Mineração de Subárvores Frequentes em Árvores Filogenéticas usando Conjuntos de Itens Frequentes Maximais*

João Vitor Moraes[1], Layse Gomes Castro[1], Isabel Rosseti[1], Daniel de Oliveira[1]

[1]Instituto de Computação – Universidade Federal Fluminense (UFF)

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

Resumo. *A análise filogenética permite inferir relações evolutivas entre sequências biológicas. Contudo, diferentes métodos podem gerar árvores distintas a partir do mesmo conjunto de dados, o que resulta em múltiplas hipóteses plausíveis. Abordagens como as árvores de consenso buscam sintetizar essa variabilidade, mas podem ocultar conflitos topológicos e descartar sinais evolutivos relevantes. Este artigo apresenta o* `PhyloTreeMiner`*, um workflow para mineração de subárvores frequentes como alternativa ao consenso. A proposta identifica padrões topológicos recorrentes em múltiplas árvores com base no valor de suporte. O problema é modelado como a mineração de conjuntos de itens frequentes maximais, o que reduz o custo computacional. Os resultados são armazenados no Neo4j, permitindo consultas que integram topologia e metadados.*

Abstract. *Phylogenetic analysis makes it possible to infer evolutionary relationships among biological sequences. However, different methods can produce distinct trees from the same dataset, resulting in multiple plausible hypotheses. Approaches such as consensus trees seek to summarize this variability, but they may obscure topological conflicts and discard relevant evolutionary signals. This paper presents the* `PhyloTreeMiner`*, a workflow for frequent subtree mining as an alternative to consensus trees. The proposed approach identifies recurrent topological patterns across multiple trees based on support values. The problem is modeled as maximal frequent itemset mining, reducing computational cost. The results are stored in Neo4j, enabling queries that integrate topology and metadata.*

1. Introdução

A *Análise Filogenética* [Felsenstein 2004] é uma tarefa fundamental em Bioinformática, utilizada para inferir relações evolutivas entre sequências biológicas e para apoiar análises em epidemiologia genômica [Baker et al. 2023]. Com o avanço das tecnologias de sequenciamento de DNA, tornou-se possível analisar grandes volumes de dados genômicos, o que ampliou o uso de métodos filogenéticos. Os resultados dessas análises são árvores filogenéticas [Felsenstein 2004], estruturas que representam hipóteses sobre as relações evolutivas entre organismos com base na variação observada em suas sequências. A inferência de árvores filogenéticas é sujeita ao chamado "*viés de heurística*" [Holder et al. 2008], uma vez que diferentes métodos podem produzir topologias distintas a partir do mesmo conjunto de dados de entrada. Por exemplo, abordagens que utilizam como critério de otimalidade da inferência a *Máxima Parcimônia* minimizam o número de eventos evolutivos, enquanto a *Máxima Verossimilhança* seleciona a árvore mais provável segundo um modelo evolutivo. Como consequência, não há uma única "melhor" árvore, mas sim um conjunto de hipóteses

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

plausíveis. A prática de selecionar apenas uma árvore pode, portanto, mascarar a variabilidade e a incerteza presentes no processo de inferência. Uma abordagem para mitigar esse problema é o uso de *Árvores de Consenso* [Margush and McMorris 1981], que sintetizam múltiplas topologias em uma única representação. No entanto, essa estratégia pode ocultar conflitos topológicos relevantes e descartar sinais evolutivos importantes, como eventos de recombinação ou de transferência horizontal de genes [Bryant 2003]. Além disso, a comparação entre múltiplas árvores para a identificação de padrões consensuais é um problema NP-Difícil, sendo frequentemente realizada por meio de inspeção manual ou do uso de ferramentas pouco escaláveis.

Nesse contexto, a *Mineração de Subárvores Frequentes* [Deepak et al. 2014] emerge como uma alternativa à geração de árvores de consenso. Em vez de condensar múltiplas árvores em uma única estrutura, essa abordagem identifica padrões recorrentes (*e.g.*, clados) presentes em um conjunto de árvores, considerando um limiar de suporte. Tal estratégia permite preservar informações relevantes e evidenciar relações evolutivas estáveis, mesmo na presença de divergências globais nas topologias. Apesar de seu potencial, a mineração de subárvores frequentes apresenta elevado custo computacional, o que frequentemente demanda ambientes de Computação de Alto Desempenho (HPC). Adicionalmente, a representação tradicional de dados filogenéticos em formatos como *FASTA* e *Newick* dificulta análises posteriores e limita a escalabilidade. Esses formatos carecem de suporte a consultas complexas, especialmente quando se deseja integrar informações topológicas a metadados, como dados epidemiológicos. Abordagens baseadas em bancos relacionais também se mostram pouco eficientes nesse cenário, devido ao custo de operações recursivas sobre estruturas hierárquicas [Bukhari et al. 2019].

Este artigo apresenta o `PhyloTreeMiner`, um *workflow* para a mineração eficiente de subárvores frequentes em florestas de árvores filogenéticas, com suporte a consultas analíticas. O *workflow* transforma o problema filogenético em um problema clássico de mineração de dados, utilizando técnicas de mineração de conjuntos de itens frequentes maximais, como o algoritmo FPMax [Grahne and Zhu 2005]. Ao focar em padrões maximais, reduz-se o espaço de busca e o custo computacional em comparação com abordagens que enumeram todos os padrões frequentes, que não requerem ambientes de HPC. Os resultados são armazenados em um banco de dados de grafos (Neo4j), permitindo consultas que integram a topologia e os metadados provenientes de bases de dados, como o NCBI (`https://www.ncbi.nlm.nih.gov/`).

2. Conceitos Importantes

A base das análises filogenéticas é composta por sequências biológicas, como DNA, RNA e proteínas. As sequências genéticas são representadas como cadeias de nucleotídeos, enquanto as proteínas são descritas como cadeias de aminoácidos [Setubal and Meidanis 1997]. O formato padrão para a representação desses dados é o *FASTA*, no qual cada sequência é organizada em uma linha de cabeçalho, iniciada por um identificador, seguida da representação textual da sequência biológica. Para a análise filogenética, as sequências devem ser previamente submetidas ao *Alinhamento Múltiplo de Sequências (MSA, do inglês Multiple Sequence Alignment)*. O MSA organiza as sequências de forma alinhada, geralmente em colunas, de modo a evidenciar regiões de similaridade que podem indicar ancestralidade comum, além de revelar eventos evolutivos, como mutações, inserções e deleções ao longo do tempo. Ferramentas como o MAFFT são amplamente utilizadas no MSA.

Uma vez que as sequências estejam alinhadas, pode-se iniciar a etapa de inferência e de geração de árvores filogenéticas. Nesse contexto, diferentes métodos podem ser aplicados no *workflow*, *e.g.*, *Neighbor-Joining* (NJ) e *UPGMA*, agrupam sequências a partir de matrizes de similaridade e apresentam alta eficiência computacional, sendo, portanto, adequados para grandes conjuntos de dados, embora possam simplificar determinados aspectos do processo evolutivo. O critério de otimalidade da inferência *Máxima Parcimônia* (MP) [Fitch 1971] assume que a hipótese evolutiva mais plausível é aquela que minimiza o número total de mutações; no entanto, pode apresentar limitações em cenários de evolução rápida, especialmente devido ao fenômeno de atração de ramos longos. Por sua vez, a *Máxima Verossimilhança* (ML) estima a probabilidade de que um modelo evolutivo específico tenha gerado os dados observados, permitindo a incorporação de modelos de substituição mais complexos.

Tradicionalmente, o resultado de múltiplas inferências, *i.e.*, múltiplas árvores filogenéticas obtidas por diferentes métodos, é sintetizado por meio de uma Árvore de Consenso [Margush and McMorris 1981]. Esse tipo de abordagem busca representar, em uma única topologia, as relações evolutivas mais frequentes ou melhor suportadas entre as árvores analisadas, geralmente por meio de critérios como o consenso estrito, a maioria (*majority-rule*) ou o consenso ponderado. Embora útil para a sumarização, essa técnica tende a suprimir conflitos topológicos relevantes, especialmente quando há discordâncias consideráveis entre as árvores de entrada. Como consequência, sinais evolutivos importantes, *e.g.*, eventos de recombinação, podem ser tratados como variações residuais ou como "ruído" em vez de serem representados explicitamente.

A *Mineração de Subárvores Frequentes* (FSM, do inglês *Frequent Subtree Mining*) [Guedes et al. 2017] propõe uma mudança de paradigma: em vez de sintetizar os resultados em uma "árvore média", busca-se identificar subestruturas recorrentes, *i.e.*, clados que aparecem de forma consistente em diferentes árvores inferidas. Essa abordagem explora a recorrência de padrões topológicos como evidência evolutiva, considerando a variabilidade entre métodos e modelos. Diferentemente dos valores de *bootstrap*, que estimam a confiança de clados restritos à topologia gerada por um único método (*e.g.*, Máxima Verossimilhança), a mineração de padrões extrai um consenso estrutural global cruzando simultaneamente métodos com diferentes vieses matemáticos, evidenciando relações que sobrevivem a essas divergências algorítmicas, o que um valor de *bootstrap* isolado não é capaz de capturar. A próxima seção apresenta o *workflow* `PhyloTreeMiner` para realizar a mineração das subárvores frequentes.

3. O *Workflow* `PhyloTreeMiner`

A Figura 1 apresenta o *workflow* `PhyloTreeMiner`, que integra inferência filogenética, mineração de padrões e modelagem orientada a grafos. Cada atividade do *workflow* executa uma transformação específica e está agrupada em uma macro-atividade. O *workflow* inicia com a macroatividade de *Coleta e Pré-processamento*, na qual os arquivos FASTA são processados por meio de rotinas do *Biopython*. Nessa fase, realizam-se a validação sintática, a remoção de duplicatas e a aplicação de filtros de qualidade definidos pelo usuário, como limites de tamanho de sequências e a exclusão de regiões ambíguas, garantindo um conjunto de dados consistente e confiável. Em seguida, ocorre a macroatividade de *Alinhamento Múltiplo de Sequências e Inferência de Árvores Filogenéticas*, etapa central da análise filogenética. Diferentes algoritmos podem ser utilizados (*e.g.*, MAFFT e ClustalW), permitindo explorar variações metodológicas. A partir dos alinhamentos (MSA), inferem-se múltiplas árvores por

abordagens complementares, como *Neighbor Joining* (NJ), Máxima Verossimilhança (ML), com ferramentas como IQ-TREE, RAxML e FastTree, e Inferência Bayesiana (BI), com MrBayes. Essa diversidade gera um conjunto de topologias alternativas, formando uma "floresta" filogenética que abrange diferentes hipóteses evolutivas a partir do mesmo *dataset*.

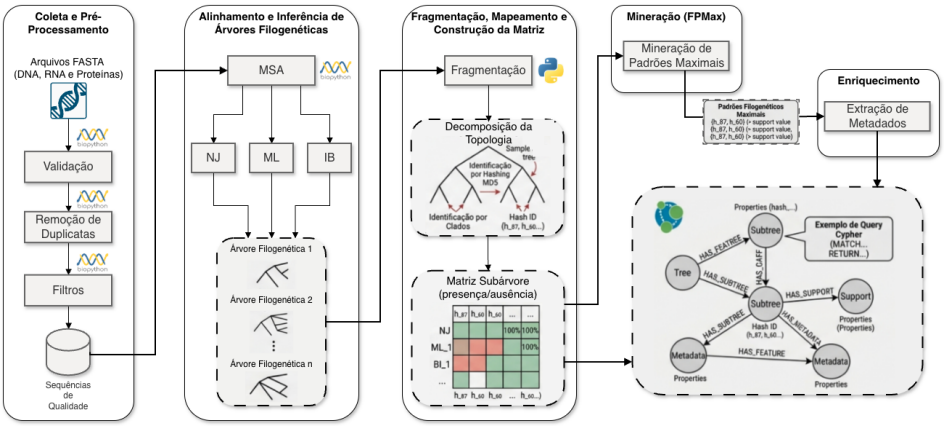

Figura 1. Especificação do *Workflow* `PhyloTreeMiner`

Uma vez obtidas as árvores, o *workflow* avança para a macroatividade de *Fragmentação, Mapeamento e Construção da Matriz*, núcleo da transformação dos dados para mineração. Cada árvore é decomposta em subárvores (*e.g.*, clados) por meio da travessia da topologia, gerando múltiplas subárvores. Essa etapa segue o modelo *bag-of-tasks*, permitindo a paralelização da extração entre os núcleos do processador. Para viabilizar a mineração, cada subárvore é convertida em um identificador único por meio de *hashing*. Assim, clados estruturalmente idênticos, mesmo provenientes de árvores distintas, recebem o mesmo identificador, o que permite mapear estruturas complexas em um espaço discreto de itens. Essa abstração permite tratar a coleção como um banco de dados transacional: cada árvore é uma transação e cada subárvore, um item. Com isso, constrói-se a *matriz Subárvore–Árvore*, uma matriz de incidência em que as linhas representam as árvores (obtidas por diferentes heurísticas) e as colunas, os identificadores das subárvores. Os valores indicam a presença ou a ausência de cada clado, servindo de base para a mineração de padrões.

Com a matriz construída, o *workflow* executa a macroatividade de *Mineração de Padrões Frequentes* com o algoritmo FPMax. Diferentemente de abordagens que enumeram todos os padrões, o FPMax identifica diretamente os conjuntos maximais (*i.e.*, não contidos em outros conjuntos frequentes), reduzindo a redundância e melhorando a eficiência, sobretudo em cenários de alta dimensionalidade (*i.e.*, grande número de subárvores). Cada padrão corresponde a um conjunto de subárvores que coocorrem em múltiplas árvores filogenéticas, e seu suporte indica a proporção de árvores em que ocorre simultaneamente. Padrões com alto suporte refletem relações filogenéticas consistentes entre diferentes métodos de inferência e são interpretados como indícios de robustez evolutiva. Assim, a comparação de árvores se transforma no problema de descoberta de *itemsets* frequentes. Após identificar os padrões, o *workflow* executa a macroatividade de *Enriquecimento Semântico*, integrando os resultados aos metadados de bases de dados externas, como a do NCBI. A partir dos identificadores das sequências, são recuperadas informações como taxonomia, localização geográfica, data de coleta e referências bibliográficas. Isso amplia o contexto dos padrões, permitindo análises que combinam estrutura evolutiva com evidências biológicas e epidemiológicas. Por fim, todos os artefatos gerados (*e.g.*, árvores, subárvores, padrões

e metadados) são organizados em um *Grafo de Conhecimento* no Neo4j. Nessa estrutura, nós representam entidades como *Tree*, *Subtree*, *Support* e *Metadata*, enquanto as arestas indicam relações como `HAS_SUBTREE`, `HAS_SUPPORT` e `HAS_METADATA`. Essa modelagem permite consultas complexas em *Cypher*, facilitando a exploração de relações filogenéticas, a análise de suporte e a integração com dados externos. Código-fonte disponível em `https://github.com/UFFeScience/phylotreeminer`.

4. Avaliação Experimental

Nesta seção, avaliamos o `PhyloTreeMiner` por meio de um estudo de caso focado na análise do vírus Zika (ZIKV) no contexto de vigilância genômica. [Zadra et al. 2024] disponibilizam um conjunto de dados composto por variantes do vírus Zika obtidas do GenBank, totalizando 479 sequências genômicas, cada uma com aproximadamente 10.000 nucleotídeos. Esse conjunto apresenta ampla diversidade geográfica, incluindo amostras da África, da Ásia, das Ilhas do Pacífico e das Américas, o que permite capturar diferentes linhagens evolutivas do vírus Zika. Todas as sequências foram previamente submetidas a etapas de curadoria, incluindo a remoção de regiões altamente variáveis e a exclusão de sequências com evidências de recombinação, o que garante maior robustez às análises filogenéticas subsequentes. Esse conjunto foi utilizado como entrada para o `PhyloTreeMiner`, a partir do qual foram inferidas 16 árvores filogenéticas (Figura 2), considerando diferentes combinações de métodos de MSA e de algoritmos de inferência filogenética. Sobre esse conjunto de árvores, foi aplicada a etapa de mineração do *workflow*, com o objetivo de identificar subárvores frequentes que representem relações evolutivas robustas e consistentes entre diferentes metodologias.

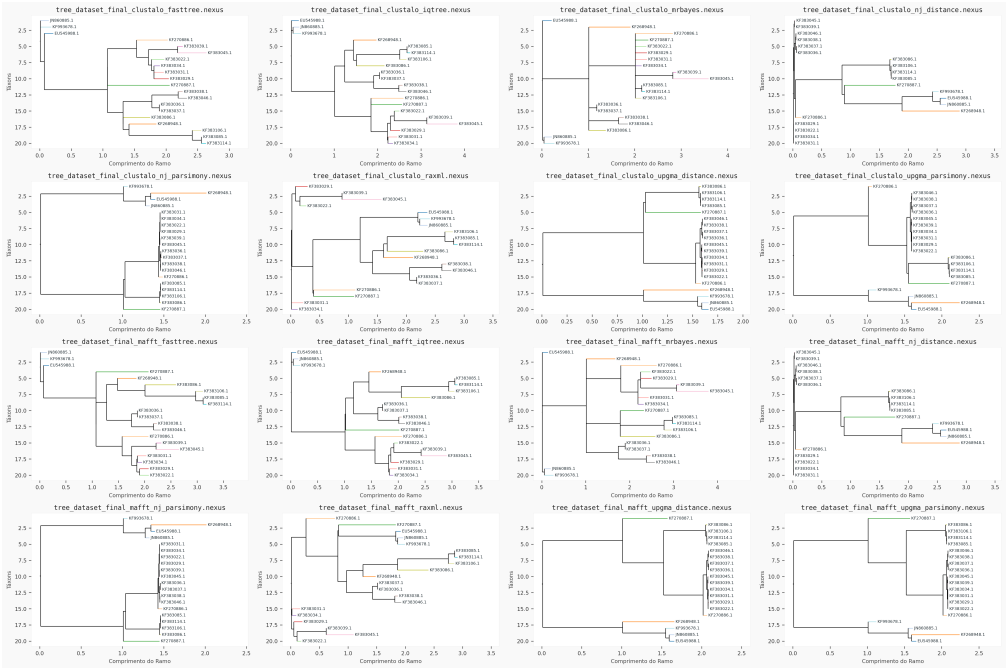

Figura 2. Conjunto de 16 árvores filogenéticas geradas por diferentes métodos.

Conforme pode ser observado na Figura 2, que apresenta fragmentos das 16 árvores geradas, as topologias variam significativamente em função do método de alinhamento e da abordagem de inferência filogenética adotada. Essa variabilidade torna esse conjunto de dados particularmente adequado como estudo de caso para a avaliação do `PhyloTreeMiner`,

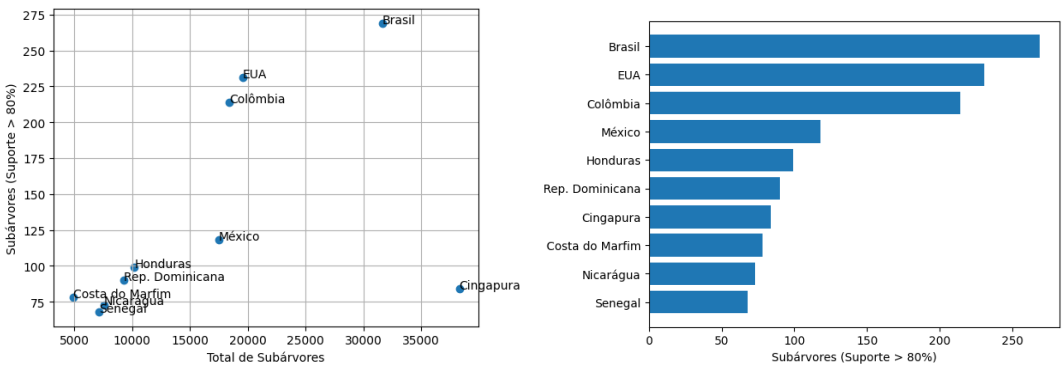

Figura 3. (a) Relação do total de subárvores com as que possuem suporte maior que 80% (b) Quantidade de subárvores com suporte maior que 80%.

pois expõe o *workflow* a diferentes hipóteses evolutivas derivadas do mesmo conjunto de sequências. Para avaliar o *workflow* no que se refere à capacidade de identificação de padrões frequentes, analisamos a distribuição geográfica das subárvores com suporte acima de um *treshold* com base em metadados recuperados do NCBI. Foram selecionadas subárvores com suporte superior a 80% por meio de consultas ao grafo de conhecimento filogenético. A partir dessa análise (Tabela 1 e Figura 3), observa-se que, para o Brasil, dentre um total de 31.633 subárvores extraídas, apenas 269 (aproximadamente 0,85%) apresentam suporte $\geq 80\%$.

Esse resultado evidencia a eficácia do `PhyloTreeMiner` como mecanismo de filtragem seletiva, capaz de reduzir a redundância e o ruído topológico, preservando apenas o núcleo das relações filogenéticas mais consistentes entre diferentes metodologias. Em outras palavras, o *workflow* descarta milhares de padrões pouco informativos para destacar aqueles com maior evidência de recorrência. A Figura 4 apresenta as *top* 3 subárvores com os maiores valores de suporte. Adicionalmente, observa-se uma alta densidade amostral em Cingapura, com 38.302 subárvores totais, o que corresponde exatamente ao que foi reportado em [Zadra et al. 2024], possivelmente associada à disseminação da linhagem asiática do ZIKV durante o surto de 2016. No entanto, esse *hotspot* apresenta o menor suporte médio no conjunto de dados (11,7%), apesar de concentrar o maior número de subárvores, e também é acompanhado por uma elevada frequência de variações topológicas. Esse cenário, identificado de forma automatizada pelo *workflow*, sugere maior complexidade evolutiva, possivelmente associada a eventos de recombinação ou a múltiplas introduções virais independentes.

Tabela 1. Subárvores com suporte $\geq$ 80% agrupadas por país de ocorrência.

País	Total de Subárvores	Subárvores com suporte $\geq$80%	Suporte Médio	Suporte Máximo
Brasil	31.633	269	0,134	0,8
EUA	19.553	231	0,137	0,8
Colômbia	18.390	214	0,132	0,8
México	17.475	118	0,131	0,9
Honduras	10.195	99	0,141	0,8
Rep. Dominicana	9.274	90	0,133	0,8
Cingapura	38.302	84	0,117	0,8
Costa do Marfim	4.875	78	0,154	0,8
Nicarágua	7.536	73	0,140	0,8
Senegal	7.113	68	0,144	0,8

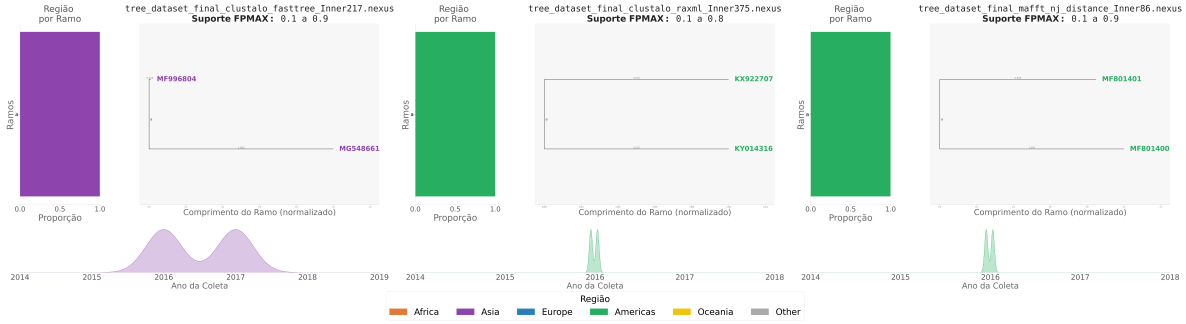

Figura 4. *Top* 3 subárvores com maior suporte.

5. Trabalhos Relacionados

Existem trabalhos na literatura que abordam o problema de mineração de subárvores. O SciPhyloMiner [Guedes et al. 2017] propõe um *workflow* composto por três etapas (geração, mineração e análise) voltado à comparação de árvores de genes ortólogos em protozoários. No entanto, a abordagem adotada pelo SciPhyloMiner apresenta limitações de escalabilidade, uma vez que se baseia na comparação exaustiva de subárvores, o que resulta em elevado custo computacional à medida que o número de árvores aumenta. Em contraste, neste trabalho, modelamos o problema como uma tarefa de mineração de conjuntos de itens frequentes maximais, reduzindo o espaço de busca e tornando a abordagem mais escalável. O EvoMiner [Deepak et al. 2014] é uma abordagem para a mineração de subárvores em bases filogenéticas, permitindo a identificação de padrões topológicos recorrentes. No entanto, ele apresenta limitações importantes. Uma primeira limitação é que a mineração é realizada diretamente sobre estruturas de árvores, exigindo operações de comparação estrutural custosas, o que compromete sua escalabilidade. Além disso, o espaço de busca cresce rapidamente à medida que aumenta o número de subárvores possíveis. Por fim, a abordagem não contempla a integração com metadados nem oferece suporte a análises exploratórias mais avançadas, o que limita sua aplicabilidade em cenários que demandam interpretações mais complexas.

O algoritmo IncGM+ [Abdelhamid et al. 2017] introduz uma abordagem incremental baseada no conceito de *fringe* (fronteira entre padrões frequentes e infrequentes), que permite restringir o espaço de busca e melhorar a eficiência da mineração, alcançando ganhos de desempenho de até três ordens de magnitude em comparação com métodos estáticos. Mais recentemente, [Bostanoğlu and Abuzayed 2024] propõem uma categorização dos algoritmos de *Dynamic Frequent Subgraph Mining*, destacando que a crescente complexidade e o volume dos dados atuais demandam abordagens capazes de lidar com atualizações contínuas, sem a necessidade de reprocessar integralmente o conjunto de dados. Diferentemente dessas abordagens, que se concentram em cenários dinâmicos de grafos gerais, este artigo foca no domínio específico de árvores filogenéticas, explorando uma modelagem alternativa baseada na mineração de conjuntos de itens frequentes maximais.

6. Conclusão

Este artigo apresentou o `PhyloTreeMiner`, um *workflow* para mineração de subárvores frequentes como alternativa às abordagens tradicionais baseadas em árvores de consenso. Ao modelar o problema de comparação entre múltiplas árvores filogenéticas como uma tarefa de mineração de conjuntos de itens frequentes maximais, foi possível identificar padrões topológicos recorrentes de forma eficiente, preservando sinais evolutivos relevantes frequentemente ocultados por técnicas de sumarização. A avaliação experimental, conduzida com

dados reais do vírus Zika, demonstraram que o *workflow* é capaz de filtrar grandes volumes de subárvores, destacando apenas aquelas com alto suporte e maior consistência entre diferentes métodos de inferência. Essa capacidade evidencia o potencial da abordagem como ferramenta para análise filogenética robusta, especialmente em cenários com elevada variabilidade topológica e múltiplas hipóteses evolutivas. Além disso, a integração com banco de dados orientado a grafos permite a exploração dos resultados de forma mais simples, combinando informações topológicas com metadados biológicos e epidemiológicos. Essa característica amplia as possibilidades analíticas, tornando o `PhyloTreeMiner` uma solução para aplicações em vigilância genômica e estudos evolutivos em larga escala. Como trabalhos futuros, pretende-se investigar estratégias para otimização adicional do processo de mineração, bem como explorar a aplicação do *workflow* em outros domínios e conjuntos de dados, incluindo cenários de análise comparativa mais ampla.

Referências

Abdelhamid, E., Canim, M., et al. (2017). Incremental frequent subgraph mining on large evolving graphs. *IEEE Trans. on Know. and Data Eng.*, 29(12):2710–2723.

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.

Bostanoğlu, B. E. and Abuzayed, N. (2024). Dynamic frequent subgraph mining algorithms over evolving graphs: a survey. *PeerJ Computer Science*, 10:e2361.

Bryant, D. (2003). A classification of consensus methods for phylogenetics. *DIMACS series in discrete mathematics and theoretical computer science*, 61:163–184.

Bukhari, S. A. C., Mandell, J., et al. (2019). A linked data graph approach to integration of immunological data. *Proc. of the IEEE BIBM*.

Deepak, A., Fernández-Baca, D., et al. (2014). Evominer: frequent subtree mining in phylogenetic databases. *Knowl. Inf. Syst.*, 41(3):559–590.

Felsenstein, J. (2004). *Inferring Phylogenies*. Sinauer Associates.

Fitch, W. M. (1971). Toward defining the course of evolution: Minimum change for a specific tree topology. *Systematic Zoology*, 20(4):406–416.

Grahne, G. and Zhu, J. (2005). Fast algorithms for frequent itemset mining using fp-trees. *IEEE Transactions on Knowledge and Data Engineering*, 17(10):1347–1362.

Guedes, T., Ocaña, K., et al. (2017). Sciphylominer: um workflow para mineração de dados filogemônicos de protozoários. In *XI BreSci*, pages 69–76, São Paulo, Brasil. SBC.

Holder, M. T., Zwickl, D. J., et al. (2008). Evaluating the robustness of phylogenetic methods to among-site variability in substitution processes. *Phil. Trans. of the Royal Society B: Biological Sciences*, 363(1512):4013–4021.

Margush, T. and McMorris, F. (1981). Consensusn-trees. *Bulletin of Mathematical Biology*, 43(2):239–244.

Setubal, J. C. and Meidanis, J. (1997). *Introduction to computational molecular biology*. PWS Publishing Company.

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.