Atividade Final de Fiologenia Molecular

Gabriel Victor Pina Rodrigues *
Universidade Estadual de Santa Cruz,
gvpina.rodrigues@gmail.com

Abstract. Atividade avaliativa realizada com a finalidade de aplicar todos os conceitos teóricos e práticos aprendidos em sala de aule e durante os estudos dirigidos. Essa atividade se propõe a fazer uma análise filogenética de 43 sequências do gene ribossomal mitochondrial 16S de abelhas sem ferrão (tribo Meliponini).

Keywords. Filogenia; MSA; Bioinformática; Abelhas

Introdução

O uso de métodos moleculares é essencial para contar a história dos eventos evolutivos que ocorreram com um táxon em determinado tempo. As evidências necessárias para descrever o precesso devem ser coletadas com critérios rígidos, para não gerar dados ruidosos em uma reconstrução filogenética.

Métodos filogenéticos são baseados em inferências, já que não é possível reconstruir os elementos já perdidos na natureza. Na maioria dos trabalhos, as amostras serão parciais. Fatores como a diversidade interna do gênero/família são chave para determinar a completude da filogenia molecular.

Para esse trabalho em particular, iremos usar o arquivo `seqs.fasta`, que possui 43 sequências do gene ribossomal mitochondrial 16S de abelhas sem ferrão (tribo Meliponini).

Escolha do Caractere

Um caractere, seja molecular ou morfológico, deve ter uma variação intrísceca entre as amostras (táxons estudados), porém sem ser muito divergente ao ponto de não poder ser utilizado como um caractere conservado entre as amostras.

Alguns dos caracteres escolhidos vão estar ausentes, no caso de amostras moleculares pode ser o caso de uma base nitrogenada não sequenciada ou com baixa qualidade.

Objeto da Análise

A subunidade 16S do RNA Ribossomal é eficaz para resconstruir relações evolutivas devido à propriedades como distribuição universal, função conservada, taxa de evolução lenta e a presença de regiões variáveis em sua composição.

*Corresponding author.

Curadoria das Sequências

Em primeiro lugar, é necessário padronizar os nomes das sequências de input para as análises subsequentes. Como os arquivos foram retirados do GenBank, eles possuem a seguinte estrutura de *header*:

```
>AF343118.1 Lepidotrigona ventralis 16S large subunit ribosomal RNA gene
TTGTATATTTGTATAATGAAATCTGGAATGAAAGGATTAATGAAATAT
```

Em arquivos FASTA, o *header* é indicado como sendo o elemento após o símbolo de $>$ (maior que), indicando o nome da sequência e outras informações relevantes. Para as análises futuras, é vantajoso reduzir o nome para uma melhor visualização da árvore e suporte dos softwares de alinhamento. Podemos reduzir para a seguinte estrutura:

```
>Lepidotrigona_ventralis
TTGTATATTTGTATAATGAAATCTGGAATGAAAGGATTAATGAAATAT
```

Propriedade das Sequêcias

No caso das sequências analisadas, tem-se uma média de 424,46 nucleotídeos com um desvio padrão de 37,6 nucleotídeos. Apesar das diferenças em tamanho, não são observadas sequências *outliers* nesse quesito (Figure 1).

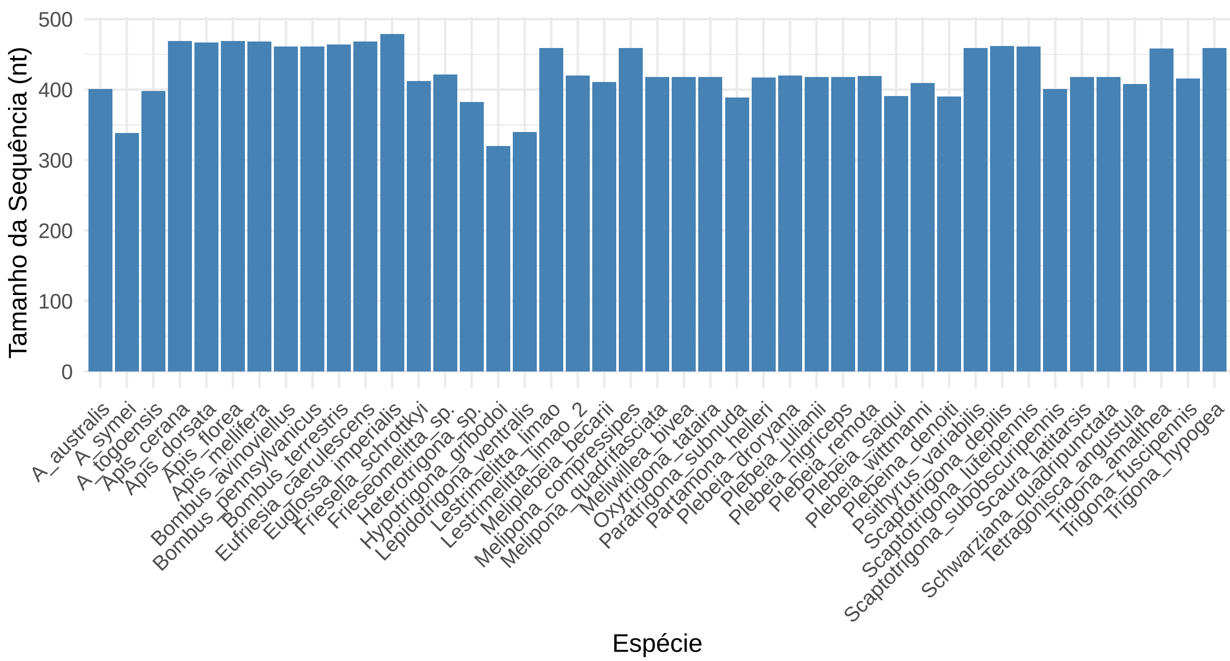

Figure 1. Tamanho das sequências em nucleotídeos por espécie analisada.

Sobre o conteúdo GC, houve bastante variação apresentada pelas sequências de forma total, com uma média de aproximadamente 19,2% de conteúdo GC. Porém o desvio padrão apresentado é de 1,92% (Figure 2).

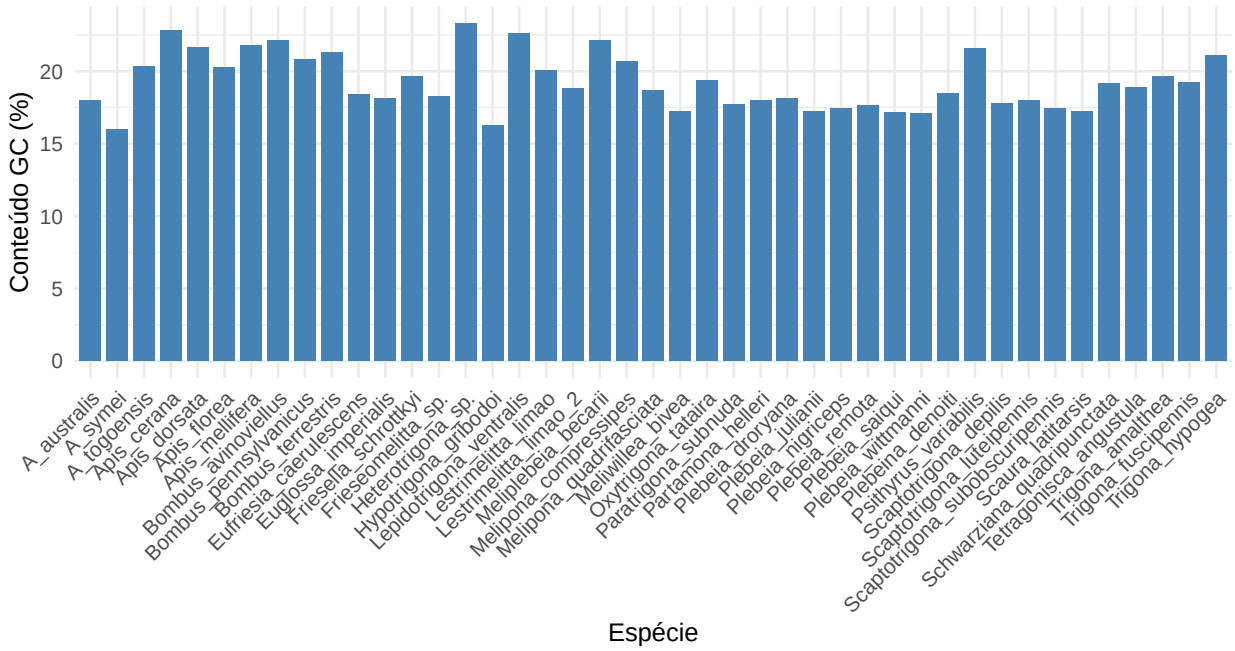

Figure 2. Conteúdo GC das sequências em porcentagem por espécie analisada.

Alinhamento Múltiplo de Sequências (MSA)

MAFFT (Multiple Alignment with Fast Fourier Transform, em inglês) é um programa de alinhamento múltiplo de sequências. O programa MAFFT implementa transformada rápida de Fourier para otimizar os alinhamentos de proteínas com base nas propriedades físicas dos aminoácidos (Katoh et al. 2002). O programa usa o alinhamento progressivo e iterativo. As sequências nucleotídicas e de aminoácidos no formato FASTA podem ser alinhadas. MAFFT é útil para as sequências mais difíceis de se alinhar com outros programas, tais como aquelas que contêm grandes lacunas, (ex: rRNA que contêm as regiões variáveis de laços).

O MAFFT foi instalado no sistema operacional ArchLinux via ambiente conda, através do comando:

```
conda install -c bioconda mafft
```

A partir da sua instalação, podemos executar o alinhamento com o MAFFT a partir do seguindo comando:

```
mafft --maxiterate 1000 --localpair seqs.fasta > alinhamento.fasta
```

O argumento `--maxiterate 1000` diz respeito ao número máximo de iterações (ciclos de refinamento) que o algoritmo deve tentar para otimizar o alinhamento. Ou seja, o MAFFT começa com um alinhamento rápido inicial e, em seguida, tenta melhorá-lo repetidamente. Definir um valor alto (como 1000) faz com que o algoritmo refine o alinhamento por mais tempo, levando a uma maior precisão em comparação com as opções rápidas

A função `--localpair` indica uma instrução para usar a estratégia de alinhamento local par a

par. Ela significa que o alinhamento usa informações de regiões locais altamente conservadas entre cada par de sequências, antes de construir a árvore-guia e o alinhamento múltiplo final.

Matriz de Distância dos Pares

Após o alinhamento de sequências de nucleotídeos ou aminoácidos, o passo subsequente em diversas abordagens filogenéticas envolve o cálculo das distâncias pareadas. Esta métrica quantifica o grau de divergência acumulada entre duas sequências ao longo do tempo evolutivo, oferecendo uma medida fundamental para elucidar as relações evolutivas entre elas. Tais distâncias são a base de métodos de construção de árvores, como UPGMA (Unweighted Pair Group Method with Arithmetic mean) e Neighbor-Joining.

A escolha do modelo de distância é crucial, pois cada um oferece uma perspectiva estatística distinta sobre a divergência. Inicialmente, distâncias mais simples como a distância de Hamming (que conta as diferenças brutas) e a p-distância (que normaliza essas diferenças brutas) podem ser utilizadas. No entanto, modelos mais sofisticados de substituição de nucleotídeos são tipicamente empregados para corrigir o problema de múltiplas substituições em um único sítio (que a p-distância não contabiliza).

Modelos de Substituição de Nucleotídeos

Os modelos de distância assumem diferentes taxas e processos de substituição, com a complexidade aumentando à medida que mais parâmetros são incorporados. O modelo mais simples é o de Jukes-Cantor (JC69), que postula uma taxa de substituição igual entre quaisquer dois nucleotídeos e assume frequências iguais para as quatro bases (A, T, C, G). Pode-se observar visuamente a matriz JC69 a partir da Figure 3.

Modelos subsequentes introduzem maior realismo biológico. O modelo de Kimura de 2 Parâmetros (K80), por exemplo, relaxa a premissa do JC69 ao distinguir as taxas entre transições (substituições entre bases do mesmo tipo, como A $\leftrightarrow$ G ou C $\leftrightarrow$ T) e transversões (substituições entre bases de tipos diferentes, como A $\leftrightarrow$ C ou G $\leftrightarrow$ T), mas ainda assume frequências iguais das bases (Figure 4).

Neighbor Joining (NJ)

O Neighbor-Joining (NJ) constitui um método de inferência filogenética classificado como baseado em distâncias. Sua aplicação ocorre após o cálculo das distâncias pareadas entre todas as sequências em um conjunto de dados (Yang and Rannala 2012).

Para fazer a seguinte árvore, foi utilizado o pacote **phangorn** no R. Para a eexecução foi necessário importar o alinhamento realizado no MAFFT, o resultado pode ser observado na Figure 5.

```
# Calcular a matriz de distância de Hamming com os dados filogenéticos
D_hamming = dist.hamming(phyDat_msa_sample)
```

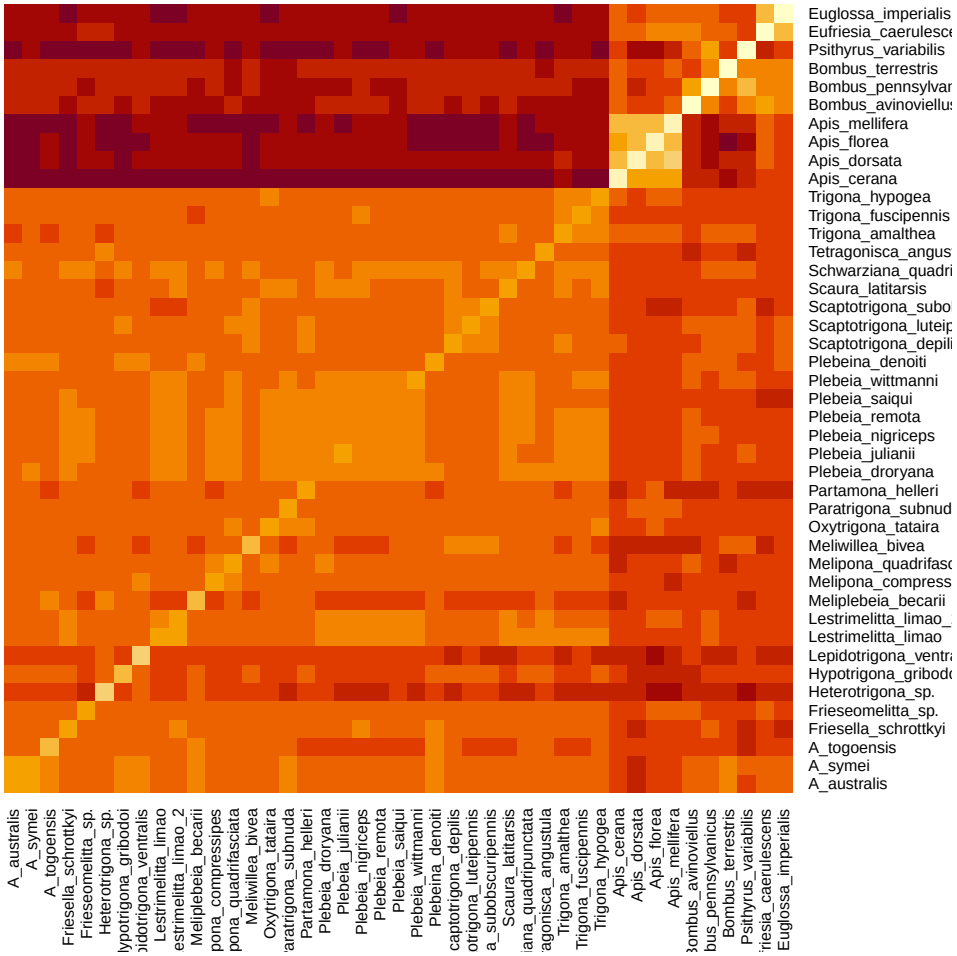

Figure 3. Matriz de distância baseada no método de Jukes-Cantor (JC69). Quanto mais claro, mais semelhante.

```
# Processar a árvore de Neighbor Joining
nj_tree = nj(D_hamming)
nj_tree$edge.length[which(nj_tree$edge.length<0)]=0
nj_tree = midpoint(multi2di(nj_tree))
```

Máxima Parcimônia (MP)

A Máxima Parcimônia (MP) é um método de inferência filogenética classificado como baseado em caracteres (character-based). Em contraste com os métodos baseados em distâncias (como o Neighbor-Joining), a parcimônia utiliza diretamente as características observadas nas sequências (os nucleotídeos ou aminoácidos) para avaliar as topologias de árvores (Yang and Rannala 2012).

O objetivo fundamental da Máxima Parcimônia é identificar a topologia de árvore que requer o menor número total de eventos evolutivos (substituições, inserções ou deleções) para explicar

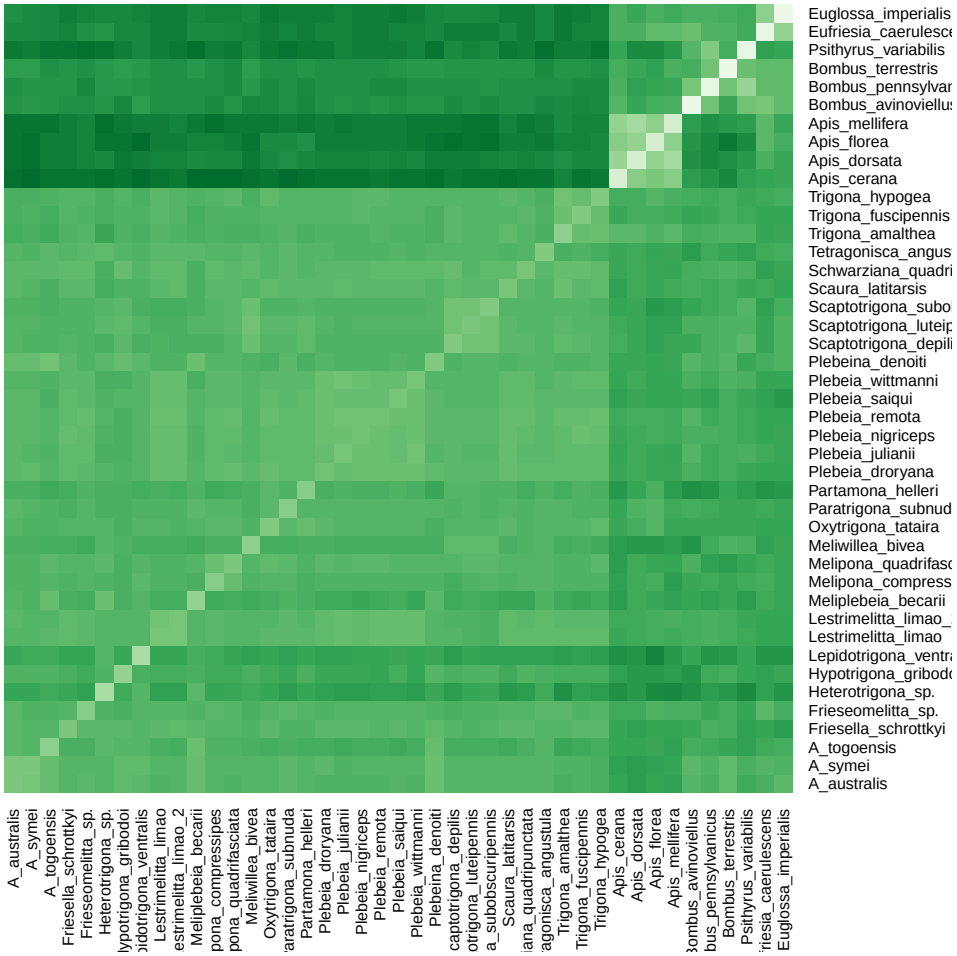

Figure 4. Matriz de distância baseada no modelo de Kimura de 2 Parâmetros (K80). Quanto mais claro, mais semelhante.

os dados observados. Em essência, o método busca a árvore que é a mais simples e mais "parcimoniosa" possível. A premissa subjacente é a de que a rota evolutiva mais provável é aquela que minimiza a ocorrência de eventos evolutivos independentes.

A árvore final com o resultado desse método nos dados pode ser observada na Figure 6.

```
Final p-score 977 after  8 nni operations
```

Modelo de Substituição Nucleotídica

IQ-TREE

Modelo do IQ-TREE: ModelFinder computes the log-likelihoods of an initial parsimony tree for many different models and the Akaike information criterion (AIC) corrected Akaike information criterion (AICc), and the Bayesian information criterion (BIC). Then ModelFinder chooses the model that minimizes the BIC score (you can also change to AIC or AICc by adding the option -AIC or -AICc, respectively).

Bibliografia

Katoh, Kazutaka, Kazuharu Misawa, Kei-ichi Kuma, and Takashi Miyata. 2002. "MAFFT: A Novel Method for Rapid Multiple Sequence Alignment Based on Fast Fourier Transform." *Nucleic Acids Research* 30 (14): 3059–66. https://doi.org/10.1093/nar/gkf436.

Yang, Ziheng, and Bruce Rannala. 2012. "Molecular Phylogenetics: Principles and Practice." *Nature Reviews Genetics* 13 (5): 303–14. https://doi.org/10.1038/nrg3186.

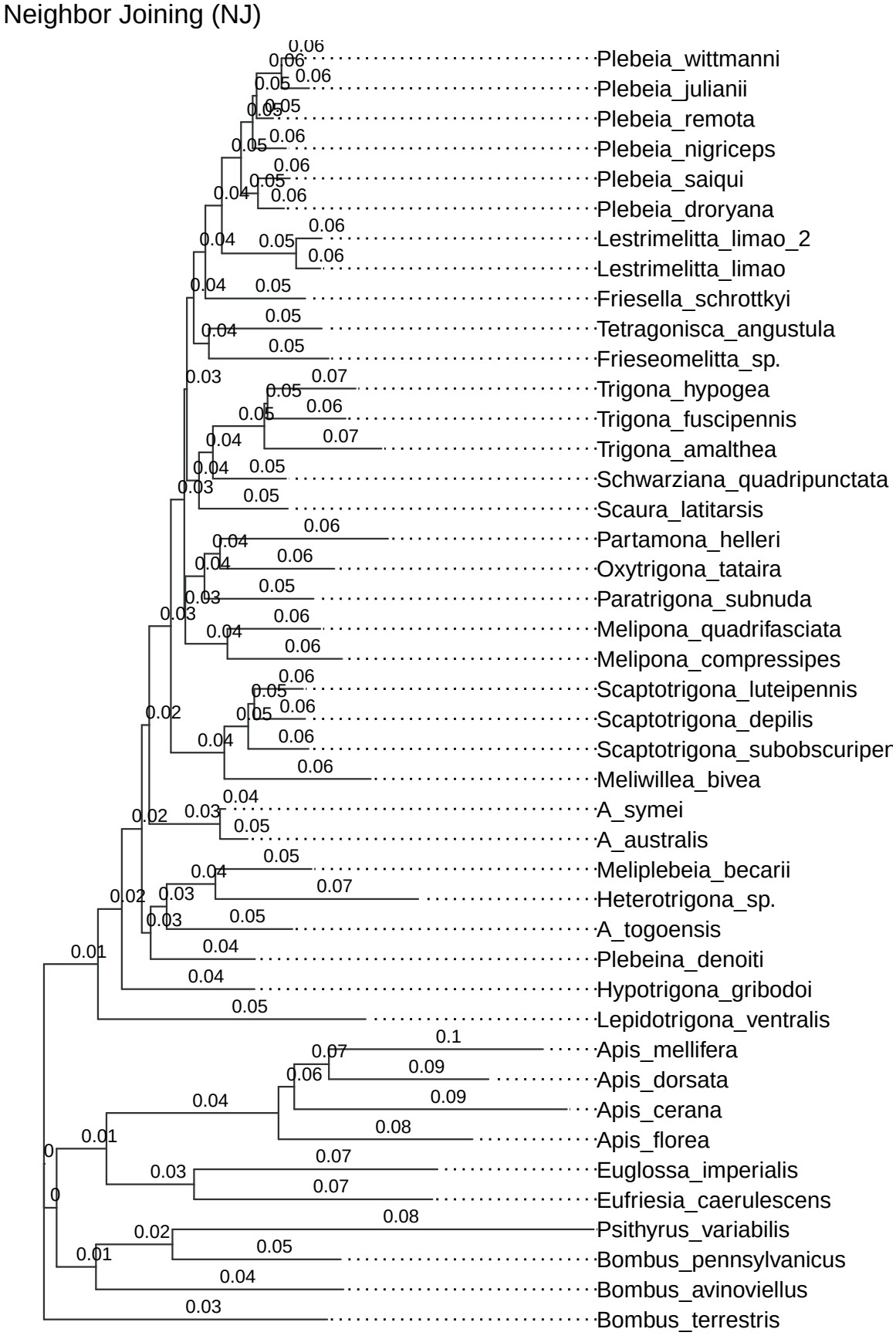

Figure 5. Árvore filogenética feita a partir do método de Neighbor-Joining (NJ)

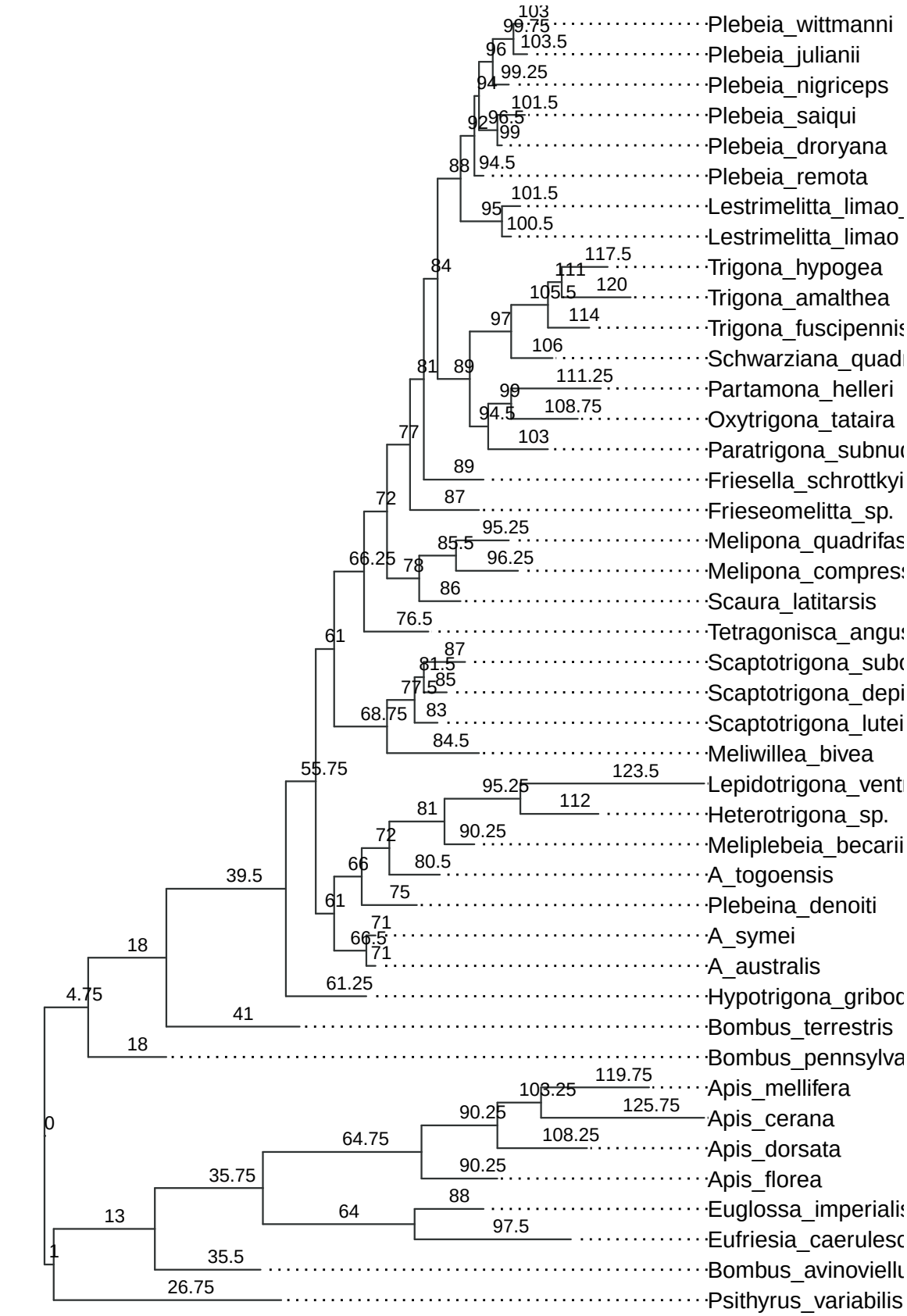

Figure 6. Árvore filogenética feita a partir do método de Máxima Parcimônia (MP).