílias que no primeiro grupo estavam classificadas como "Outras", e que possuíam poucas espécies na MAM, sendo elas: Agroecomyrmecinae, Aneuretinae, Apomyrminae, Ectatomminae, Heteroponerinae, Leptanillinae, Martialinae, Myrmeciinae, Paraponerinae e Proceratiinae.

Para o treinamento das duas redes, a divisão do conjunto de dados em conjunto de treinamento e teste, a estrutura da rede e os parâmetros de avaliação foram os mesmos usados na rede para a classificação dos grupos Poneroide e Formicoide (seção 7.1.1). Por fim, para a predição dos dados, as redes foram usadas em dois passos:

1. Classificação dos dados utilizando a rede de classificação das subfamílias maiores.

2. Para os vetores que foram classificados como "Outras", classificação dos dados utilizando a rede de subfamílias menores.

FIGURA 7.2: REDE DE CLASSIFICAÇÃO - SUBFAMÍLIAS



FONTE: a Autora (2020)

7.1.3 Fenótipo

Os dados fenotípicos foram retirados de duas fontes distintas: do banco de dados GLAD (PARR *et al.*, 2017) e do *website* AntWiki (2020). Abaixo segue a metodologia utilizada para a coleta em cada um desses bancos de dados:

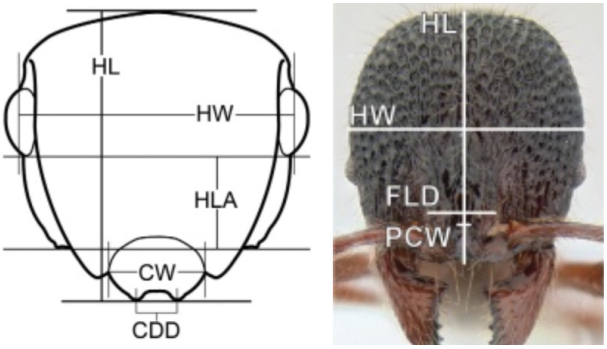
- **GLAD[1]:** *Global Ants Database* (GLAD)[2] é uma colaboração mundial entre ecólogos para unir dados acerca da abundância e características de comunidades locais de formigas. Os dados ficam disponíveis para *download* mediante o cadastro e envio de projeto descrevendo como os dados serão utilizados. Os dados fenotípicos são disponibilizados no formato de arquivo de texto CSV (*comma-separated-values*).

- **Antwiki[3]:** banco de dados *online*, público e de fácil acesso que possui uma página *web* para cada espécie de formiga descrita. Os dados são alimentados através da integração com outros bancos de dados e das contribuições de especialista e amadores. Não há como fazer o *download* de todos os dados e, por este motivo, foi utilizado um robô desenvolvido em MATLAB®[4] que realizou o *download* automático das características fenotípicas disponíveis no texto da página de cada espécie.

[1]Dados de 05/10/2020
[2]Disponível no endereço <http://globalants.org/>.
[3]Dados de 28/12/2020
[4]Disponível no endereço <https://github.com/giupasqualato/Antwiki-Extractor/>

A característica escolhida para o treinamento das redes neurais foi o comprimento da cabeça (HL - *head length*) das operárias menores, variável numérica contínua apresentada em milímetros, pois era a variável com menor quantidade de registros nulos. Essa medida é o comprimento mensurado por uma linha média vertical da cápsula da cabeça da formiga em "vista de rosto inteiro" (frontal), excluindo as mandíbulas (FIGURA 7.3). Caso se apresente alguma concavidade tanto na parte posterior da cápsula quanto no clípeo, é traçado um segmento de reta que toca os ápices da região posterior e/ou anterior e, a partir da metade desse segmento, será traçada a linha vertical.

FIGURA 7.3: MEDIDA DE COMPRIMENTO DA CABEÇA

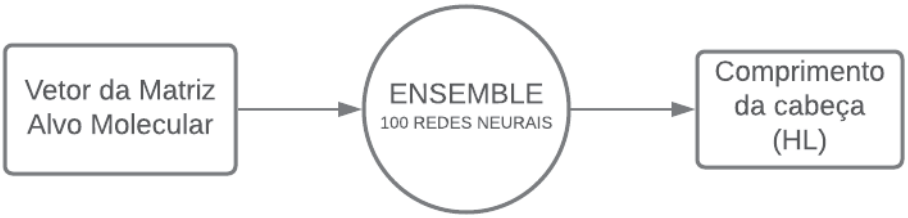


FONTE: Adaptado de *Morphological Measurements* de AntWiki (2020)
LEGENDA: A medida comprimento da cabeça corresponde ao HL.

O GLAD possui um total de 4.154 registros, porém nem todos identificados a nível de espécie e com valor de HL preenchido. Para a montagem da tabela a ser usada para o treinamento da rede, o primeiro passo foi excluir os registros que não possuíam informação de HL. Após isso, espécies repetidas foram concatenadas através do cálculo da média dos valores de HL. A seguir, os registros restantes do GLAD foram unidos aos valores de HL do Antiwiki minerados pelo robô. Por último, foram selecionadas somente as espécies que estavam contidas na MAM, resultando numa tabela final com valores de HL para 1332 espécies (294 do GLAD + 1038 do Antiwiki).

Para o treinamento das redes neurais, foi realizado um *ensemble* de 100 redes. A entrada de cada rede individual foi a MAM e, a saída, o valor de HL (FIGURA 7.4). Cada rede foi composta de 3 camadas ocultas, cada uma com 15 neurônios e função de treino LM. O conjunto de treinamento constituiu-se de 80% dos dados e, o conjunto de teste, 20 %. Os registros foram distribuídos aleatoriamente entre os grupos de treino e teste. Para a validação da rede, foi utilizada a correlação de Spearman entre os dados inferidos e os dados reais, além da média de MSE das redes que compõe o *ensemble*.

FIGURA 7.4: REDE DE CLASSIFICAÇÃO - FENÓTIPO



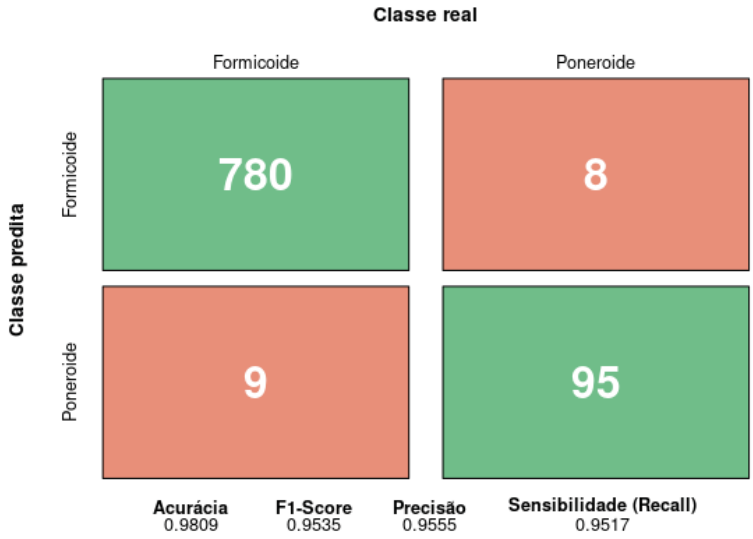
FONTE: a Autora (2020)

7.2 RESULTADOS

7.2.1 Complexos Poneroide e Formicoide

A rede obteve acurácia de 0,9809, F1-Score de 0,9535, precisão de 0,9555 e sensibilidade de 0,9517 na classificação dos grupos Formicoide e Poneroide nos dados de teste (FIGURA 7.5).

FIGURA 7.5: MATRIZ DE CONFUSÃO - PONEROIDE E FORMICOIDE



FONTE: a Autora (2021)

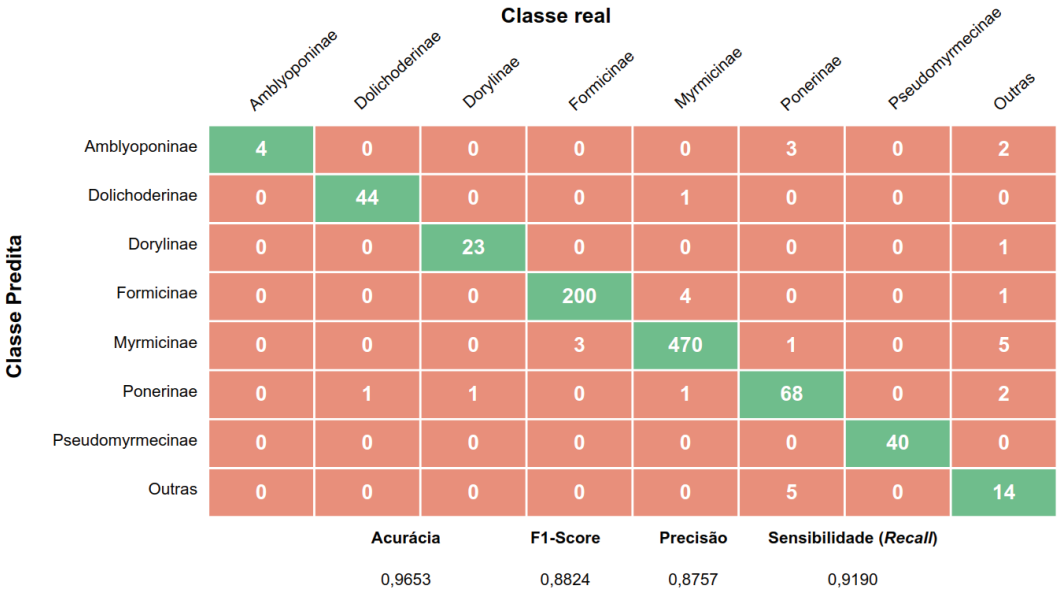
LEGENDA: a matriz mostra os acertos e erros cometidos pela rede. Os quadrados em verde mostram a quantidade de predições corretas (a classe real e a predita são as mesmas) e os quadrados em vermelho mostram a quantidade de erros (a classe real e a predita diferem).

7.2.2 Subfamílias

A primeira rede obteve uma acurácia de 0,9653, F1-Score de 0,8824, precisão de 0,8757 e sensibilidade de 0,9190 na classificação das subfamílias grandes (FIGURA 7.6). Já a segunda rede, para classificação das subfamílias pequenas, a acurácia foi de 0,6957 e precisão 0,4688 (FIGURA 7.7). Como algumas subfamílias possuíam apenas um representante, não houve dados

suficiente dessas subfamílias nos grupos de treinamento e teste e o F1-Score e a sensibilidade não puderam ser calculados.

FIGURA 7.6: MATRIZ DE CONFUSÃO - SUBFAMÍLIAS GRANDES



FONTE: a Autora (2021)

LEGENDA: a matriz mostra os acertos e erros cometidos pela rede. Os quadrados em verde mostram a quantidade de predições corretas (a classe real e a predita são as mesmas) e os quadrados em vermelho mostram a quantidade de erros (a classe real e a predita diferem).

FIGURA 7.7: MATRIZ DE CONFUSÃO - SUBFAMÍLIAS PEQUENAS

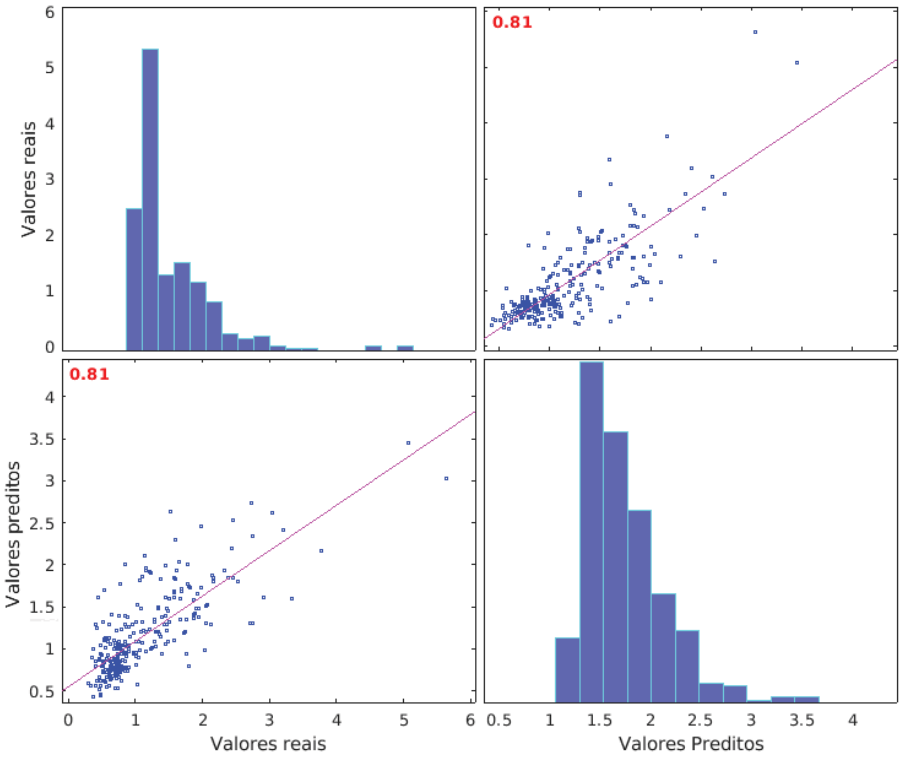
FONTE: a Autora (2021)

LEGENDA: a matriz mostra os acertos e erros cometidos pela rede. Os quadrados em verde mostram a quantidade de predições corretas (a classe real e a predita são as mesmas) e os quadrados em vermelho mostram a quantidade de erros (a classe real e a predita diferem).

7.2.3 Fenótipo

A correlação entre os valores de comprimento de cabeça preditos pelo *ensemble* e os valores reais foi alta, positiva e significativa ($\rho = 0,8104$, p-value = $2,6568.10^{-76}$) (FIGURA 7.8). Além disso, o MSE médio das redes individuais foi de 0,0370.

FIGURA 7.8: CORRELAÇÃO DOS FENÓTIPOS REAIS E PREDITOS



FONTE: a Autora (2021)

LEGENDA: os gráficos superior esquerdo e inferior direito representam a distribuição dos valores reais e preditos, respectivamente. Os gráficos inferior esquerdo e superior direito mostram o diagrama de dispersão entre os valores reais e preditos. O valor em vermelho representa o coeficiente de correlação.

7.3 DISCUSSÃO

A rede usada para a classificação dos clados Poneroide e Formicoide obteve performance excelente (FIGURA 7.5). Tal fato mostra que a MAM gerada neste trabalho conseguiu sumarizar os padrões necessários para representar as propriedades dessas duas classes. Na literatura, como já discutido, a divisão entre os dois grupos é clara (BRANSTETTER *et al.*, 2017; FERNÁNDEZ; LATTKE, 2015).

Para as subfamílias, a acurácia das classificações se relacionou com a quantidade de dados disponíveis em cada subfamília. A primeira rede classificou as subfamílias com mais representantes na MAM e colocou as subfamílias com menos representantes numa classe única, chamada "Outras". Como havia quantidade suficiente de dados para a rede entender os padrões,

a acurácia dessa primeira rede foi de 0,9653 (FIGURA 7.6). Para as subfamílias com poucos dados, a acurácia diminuiu para 0,6957 pois não haviam dados suficientes para a rede entender os padrões de forma satisfatória (FIGURA 7.7). Como já explicado na seção 5.3, algumas subfamílias possuem apenas um representante, o que não é suficiente para o treinamento de uma rede. Apesar disso, como discutido na seção 1.2, a maioria das espécies se encontra nas subfamílias abrangidas pela primeira rede, então a diminuição da acurácia na segunda rede não apresenta um problema.

A rede para a predição do fenótipo conseguiu acurácia alta e erro baixo (FIGURA 7.4). Em formigas, o fenótipo além de estar associado à herança genética dos organismos ancestrais, também está muito associado com o ambiente que essas formigas habitam (WEISER; KASPARI, 2006; ECONOMO *et al.*, 2015). Ou seja, alguns grupos sofrem convergência evolutiva e apresentam fenótipos similares apesar de não estarem tão próximos filogeneticamente. Mesmo com essas variações, a informação presente na MAM foi suficiente para o *ensemble* de redes conseguir aprender os padrões e predizer os fenótipos das espécies com acurácia alta.

É importante ressaltar que, novamente, problemas na qualidade dos dados acabam diminuindo a qualidade de predição e classificação das redes. Além da qualidade dos dados moleculares que já foram discutidos na seção 5.3, os bancos de dados de fenótipos para formigas são de qualidade contestável. No GLAD há muita informação incompleta e pouquíssimas espécies não apresentam valores nulos. No AntWiki, o grande empecilho é a falta de padronização dos dados e erros de digitação que dificultam a extração automática de informações. Há também outras bases de dados, como o <AntBase.org> que não estão atualizadas ou que as buscas não funcionam, talvez por falta de manutenção no *website*.

Tendo em vista as dificuldades trazidas pelos dados, os testes feitos nesse capítulo mostram a qualidade da informação contida na MAM. O modelo é capaz de compensar a incompletude dos dados e reconhecer padrões importantes acerca das espécies.

8 EXPANSÃO DO MODELO

O presente capítulo tem dois objetos principais: analisar como o modelo se comporta utilizando dados moleculares novos e não identificados e integrar à filogenia as espécies que não possuem nenhuma informação molecular.

8.1 MATERIAL E MÉTODOS

8.1.1 Uso do modelo com dados novos e não identificados

Para testar o comportamento do modelo com dados novos e com subfamílias desconhecidas, foi feito o download de novas sequências no banco de dados de proteínas do NCBI (SCHOCH *et al.*, 2020). Como já mencionado no capítulo 5, as sequências baixadas para a obtenção do modelo correspondem as todas as sequências de proteínas para Formicidae cadastradas no NCBI até o dia 26 de junho de 2020. Para o presente teste, foram baixadas as sequências cadastradas a partir do dia 10/10/2020 até o dia 26/01/2021, garantindo que nenhuma sequência usada nesse teste tenha sido usada na obtenção do modelo. Com as sequências novas em mãos, todo o processo descrito no capítulo 5 foi repetido, até chegar na geração da MAM.

A classificação real das subfamílias dos vetores não foi usada. No lugar, usou-se as redes descritas na seção 7.1.2 para inferir essas subfamílias. Com as subfamílias inferidas, deu-se continuidade no processo e a MAF foi gerada a partir das médias já calculadas na etapa de obtenção do modelo (capítulo 5). A árvore filogenética foi gerada com a MAF, como descrito na seção 5.1.6. Por fim, a nova filogenia foi comparada com a literatura.

8.1.2 Inserção das Espécies sem Dados Moleculares

Para alcançar uma **Filogenia Global** que inclua a totalidade de espécies da família Formicidae, é necessário incluir as espécies que não possuem nenhum dado molecular. Como mostrado no capítulo 5, somente 21,58% das espécies possuem dados moleculares de proteínas cadastrados no NCBI. Para incluir as demais 10.831 espécies, foi acrescentada a informação taxonômica de forma semelhante à geração da MAF. Cada vetor da matriz com todas as espécies é formado por duas partes:

1. Vetor de tamanho 200 que corresponde às médias normalizadas das subfamílias geradas na obtenção do modelo (seção 5.1.6)

2. Vetor de tamanho 200 que varia de acordo com as informações que a espécie possui, podendo ser:

    - *Para espécies que se encontram na MAM*: vetor correspondente na MAM.

- *Para espécie que não se encontram na MAM, porém outras espécies do gênero se encontram*: média normalizada do gênero.
- *Para espécies de gêneros que não se encontram na MAM:* média normalizada das subfamílias.

Desse modo, conseguiu-se uma matriz de tamanho 13812X400, que contemplou todas as espécies descritas, inclusive as sem dado molecular algum. Para a elaboração da filogenia seguiu-se os passos de cálculo de distância e construção da árvore descritos na seção 5.1.6.

8.2 RESULTADOS

8.2.1 Inserção de dados novos

Foram baixadas um total de 11.888 novas sequências. Desse total, das proteínas utilizadas na obtenção do modelo, quatro estavam presentes nessas novas sequências, sendo elas: citocromo C oxidase subunidade 1, rodopsina (comprimento de onda longo), fator de elongação 1-alpha e *Wingless*. Na tabela abaixo (TABELA 8.1), encontram-se maiores informações acerca das espécies e gêneros, além de informações sobre a fragmentação dessas proteínas.

Quatro matrizes alvos foram inferidas:

1. Matriz alvo de citocromo C oxidase subunidade 1 abrangendo 181 espécies;
2. Matriz alvo de rodopsina (comprimento de onda longo) abrangendo 15 espécies;
3. Matriz alvo de *wingless* abrangendo 15 espécies;
4. Matriz alvo de fator de elongação 1-alpha abrangendo 15 espécies;

TABELA 8.1: Informações acerca das sequências novas

Proteína	Gêneros abrangidos	Espécies abrangidas	Total de sequências	Total de seq. parciais	Total de seq. completas
Citocromo C oxidase subunidade 1	**54**	**181**	**4624**	**4584**	**40 (0,865%)**
Rodopsina (comprimento de onda longo)	2	15	720	720	0
Wingless	2	15	352	352	0
Fator de elongação 1-alpha	2	15	704	704	0

FONTE: a Autora (2021)
NOTA: Na coluna "Total de seq. completas", entre parênteses se encontram as porcentagens de sequências completas em relação ao número total de sequências. Em negrito, destacam-se os maiores valores da coluna.

A sumarização das matrizes alvo gerou uma MAM contendo 195 espécies. Após a classificação das subfamílias dos vetores da MAM, pode-se gerar a MAF e, por fim, a filogenia

(FIGURA 8.1). A filogenia obteve sucesso na separação dos clados Formicoide e Poneroide. A organização das subfamílias corresponde à literatura (BRANSTETTER *et al.*, 2017) e espécies do mesmo gênero ficaram bem agrupadas. A única exceção foi a espécie *Paratrachymyrmex irmgardae* que acabou misturando-se à subfamília Formicinae.

8.2.2 Inserção das Espécies sem Dados Moleculares

A filogenia global englobou todas as 13.812 espécies da família Formicidae. Apesar de mostrar os clados Poneroide e Formicoide como parafiléticos, a árvore conseguiu separar bem esses dois grupos (FIGURA 8.2). Com exceção da subfamília Myrmicinae, todas as subfamílias se encontram em grupos monofiléticos (FIGURA 8.3).

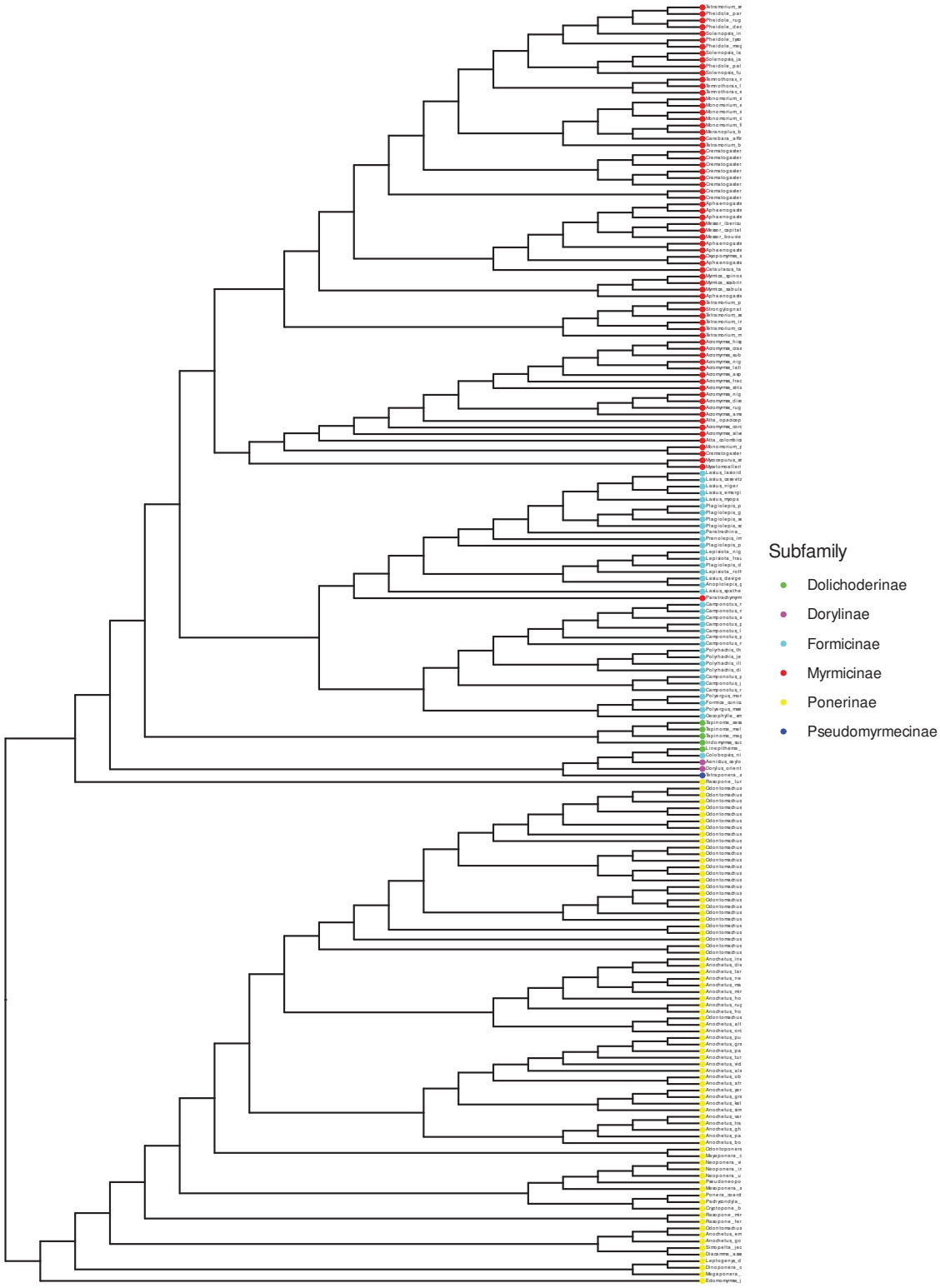
8.3 DISCUSSÃO

A inserção de novos dados mostrou que o modelo pode ser extrapolado para dados que não participaram das etapas de treinamento (capítulo 5), ou seja, o modelo obteve sucesso em extrair os padrões necessários para a reconstrução da relações filogenéticas da família Formicidae, mesmo com dados incompletos como mostra a TABELA 8.1. Além disso, comprovou-se que a informação taxonômica pode ser usada mesmo em dados em que a mesma é desconhecida: a rede de classificação de subfamílias usada para enriquecer a MAM conseguiu adicionar informação confiável para a geração da MAF, o que refletiu numa filogenia coesa (FIGURA 8.1).

Em relação à filogenia global (com todas as espécies pertencentes à subfamília Formicidae), com exceção da subfamília Agroecomyrmecinae, todas as subfamílias do clado Poneroide e a subfamília Leptanilinae se agruparam em um único ramo, mostrando que mesmo com a falta de dados, o modelo ainda consegue capturar os padrões necessários para a caracterização do clado Poneroide, como foi discutido na seção 7.3. Porém, diferente da literatura, o clado Poneroide não aparece como um clado basal e irmão ao clado Formicoide. As subfamílias mais basais Leptanilinae e Martialine que deveriam ser um grupo irmão à todas as outras formigas aparecem em ramos mais apicais. Ainda assim, a subfamília Leptanilinae aparece em um ramo basal dentro do clado Poneroide. A subfamília Martialinae está dentro do clado Formicoide e as razões para esse deslocamento já foram discutidas na seção 5.3.

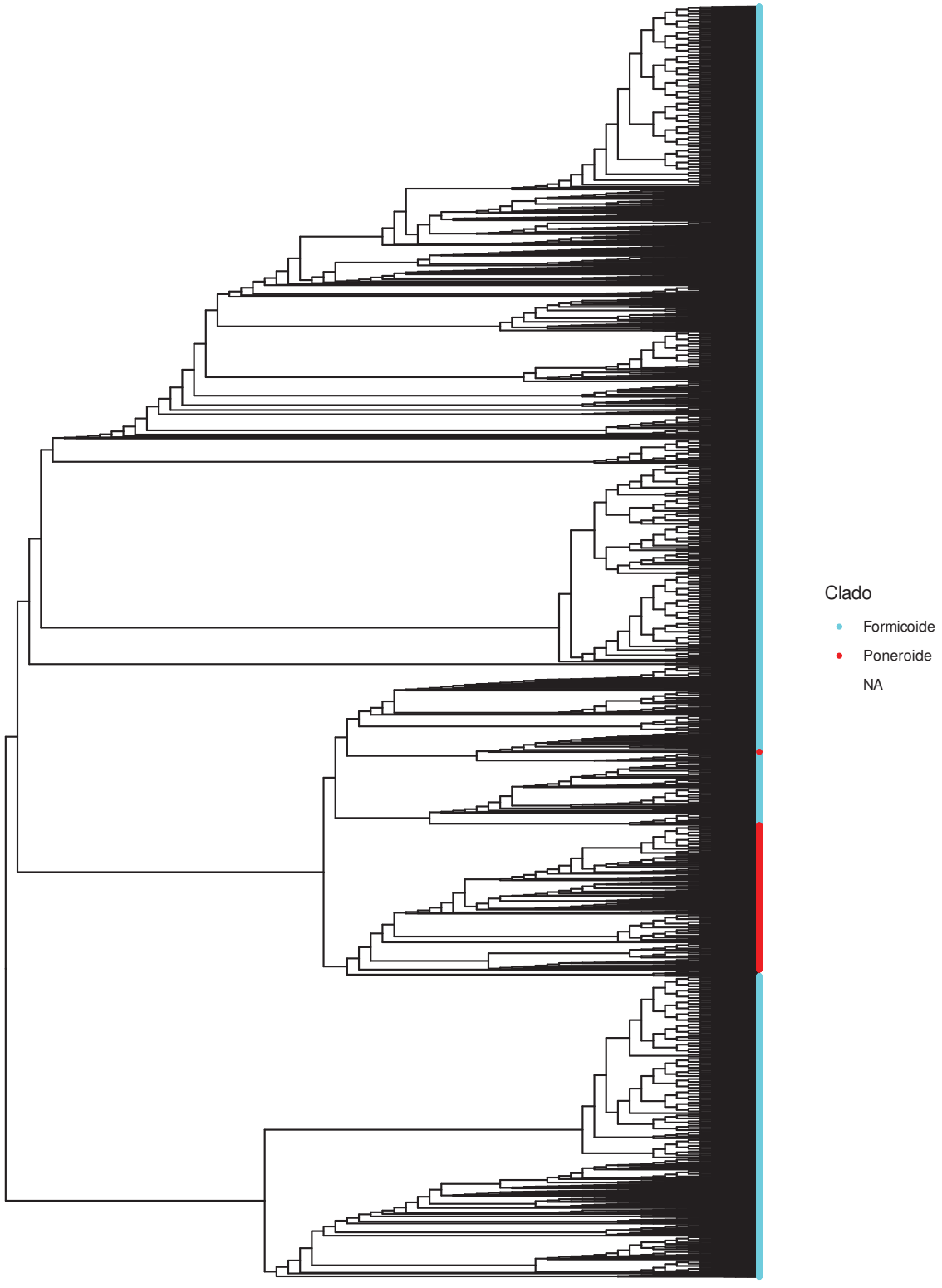
Com esses resultados fica claro que o modelo consegue lidar com dados de má qualidade e sem identificação taxonômica adequada. Além disso, o modelo conseguiu integrar espécies sem nenhum dado molecular à filogenia, gerando uma filogenia que contempla todas as espécies viventes. Porém, é evidente que a qualidade da informação utilizada afeta diretamente o resultado alcançado pelo modelo e a organização da filogenia acaba sendo prejudicada com a falta de dados.

FIGURA 8.1: FILOGENIA COM DADOS NOVOS



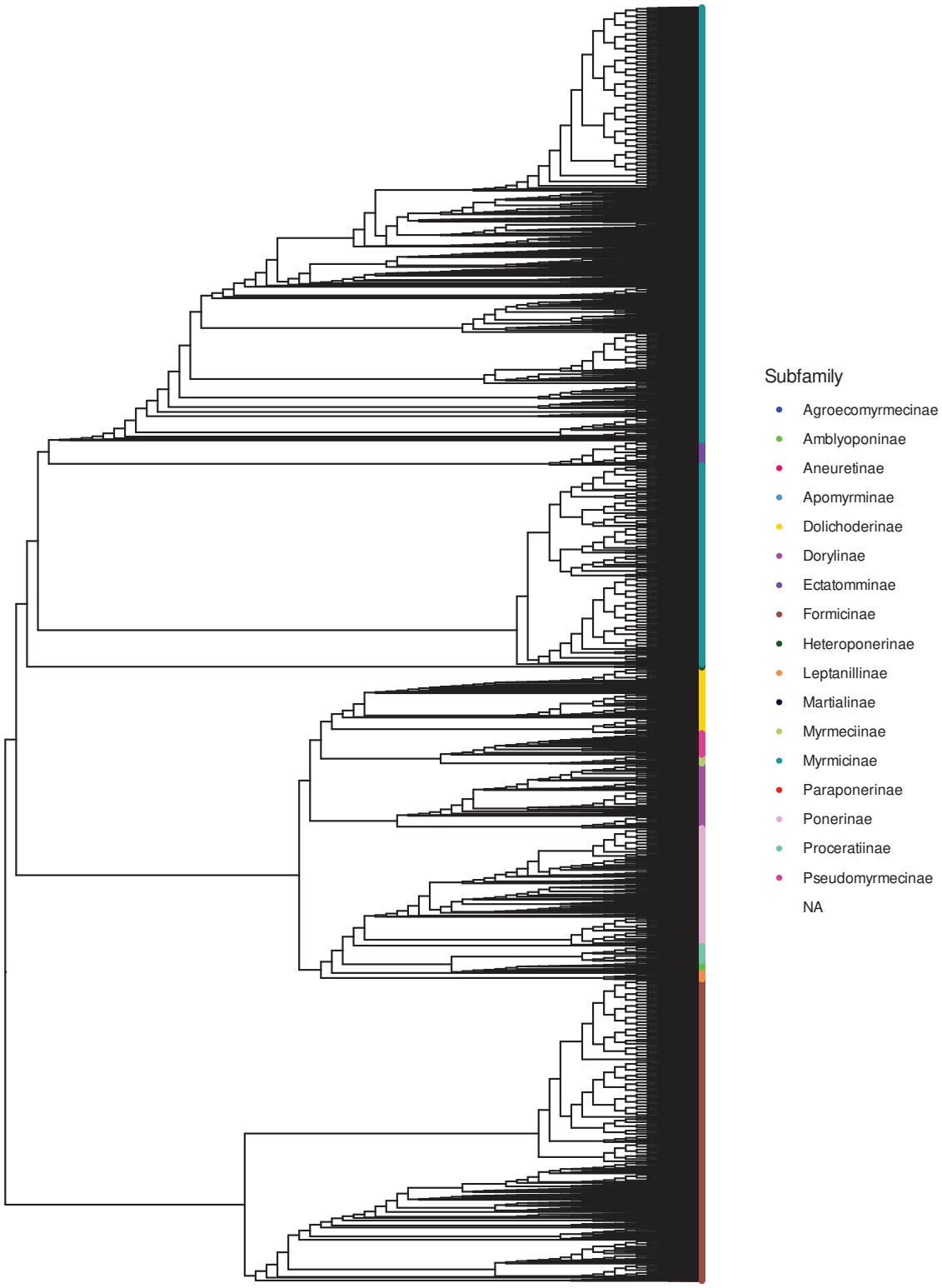
FONTE: a Autora (2021)

FIGURA 8.2: FILOGENIA COM 13812 ESPÉCIES - PONEROIDE E FORMICOIDE



FONTE: a Autora (2021)

FIGURA 8.3: FILOGENIA COM 13812 ESPÉCIES - SUBFAMÍLIAS



FONTE: a Autora (2021)

9 CONCLUSÃO

O modelo desenvolvido nesse trabalho conseguiu alcançar os objetivos propostos. O método SweeP possibilitou a vetorização e redução da dimensionalidade das sequências sem a perda da informação e comparabilidade entre elas. Através do aprendizado de máquina foi possível criar um modelo que conseguiu unir informação incompleta de diferentes proteínas em uma matriz (MAM) que contemplava toda a informação. Ou seja, o modelo foi capaz de completar lacunas de informação que inviabilizariam a construção da filogenia.

Não foi possível recuperar toda a informação filogenética através somente da MAM, mas esse problema foi contornado concatenando um vetor com informação taxonômica já existente, gerando a MAF. Esse vetor criou um viés para que organismos da mesma subfamílias se agrupassem, o que diminuiu erros e permitiu a recuperação das relações filogenéticas corretas. Para o caso da subfamília ser desconhecida, criou-se uma rede que fornece a classificação da mesma. Além disso, criou-se uma metodologia para inserir espécies sem dado molecular na filogenia global, porém notou-se que a falta de dados moleculares acabou distorcendo a informação filogenética.

A filogenia com melhor resolução alcançada foi a feita a partir da MAF, ou seja, a matriz com informações moleculares e taxonômicas concatenadas, englobando 2.981 espécies de formigas de 281 gêneros e das 17 subfamílias (21,58% do total de formigas). Também foi alcançada a filogenia global, com todas as 13.812 espécies descritas, 338 gêneros e 17 subfamílias

Os testes realizados mostraram que, apesar de dados incompletos, desbalanceados e heterogêneos, a MAM conseguiu representar os padrões taxonômicos e fenotípicos. Apesar de termos alcançado resultados ótimos mesmo com essas condições dos dados, ficou clara a relação direta entre a qualidade do dado e a qualidade dos resultados do modelo. Esse trabalho é um exemplo de como banco de dados com má manutenção, má curadoria de dados e informações incompletas acabam por prejudicar e dificultar estudos com contribuições importantes.

Por fim, este trabalho mostra o potencial da integração do método SWeeP à técnicas de aprendizado de máquina na resolução de problemas biológicos complexos. As matrizes geradas pelo SWeeP, ao servirem de entrada para treinamento do modelo, não só se demonstraram eficientes para a construção de filogenias de táxons extensos sem a necessidade de alinhamento, mas também proporcionou informação o bastante para que fosse possível estimar fenótipo a partir registros moleculares de qualidade média. Como perspectivas futuras, pretende-se testar o modelo em outros grupos taxonômicos.

REFERÊNCIAS

AGGARWAL, C. C. **Data Mining: The Textbook**. New York, NY: Springer, 2015. 734 p.

AMORIM, D. d. S. **Fundamentos de Sistemática Filogenética**. 3. ed. Ribeirão Preto: Holos Editora, 2002. 156 p.

ANGOTTI, M. A. *et al.* Seed removal by ants in Brazilian savanna: optimizing fieldwork. **Sociobiology**, v. 65, n. 2, p. 155–161, 2018. ISSN 03616525.

ANTWIKI. **AntWiki: Reports**. 2020. Disponível em: <https://antwiki.org/wiki/Help:AntWiki_Reports>. Acesso em: 01/05/2020.

BATEMAN, A. UniProt: A worldwide hub of protein knowledge. **Nucleic Acids Research**, v. 47, n. D1, p. D506–D515, 2019. ISSN 13624962.

BAUER, E.; KOHAVI, R. Empirical comparison of voting classification algorithms: bagging, boosting, and variants. **Machine Learning**, v. 36, n. 1, p. 105–139, 1999. ISSN 08856125.

BOROWIEC, M. L. *et al.* Compositional heterogeneity and outgroup choice influence the internal phylogeny of the ants. **Molecular Phylogenetics and Evolution**, Elsevier, v. 134, n. August 2017, p. 111–121, 2019. ISSN 10959513.

BRANSTETTER, M. G. *et al.* Enriching the ant tree of life: enhanced UCE bait set for genome-scale phylogenetics of ants and other Hymenoptera. **Methods in Ecology and Evolution**, v. 8, n. 6, p. 768–776, 2017. ISSN 2041210X.

BURGIO, K. R. *et al.* Phylogenetic supertree and functional trait database for all extant parrots. **Data in Brief**, v. 24, 2019. ISSN 23523409.

CHAKRABORTY, I.; CHOUDHURY, A. Artificial Intelligence in Biological Data. **Journal of Information Technology & Software Engineering**, v. 07, n. 04, 2017.

CHAN, K. H.; GUÉNARD, B. Ecological and socio-economic impacts of the red import fire ant, Solenopsis invicta (Hymenoptera: Formicidae), on urban agricultural ecosystems. **Urban Ecosystems**, Urban Ecosystems, v. 23, n. 1, p. 1–12, 2020. ISSN 15731642.

CHICCO, D. Ten quick tips for machine learning in computational biology. **BioData Mining**, BioData Mining, v. 10, n. 1, p. 1–17, 2017. ISSN 17560381.

DAVIS, K. E. *et al.* Freshwater transitions and symbioses shaped the evolution and extant diversity of caridean shrimps. **Communications Biology**, Springer US, v. 1, n. 1, p. 1–7, 2018. ISSN 2399-3642.

DE PIERRE, C. R. *et al.* SWeeP: representing large biological sequences datasets in compact vectors. **Scientific Reports**, v. 10, n. 1, p. 1–10, 2020. ISSN 20452322.

DIETTERICH, T. G. Experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. **Machine Learning**, v. 40, n. 2, p. 139–157, 2000. ISSN 08856125.

DONOGHUE, M. **Travels in the Great Tree**. 2009. Disponível em: <https://peabody.yale.edu/exhibits/tree-of-life/credits>. Acesso em: 04/09/2020.

ECONOMO, E. P. *et al.* Evolution of the latitudinal diversity gradient in the hyperdiverse ant genus Pheidole. **Global Ecology and Biogeography**, v. 28, n. 4, p. 456–470, 2019. ISSN 14668238.

ECONOMO, E. P. *et al.* Global phylogenetic structure of the hyperdiverse ant genus Pheidole reveals the repeated evolution of macroecological patterns. **Proceedings of the Royal Society B: Biological Sciences**, v. 282, n. 1798, 2015. ISSN 14712954.

EMERY, L. **Phylogenetics: an introduction**. 2019. Disponível em: <https://www.ebi.ac.uk/training/online/course/introduction-phylogenetics/why-phylogenetics-important>. Acesso em: 31/08/2020.

ESTRADA, M. A. **A Diversidade e o Papel da Fauna de FOrmigas em Áreas Agrícolas Submetidas ao CUltivo Orgânico e Convencional**. Tese (Doutorado) — Universidade Federal Rural do Rio de Janeiro, 2017.

FAIRCLOTH, B. C. *et al.* Target enrichment of ultraconserved elements from arthropods provides a genomic perspective on relationships among hymenoptera. **Molecular Ecology Resources**, v. 15, n. 3, p. 489–501, 2015. ISSN 17550998.

FERNANDES, D. R. *et al.* rSWeeP: A R/Bioconductor package deal with SWeeP sequences representation. **bioRxiv**, p. 2020.09.09.290247, 2020.

FERNÁNDEZ, F.; LATTKE, J. E. Filogenia e sistemática das formigas poneromorfas. In: **As formigas poneromorfas do Brasil**. Ilhéus: Editus, 2015. cap. 8, p. 85–93. ISBN 9788574554419.

FIORAVANTI, D. *et al.* Phylogenetic convolutional neural networks in metagenomics. **BMC Bioinformatics**, BMC Bioinformatics, v. 19, n. Suppl 2, p. 1–13, 2018. ISSN 14712105.

FOLGARAIT, P. J. Ant biodiversity and its relationship to ecosystem functioning: A review. **Biodiversity and Conservation**, v. 7, n. 9, p. 1221–1244, 1998. ISSN 09603115.

FREUND, Y.; SCHAPIRE, R. E. Experiments with a New Boosting Algorithm. **Proceedings of the 13th International Conference on Machine Learning**, p. 148–156, 1996. ISSN 0706-652X, 1205-7533.

GHAHRAMANI, Z. Probabilistic machine learning and artificial intelligence. **Nature**, v. 521, n. 7553, p. 452–459, 2015. ISSN 14764687.

HAESELER, A. V. Do we still need supertrees? **BMC Biology**, v. 10, p. 2–5, 2012. ISSN 17417007.

HALGASWATHTHA, T. *et al.* Neural network based phylogenetic analysis. In: **2012 International Conference on Biomedical Engineering, ICoBE 2012**. [S.l.: s.n.], 2012. p. 155–160. ISBN 9781457719899.

HAUBOLD, B. Alignment-free phylogenetics and population genetics. **Briefings in Bioinformatics**, v. 15, n. 3, p. 407–418, 2014. ISSN 14774054.

HILL, J.; DAVIS, K. E. The Supertree Toolkit 2: A new and improved software package with a Graphical User Interface for supertree construction. **Biodiversity Data Journal**, v. 2, n. 1, 2014. ISSN 13142828.

HILLIS, D. M. *et al.* **Molecular Systematics**. 2. ed. Sunderland, MA: Sinauer Associates, Inc, 1996.

JAMES, G. *et al.* **An Introduction to Statistical Learning: with Applications in R**. New York, NY: Springer, 2013. 618 p. ISSN 01621459. ISBN 9780387781884.

JETZ, W. *et al.* The global diversity of birds in space and time. **Nature**, v. 491, n. 7424, p. 444–448, 2012. ISSN 00280836.

JORDAN, M. I.; MITCHELL, T. M. Machine learning: Trends, perspectives, and prospects. **Science**, v. 349, n. 6245, p. 255–260, 2015. ISSN 10959203.

KEMENA, C. *et al.* DOGMA: A web server for proteome and transcriptome quality assessment. **Nucleic Acids Research**, Oxford University Press, v. 47, n. W1, p. W507–W510, 2019. ISSN 13624962.

KHAN, A. *et al.* A Review of Machine Learning Algorithms for Text-Documents Classification. **Journal of Advances in Information Technology**, v. 1, n. 1, p. 4–20, 2010. ISSN 22991093.

KÜCK, P. *et al.* Improved phylogenetic analyses corroborate a plausible position of martialis heureka in the ant tree of life. **PLoS ONE**, v. 6, n. 6, 2011. ISSN 19326203.

KUSUMOTO, D.; YUASA, S. The application of convolutional neural network to stem cell biology. **Inflammation and Regeneration**, v. 39, n. 1, 2019. ISSN 18808190.

LEMEY, P. *et al.* (Ed.). **The Phylogenetic Handbook**. [S.l.]: Cambridge University Press, 2009. ISBN 9780521877107.

MARX, V. The big challenges of big data. **Nature**, v. 498, n. 7453, p. 255–260, 2013. ISSN 00280836.

MNIH, V. *et al.* Human-level control through deep reinforcement learning. **Nature**, v. 518, n. 7540, p. 529–533, 2015. ISSN 14764687.

MOREAU, C. S.; BELL, C. D. Testing The Museum Versus Cradle Tropical Biological Diversity Hypothesis: Phylogeny, Diversification, And Ancestral Biogeographic Range Evolution Of The Ants. **Evolution**, v. 67, n. 8, p. 2240–2257, 2013. ISSN 00143820.

MOREAU, C. S. *et al.* Phylogeny of the ants: Diversification in the age of angiosperms. **Science**, v. 312, n. 5770, p. 101–104, 2006. ISSN 00368075.

NOVELLO, A. *et al.* **The Phylogenetic Handbook. A Practical Approach to Phylogenetic Analysis and Hypothesis Testing**. 2. ed. [S.l.]: Cambridge University Press, 2009. ISSN 14248581. ISBN 9780521877107.

ODUM, E. P. **Fundamentos de Ecologia**. 6. ed. [S.l.: s.n.], 2007. v. 4. 612 p. ISBN 9788522105410.

OFFENBERG, J. Ants as tools in sustainable agriculture. **Journal of Applied Ecology**, v. 52, n. 5, p. 1197–1205, 2015. ISSN 13652664.

OLIVEIRA, J. C. de. **Fundamentos De Sistematica Filogenetica Para Professores de Ciencias e Biologia**. 2010. 1–9 p.

PARR, C. L. *et al.* GlobalAnts: a new database on the geography of ant traits (Hymenoptera: Formicidae). **Insect Conservation and Diversity**, v. 10, n. 1, p. 5–20, 2017. ISSN 17524598.

PETERS, R. S. *et al.* Evolutionary History of the Hymenoptera. **Current Biology**, Elsevier Ltd., v. 27, n. 7, p. 1013–1018, 2017. ISSN 09609822.

RAITTZ, R. T. *et al.* Article comparative genomics provides insights into the taxonomy of azoarcus and reveals separate origins of nif genes in the proposed azoarcus and aromatoleum genera. **Genes**, v. 12, n. 1, p. 1–21, 2021. ISSN 20734425.

RUSSELL, S. J.; NORVIG, P. **Inteligência artificial**. 3ª. ed. Rio de Janeiro: Elsevier, 2013.

SCHOCH, C. L. *et al.* NCBI Taxonomy: a comprehensive update on curation, resources and tools. **Database : the journal of biological databases and curation**, v. 2020, n. 2, p. 1–21, 2020. ISSN 17580463.

SHAKIL, K. A.; ALAM, M. Cloud computing in bioinformatics and big data analytics: Current status and future research. **Advances in Intelligent Systems and Computing**, v. 654, n. January, p. 629–640, 2018. ISSN 21945357.

SILVA, B. L.; ROSA, A. C. M. **Controle da formiga cortadeira (Atta sexdens rubropilosa) em agricultura orgânica no bioma Cerrado**. 1–20 p. Tese (Doutorado) — Universidade Federal de Santa Catarina, 2017.

SOBER, E. The contest between parsimony and likelihood. **Systematic Biology**, v. 53, n. 4, p. 644–653, 2004. ISSN 10635157.

TANG, P. *et al.* Mitochondrial phylogenomics of the Hymenoptera. **Molecular Phylogenetics and Evolution**, Elsevier, v. 131, n. June 2018, p. 8–18, 2019. ISSN 10959513.

THOMAS, G. H. *et al.* PASTIS: An R package to facilitate phylogenetic assembly with soft taxonomic inferences. **Methods in Ecology and Evolution**, v. 4, n. 11, p. 1011–1017, 2013. ISSN 2041210X.

VALAFAR, F. Neural Network Applications in Biological Sequencing. In: DEPARTMENT OF COMPUTER SCIENCE, SAN DIEGO STATE UNIVERSITY,. **Proceedings of the 2003 International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences**. San Diego, California, 2003. p. 24—-27.

VALAN, M. *et al.* Automated Taxonomic Identification of Insects with Expert-Level Accuracy Using Effective Feature Transfer from Convolutional Networks. **Systematic Biology**, v. 68, n. 6, p. 876–895, 2019. ISSN 1076836X.

VALENTINI, G.; MASULLI, F. Ensembles of learning machines. **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**, v. 2486 LNCS, p. 3–20, 2002. ISSN 16113349.

VAN DER MAATEN, L. J. P. *et al.* Dimensionality Reduction: A Comparative Review. **Journal of Machine Learning Research**, v. 10, p. 1–41, 2009. ISSN 0169328X.

VINGA, S. Editorial: Alignment-free methods in computational biology. **Briefings in Bioinformatics**, v. 15, n. 3, p. 341–342, 2014. ISSN 14774054.

Ward P. Phylogeny, classification, and species-level taxonomy of ants (Hymenoptera: Formicidae)*. **Zootaxa**, v. 563, p. 549 – 563, 2007. ISSN 11755326.

WARNOW, T. Supertree Construction: Opportunities and Challenges. **arXiv**, 2018.

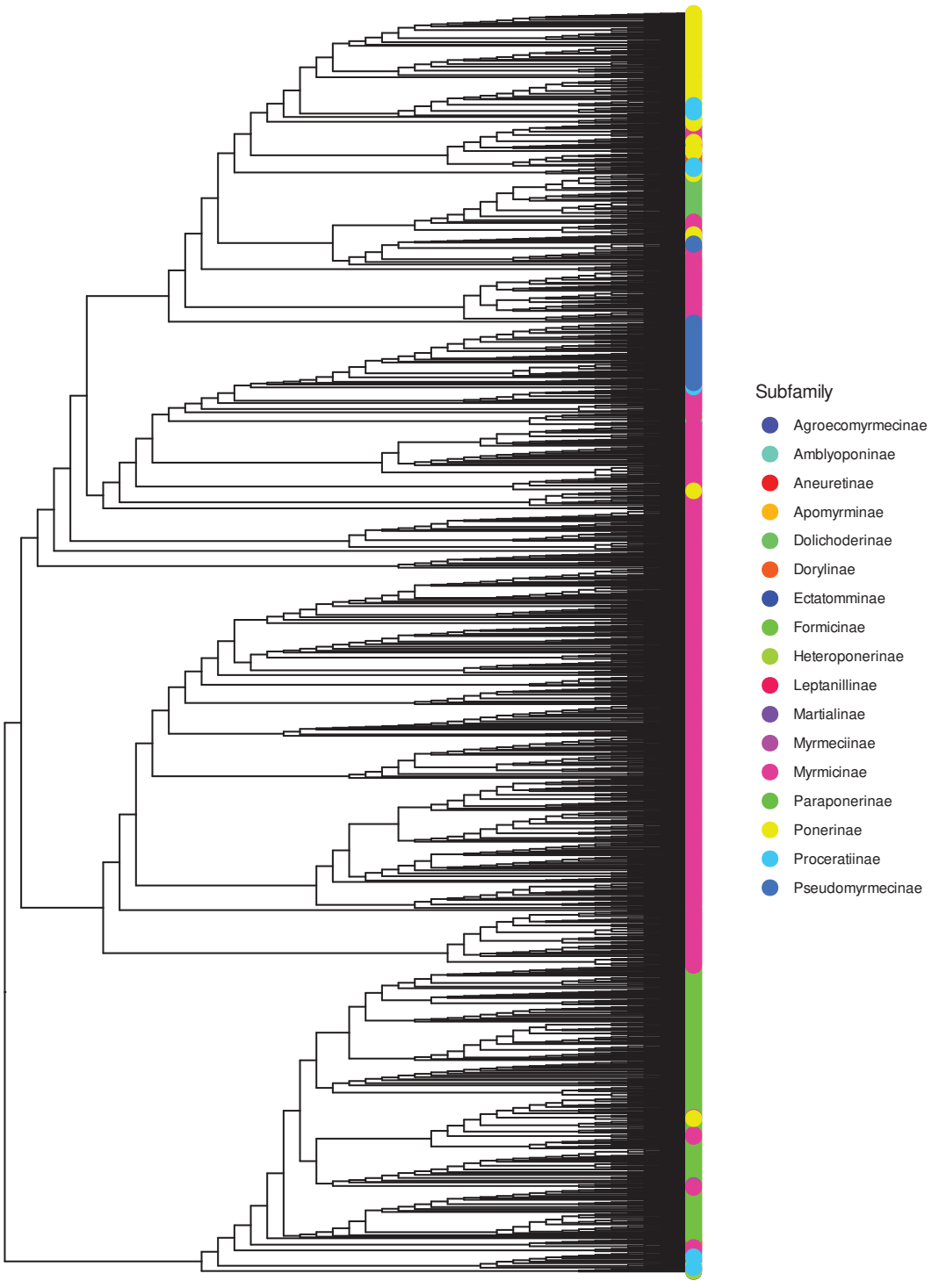
WEISER, M. D.; KASPARI, M. Ecological morphospace of New World ants. **Ecological Entomology**, v. 31, n. 2, p. 131–142, 2006. ISSN 03076946.

XIA, X. Phylogenetic bias in the likelihood method caused by missing data coupled with among-site rate variation: An analytical approach. **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**, v. 8492 LNBI, p. 12–23, 2014. ISSN 16113349.

YIP, K. Y. *et al.* Machine learning and genome annotation: A match meant to be? **Genome Biology**, v. 14, n. 5, 2013. ISSN 1474760X.

ZIELEZINSKI, A. *et al.* Alignment-free sequence comparison: Benefits, applications, and tools. **Genome Biology**, Genome Biology, v. 18, n. 1, p. 1–17, 2017. ISSN 1474760X.

APÊNDICE A – ÁRVORE GERADA PELA MAM



FONTE: A Autora (2020)