

PhyloTreeMiner: Integração de Dados Heterogêneos e Consultas em Grafos para Mineração de Padrões Filogenéticos^{*†}

João Vitor Moraes¹, Layse Gomes Castro¹, Isabel Rosseti¹, Daniel de Oliveira¹

¹Instituto de Computação – Universidade Federal Fluminense (UFF)

{joaovitormoraes, laysegomes}@id.uff.br, {danielcmo, rosseti}@ic.uff.br

Resumo. A análise filogenética permite compreender relações evolutivas, mas diferentes métodos de inferência frequentemente geram múltiplas árvores a partir do mesmo conjunto de dados. Este artigo apresenta o *PhyloTreeMiner*, um sistema de mineração de padrões filogenéticos a partir de múltiplas árvores filogenéticas. A abordagem modela o problema como a mineração de itemsets frequentes maximais, identificando subárvores recorrentes que representam relações evolutivas robustas. Os resultados são enriquecidos e armazenados em um grafo de conhecimento, permitindo consultas analíticas e exploração interativa.

1. Introdução

Com o avanço das tecnologias de sequenciamento de DNA nas últimas décadas, um grande volume de dados biológicos tornou-se disponível [Metzker 2010]. Esse cenário tem impulsionado o desenvolvimento de diversas aplicações, dentre as quais se destaca a *Análise Filogenética* [Felsenstein 2004], fundamental para a compreensão das relações evolutivas entre organismos. A análise filogenética busca inferir, a partir de sequências biológicas (*e.g.*, DNA, RNA ou aminoácidos), padrões evolutivos entre os organismos analisados. Tais relações são tipicamente representadas por árvores filogenéticas, estruturas hierárquicas que modelam hipóteses sobre a evolução e a divergência das espécies ao longo do tempo.

Diversos trabalhos exploram análises filogenéticas em diferentes contextos, incluindo o desenvolvimento de novos fármacos [Ocaña et al. 2012] e a vigilância genômica de patógenos, especialmente em cenários de surtos infecciosos [Baker et al. 2023]. Entretanto, a análise filogenética está sujeita a variações decorrentes do uso de diferentes métodos, modelos evolutivos e parâmetros, o que frequentemente resulta em múltiplas árvores plausíveis a partir de um mesmo conjunto de dados. A prática de selecionar uma única árvore para análise pode ocultar conflitos topológicos e limitar a interpretação dos padrões evolutivos.

Nesse contexto, torna-se necessário analisar conjuntamente as múltiplas árvores geradas. Uma abordagem promissora consiste em identificar subárvores recorrentes em uma floresta filogenética [Deepak et al. 2014], uma vez que padrões de alta coocorrência tendem a refletir relações evolutivas mais robustas. Contudo, essa tarefa apresenta desafios importantes. Primeiramente, cada árvore pode gerar dezenas ou centenas de subárvores, o que torna a comparação entre elas computacionalmente custosa. Além disso, a análise dos padrões extraídos é dificultada pelo grande volume de resultados e pela ausência de metadados associados, o que frequentemente leva à implementação de *scripts ad-hoc*, pouco escaláveis e suscetíveis a erros [Marx 2013].

^{*}Vídeo demonstrativo da ferramenta pode ser acessado em <https://youtu.be/qh-8OCVBun0>

[†]Os autores agradecem à CAPES, ao CNPq e à FAPERJ pelo apoio parcial à realização deste trabalho.

Para enfrentar esses desafios, este artigo apresenta o PhyloTreeMiner, um sistema de mineração de padrões filogenéticos baseado em um *workflow* que extrai e identifica subárvores frequentes e integra os resultados da mineração com metadados biológicos extraídos de bases de dados públicas. O PhyloTreeMiner modela o problema de comparação entre subárvores como uma tarefa de mineração de *itemsets* [Agrawal and Srikant 1994], permitindo identificar, de forma eficiente, relações evolutivas recorrentes por meio da descoberta de padrões maximais. Os resultados são organizados em um grafo de conhecimento [dos Santos et al. 2025], instanciado no Neo4j, o que viabiliza a execução de consultas analíticas em Cypher que integram informações topológicas e metadados, incluindo dados geográficos. A demonstração do PhyloTreeMiner utiliza um estudo de caso de vigilância genômica com dados do vírus Zika, abrangendo diferentes regiões do mundo durante o surto de 2016 [Zadra et al. 2024].

2. O Sistema PhyloTreeMiner

O PhyloTreeMiner segue uma arquitetura modular composta por três componentes principais, como ilustrado na Figura 1: (i) o portal PhyloTreeMiner, responsável pela interação com o usuário; (ii) o *workflow* PhyloTreeMiner, encarregado da execução, orquestração e paralelização das análises; e (iii) o repositório de dados, que integra mecanismos de armazenamento de proveniência e um grafo de conhecimento para suporte à consulta e à análise dos resultados.

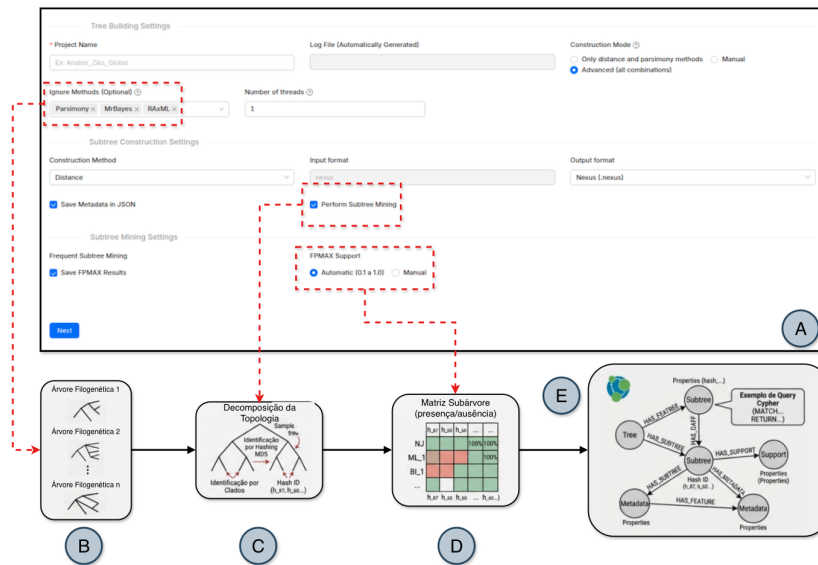


Figura 1. A Arquitetura do PhyloTreeMiner.

O portal PhyloTreeMiner (região A na Figura 1) constitui a camada de apresentação do sistema, sendo responsável por acionar o *workflow* que executa a mineração de subárvores filogenéticas e por apoiar a visualização dos resultados. Além disso, permite definir projetos experimentais que encapsulam múltiplas execuções desse *workflow* de análise filogenética. Por meio dessa interface, o usuário pode especificar os conjuntos de dados de entrada, os parâmetros de execução e as estratégias analíticas. Entre essas configurações, incluem-se a seleção ou exclusão de métodos de inferência filogenética e o controle do grau de paralelismo, por meio da definição do número de *threads*, conforme pode ser visto na área A na Figura 1. Essa abordagem de abstração permite que usuários com diferentes níveis

de experiência configurem experimentos complexos de forma simples, assegurando a consistência e a reprodutibilidade dos resultados.

Uma vez que os parâmetros são definidos pelo usuário, o *workflow* PhyloTreeMiner é invocado (parte inferior da Figura 1) e executa as etapas de processamento de dados biológicos, inferência filogenética, mineração de padrões, integração de dados e enriquecimento semântico. Na etapa inicial, denominada *Geração das Árvores Filogenéticas* (B), sequências biológicas (*e.g.*, DNA, RNA e aminoácidos) no formato FASTA são processadas por meio de rotinas da biblioteca *Biopython*. Nessa fase, são realizados o *Alinhamento Múltiplo de Sequências* e a *Inferência de Árvores Filogenéticas*, empregando diferentes algoritmos de alinhamento e múltiplas abordagens de inferência, *e.g.*, *Neighbor Joining* (NJ), *Máxima Verossimilhança* (ML) e *Inferência Bayesiana* (BI), com o suporte de ferramentas como IQ-TREE, RAxML, FastTree e MrBayes. Essa diversidade metodológica possibilita a geração de múltiplas árvores a partir do mesmo conjunto de dados, constituindo uma “floresta” filogenética que captura diferentes hipóteses evolutivas. Uma vez executadas essas atividades do *workflow*, o portal PhyloTreeMiner oferece recursos para a análise visual dos resultados, permitindo ao usuário inspecionar tanto os alinhamentos quanto as árvores filogenéticas geradas (Figura 2).

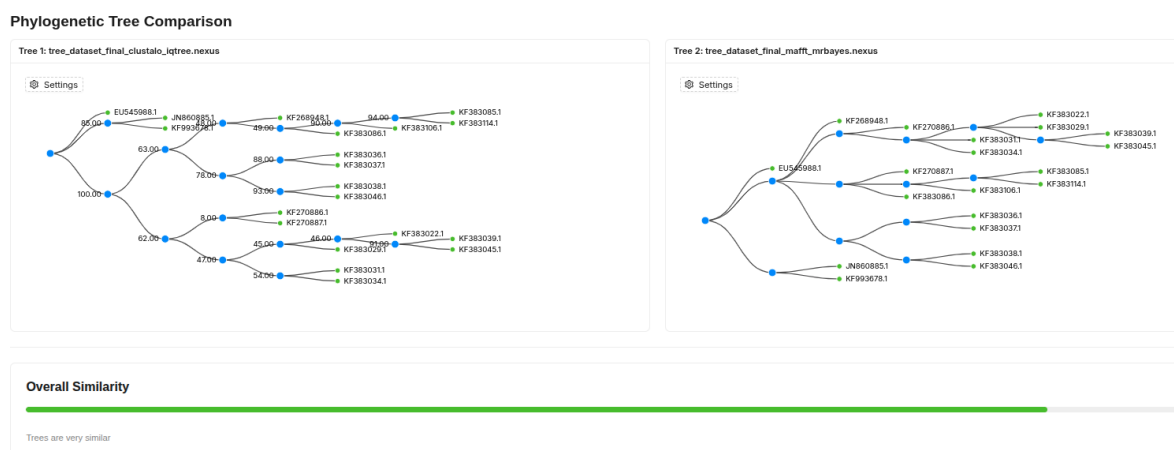


Figura 2. Comparação de duas árvores filogenéticas geradas no PhyloTreeMiner.

Uma vez geradas, as árvores são decompostas na atividade de *Fragmentação das Árvores* (C). Nessa etapa, cada árvore é decomposta em subárvores (*i.e.*, clados) por meio de travessias topológicas. As subárvores extraídas são convertidas em identificadores únicos por meio de funções *hash* aplicadas sobre os metadados brutos gerados pela biblioteca *Biopython*. É fundamental destacar que a equivalência é estrita: subárvores com a mesma topologia, mas com parâmetros distintos, *e.g.*, diferentes comprimentos de ramo, geram *hashes* distintos. Isso garante que não haja colapso de clados divergentes no mesmo item, mapeando estruturas com precisão em um espaço discreto. Essa transformação reduz a complexidade da análise ao representar os dados filogenéticos como um modelo transacional, no qual cada árvore é tratada como uma transação e cada subárvore como um item dessa transação, como no problema de mineração de *itemsets*. Com base nessa representação, o PhyloTreeMiner constrói uma matriz de incidência Subárvore–Árvore na atividade *Construção da Matriz* (D), em que as linhas correspondem às árvores e as colunas aos identificadores das subárvores, com valores binários que indicam a presença ou a ausência. Essa estrutura serve como entrada para a atividade de *Mineração de Padrões Frequentes*, realizada com o algoritmo

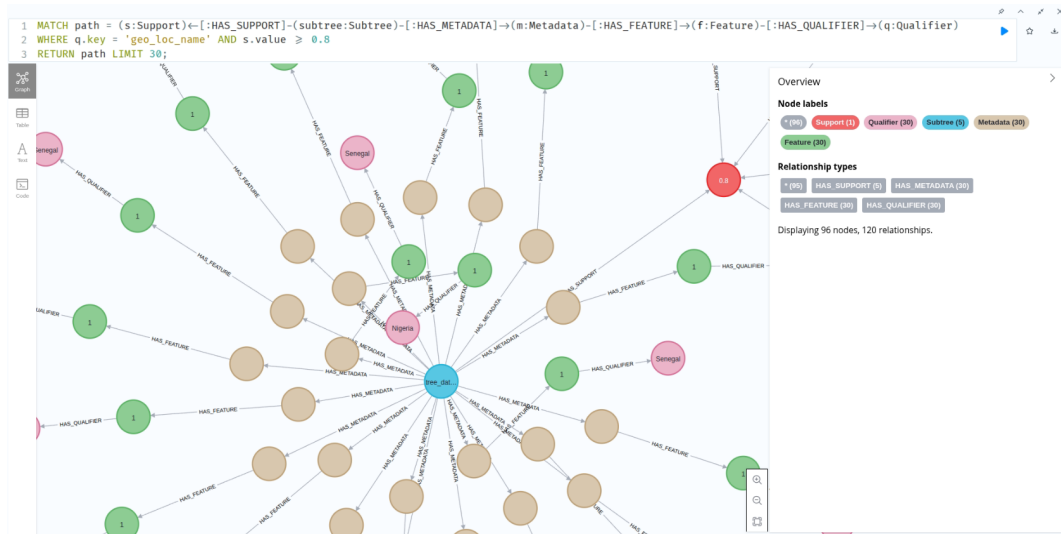


Figura 3. Exemplo de consulta estrutural no Neo4j relacionando clados de alto suporte ($\geq 80\%$) aos metadados geográficos.

FPMMax [Grahne and Zhu 2003], que extrai conjuntos maximais. Os padrões obtidos representam conjuntos de subárvores que coocorrem em múltiplas árvores, sendo interpretados como relações filogenéticas consistentes entre diferentes métodos de inferência.

O resultado da mineração das subárvores **E** é enriquecido com bases de dados externas, em especial com o NCBI [O’Leary et al. 2024]. Com base nos identificadores das sequências presentes nas subárvores, o PhyloTreeMiner recupera metadados relevantes, como a taxonomia a que pertencem, a localização geográfica da sequência, a data de coleta e informações bibliográficas, por meio de APIs públicas. Esse processo de enriquecimento conecta os padrões estruturais identificados na mineração ao contexto biológico e epidemiológico, ampliando o potencial analítico. Todas as sequências, árvores filogenéticas, subárvores, padrões frequentes, parâmetros experimentais e metadados enriquecidos são integrados ao repositório de dados. Esse repositório combina um banco de dados de proveniência, responsável por registrar o histórico completo das execuções, com um grafo de conhecimento implementado no Neo4j. Nesse grafo, entidades como TREE, SUBTREE, SUPPORT e METADATA são representadas como nós, enquanto relações como HAS_SUBTREE, CO_OCCURS_WITH e HAS_METADATA são modeladas como arestas. Assim, o grafo gerado (Figura 3) combina características topológicas com metadados espaciais. Por exemplo, pode-se identificar regiões geográficas que concentram os padrões evolutivos mais confiáveis agregando a média do valor de suporte (*s.value*) dos clados associados a um país específico (*geo_loc_name*).

3. Demonstração

A demonstração do PhyloTreeMiner baseia-se em um cenário real de vigilância genômica do vírus Zika (ZIKV) [Zadra et al. 2024]. A vigilância genômica de patógenos desempenha um papel fundamental no monitoramento da evolução e da disseminação de agentes infecciosos, permitindo a identificação de variantes e a análise de sua distribuição em diferentes contextos geográficos. Por meio da mineração de subárvores frequentes, o PhyloTreeMiner identifica padrões evolutivos consistentes entre diferentes metodologias de inferência, reduzindo o ruído e as variações topológicas.

Na demonstração, o PhyloTreeMiner consome um conjunto de 479 sequências genômicas completas, com aproximadamente 10.000 nucleotídeos cada, previamente curadas para remover regiões altamente variáveis e sinais de recombinação. Esse conjunto apresenta diversidade geográfica, incluindo amostras da África, da Ásia, das Ilhas do Pacífico e das Américas, o que permite a detecção de diferentes linhagens evolutivas do ZIKV. A demonstração segue o *workflow* PhyloTreeMiner, composto pelas seguintes etapas: (i) o sistema constrói automaticamente múltiplas árvores filogenéticas a partir de diferentes combinações de métodos de alinhamento múltiplo e algoritmos de inferência; (ii) as árvores geradas têm suas subárvores extraídas e transformadas em representações adequadas à mineração, possibilitando a identificação de padrões frequentes; (iii) os padrões são integrados a um grafo de conhecimento enriquecido com metadados; e (iv) o usuário pode consultar, filtrar e visualizar padrões evolutivos relevantes por meio de consultas ao grafo.

Durante a demonstração, serão geradas 16 árvores, correspondentes a todas as combinações entre métodos de alinhamento e algoritmos de inferência filogenética. O portal PhyloTreeMiner permite inspecionar as árvores resultantes e comparar suas topologias (como na Figura 2). O usuário poderá observar variações significativas entre as árvores, decorrentes das diferentes escolhas metodológicas. Essa variabilidade é explorada pelo sistema para identificar padrões robustos que se repetem independentemente da abordagem utilizada. Na demonstração, serão consideradas subárvores com suporte $\geq 80\%$. Por exemplo, para o Brasil, de um total de 31.633 subárvores extraídas, apenas 269 (aproximadamente 0,85%) atendem a esse critério. Valores de suporte elevados (*e.g.*, $\geq 80\%$) atuam como um filtro rigoroso que prioriza a precisão, isolando padrões altamente conservados e mitigando o ruído topológico inerente às divergências entre métodos de inferência. Esse resultado pode ser explorado dinamicamente pelo usuário, que pode ajustar o limiar de suporte e observar seu impacto no número de padrões retornados.

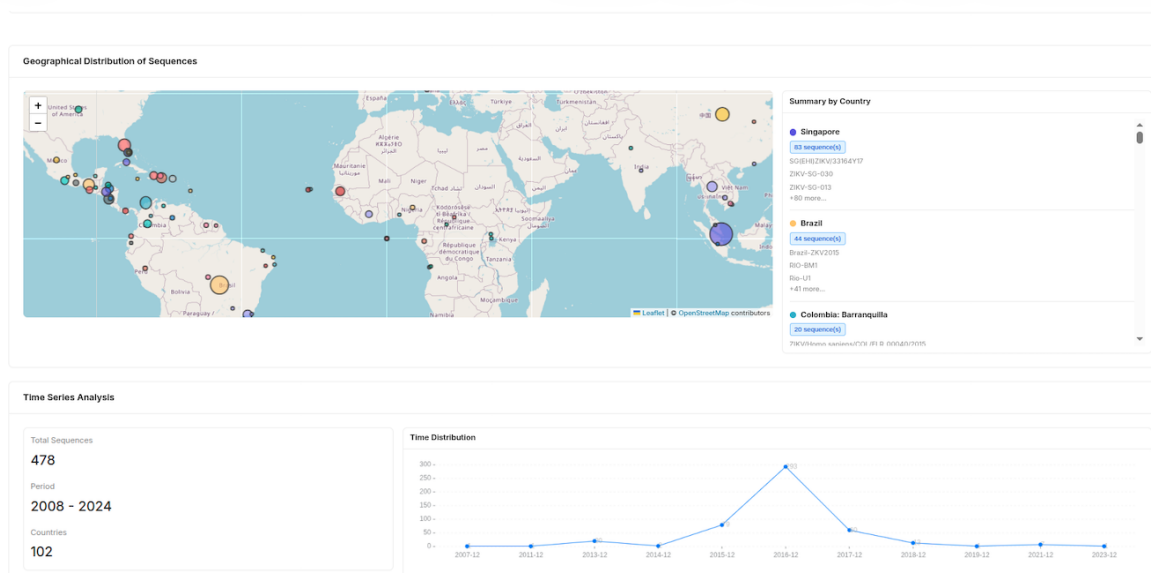


Figura 4. Distribuição geográfica do ZIKV no PhyloTreeMiner.

Além disso, o sistema apresenta metadados geográficos, permitindo análises por país e a visualização da distribuição geográfica das variantes do ZIKV (Figura 4). A demonstração evidencia diferenças relevantes entre as regiões. Em particular, Cingapura apresenta o maior número total de subárvores (38.302), porém com o menor suporte médio (11,70%), indi-

cando alta variabilidade topológica. Essa funcionalidade permite identificar automaticamente regiões com maior complexidade evolutiva, possivelmente associadas a múltiplas introduções virais ou a maior diversidade genética. Por fim, o sistema destaca automaticamente as subárvores com maior suporte, permitindo a inspeção direta dos padrões evolutivos mais consistentes.

4. Conclusões

Este artigo apresentou o PhyloTreeMiner, um sistema de mineração de padrões filogenéticos que integra múltiplas árvores inferidas a partir do mesmo conjunto de sequências. Ao modelar o problema como uma tarefa de mineração de *itemsets* frequentes maximais, o PhyloTreeMiner permite identificar subestruturas recorrentes que representam relações evolutivas mais robustas, mitigando as limitações associadas à análise baseada em uma única árvore. Além disso, a integração dos padrões extraídos a um grafo de conhecimento enriquecido com metadados biológicos e geográficos amplia as possibilidades analíticas, permitindo consultas expressivas e a exploração interativa dos resultados. A demonstração com dados reais do ZIKV evidenciou a capacidade do sistema de capturar padrões consistentes e de revelar variações relevantes entre diferentes contextos geográficos. Como trabalhos futuros, destacam-se a incorporação de novos métodos de inferência filogenética, a integração com bases de dados adicionais e a exploração de técnicas avançadas de visualização para apoiar ainda mais a interpretação dos padrões descobertos.

Referências

- Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules. In *Proceedings of the 20th VLDB*, pages 487–499.
- Baker, K. S., Jauneikaite, E., et al. (2023). Genomics for public health and international surveillance of antimicrobial resistance. *The Lancet Microbe*, 4(12):e1047–e1055.
- Deepak, A., Fernández-Baca, D., et al. (2014). Evominer: frequent subtree mining in phylogenetic databases. *Knowl. Inf. Syst.*, 41(3):559–590.
- dos Santos, V., de Oliveira, D., et al. (2025). Mind the context: Enriching knowledge graphs with rules and mappings. In *IEEE ICKG, Cyprus, 2025*, pages 59–66. IEEE.
- Felsenstein, J. (2004). *Inferring Phylogenies*. Sinauer Associates.
- Grahne, G. and Zhu, J. (2003). Efficiently using prefix-trees in mining frequent itemsets. In *Proc. of FIMI '03*, CEUR Workshop Proceedings.
- Marx, V. (2013). The big challenges of big data. *Nature*, 498(7453):255–260.
- Metzker, M. L. (2010). Sequencing technologies — the next generation. *Nature Reviews Genetics*, 11(1):31–46.
- Ocaña, K. A. C. S., de Oliveira, D., et al. (2012). Discovering drug targets for neglected diseases using a pharmacophylogenomic cloud workflow. In *8th IEEE e-Science*, pages 1–8. IEEE Computer Society.
- O’Leary, N. A., Cox, E., et al. (2024). Exploring and retrieving sequence and metadata for species across the tree of life with ncbi datasets. *Scientific Data*, 11(1):732.
- Zadra, N., Rizzoli, A., and Rota-Stabelli, O. (2024). Comprehensive phylogenomic analysis of zika virus: Insights into its origin, past evolutionary dynamics, and global spread. *Virus Research*, 350:199490. Epub 2024 Nov 8.