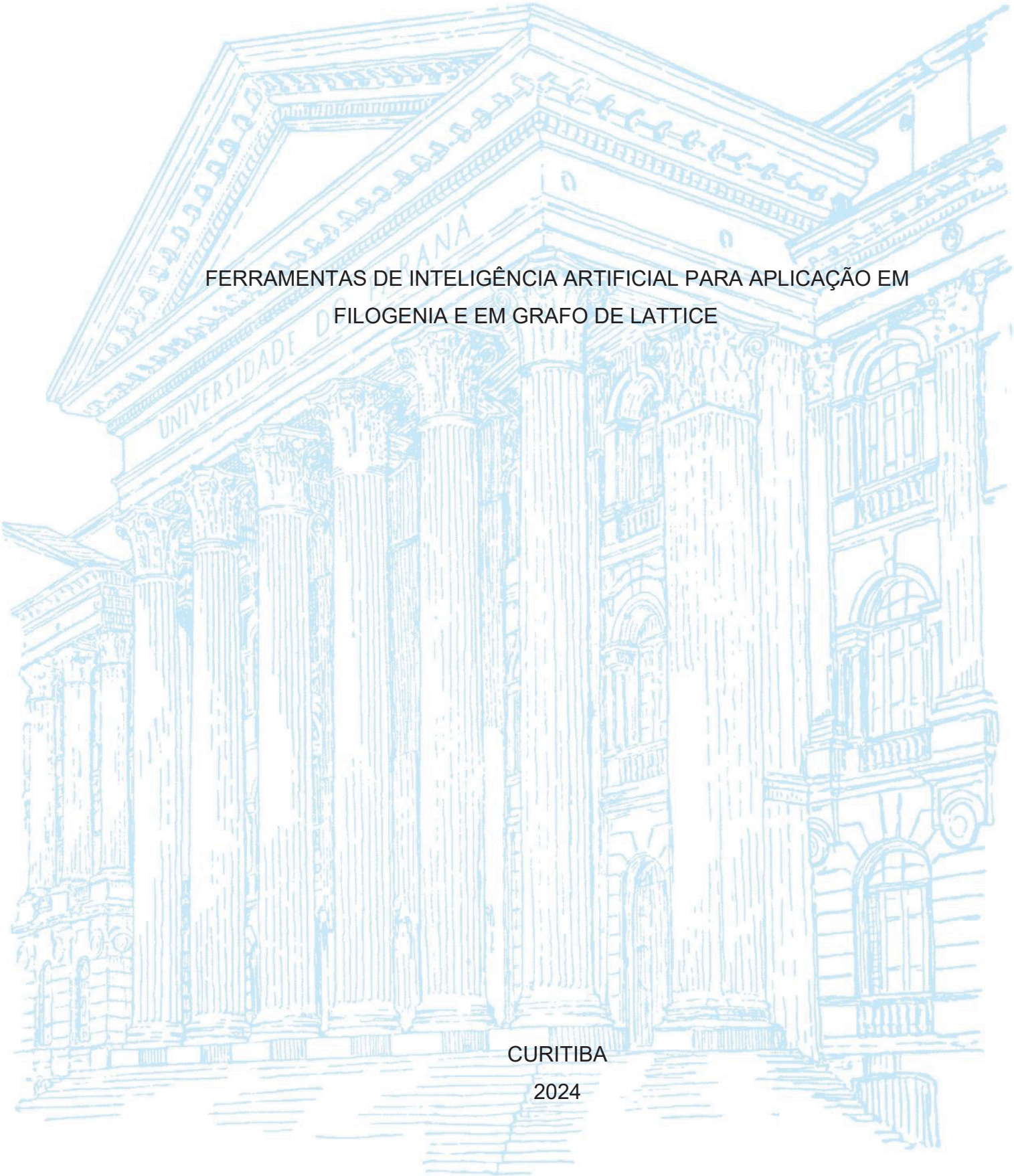
UNIVERSIDADE FEDERAL DO PARANÁ

DANRLEY RAFAEL FERNANDES



FERRAMENTAS DE INTELIGÊNCIA ARTIFICIAL PARA APLICAÇÃO EM FILOGENIA E EM GRAFO DE LATTICE

CURITIBA

2024

DANRLEY RAFAEL FERNANDES

FERRAMENTAS DE INTELIGÊNCIA ARTIFICIAL PARA APLICAÇÃO EM FILOGENIA E EM GRAFO DE LATTICE

Dissertação apresentada ao curso de Pós-Graduação em bioinformática, Setor de Educação Profissional e Tecnológica, Universidade Federal do Paraná, como requisito parcial à obtenção do título de mestre em bioinformática.

Orientador: Prof. Dr. Roberto Raittz

CURITIBA

2024

Catalogação na publicação
Sistema de Bibliotecas UFPR

F363

Fernandes, Danrley Rafael
Ferramentas de Inteligência Artificial para Aplicação em Filogenia e em Grafo de Lattice / Danrley Rafael Fernandes. - Curitiba, 2025.
1 recurso on-line : PDF.

Dissertação (Mestrado) – Universidade Federal do Paraná, Setor de Educação Profissional e Tecnológica, Programa de Pós-Graduação em Bioinformática, 2025.

Orientador: Dr. Roberto Raittz

1. Análise de sequências. 2. Filogenia. 3. Análise de sequências livre de alinhamento. 4. Algorítmos genéticos. 5. Problema do caminho mais curto. I. Raittz, Roberto. II. Fernandes, Danrley Rafael. III. Universidade Federal do Paraná. IV. Título.

Bibliotecário: André A. Oliveira — CRB 9/2019

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação BIOINFORMÁTICA da Universidade Federal do Paraná foram convocados para realizar a arguição da Dissertação de Mestrado de **DANRLEY RAFAEL FERNANDES**, intitulada: **"Ferramentas de Inteligência Artificial para aplicação em Filogenia e em Grafos de *Lattice*"**, sob orientação do Prof. Dr. ROBERTO TADEU RAITTZ, que após terem inquirido o aluno 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, 14 de Junho de 2024.

Assinatura Eletrônica
05/09/2025 15:46:40.0
ROBERTO TADEU RAITTZ
Presidente da Banca Examinadora

Assinatura Eletrônica
05/09/2025 16:44:01.0
JERONIZA NUNES MARCHAUKOSKI
Avaliador Interno (UNIVERSIDADE FEDERAL DO PARANÁ)

Assinatura Eletrônica
09/09/2025 13:28:09.0
EDUARDO TIEPPO
Avaliador Externo (UNIVERSIDADE FEDERAL DO PARANÁ)

Dedico esse trabalho aos meus grupos de amigos queridos que, de alguma maneira, me encorajaram ao longo do Mestrado. Realmente, eu tirei a sorte grande.

AGRADECIMENTOS

Gostaria de expressar minha profunda gratidão ao meu orientador, Roberto, por sua orientação, paciência e apoio contínuo ao longo deste projeto. Sua expertise e dedicação foram fundamentais para o desenvolvimento desta dissertação.

Agradeço de forma especial à minha esposa, Bruna, pelo seu amor, compreensão e paciência ao longo desta jornada. Sua presença constante e apoio incondicional foram fundamentais para que eu pudesse superar os desafios e continuar perseverando. Sem o seu incentivo e sacrifício, este trabalho não teria sido possível.

Agradeço imensamente aos meus amigos, que me incentivaram e apoiaram em cada etapa deste trabalho. Suas palavras de encorajamento e companhia nos momentos difíceis foram essenciais para a conclusão deste estudo.

Meu sincero agradecimento à empresa Lactec e aos órgãos de fomento à pesquisa, cujo suporte foi indispensável para a realização deste projeto. O financiamento e os recursos proporcionados permitiram que esta dissertação se concretizasse.

“Nada de lamúrias, camaradas, haveremos de vencer, porque temos razão”.
(LÊNIN V. I.)

RESUMO

Esta dissertação apresenta as ferramentas rSWeeP e R3TO como inovações no campo da bioinformática e otimização de redes. O rSWeeP é uma implementação em R do método SWeeP, desenvolvida para popularizar esta metodologia no campo da bioinformática. Ao utilizar o SWeeP para representar sequências biológicas em vetores de baixa dimensão, o rSWeeP facilita análises rápidas e precisas de grandes volumes de dados genômicos em uma das plataformas mais populares na bioinformática: a linguagem R. O R3TO é introduzido como uma alternativa eficiente aos métodos tradicionais de cálculo de distância, como o algoritmo de Dijkstra, demonstrando superioridade ao lidar com redes de distribuição elétrica em áreas rurais. Esta pesquisa ainda propõe um uso sinérgico das capacidades do rSWeeP e do R3TO, que pode resultar em ganhos significativos de precisão e eficiência. O estudo, portanto, evidencia o potencial de ambas as ferramentas, isoladas e em conjunto, para impulsionar avanços na análise filogenética, ampliando as possibilidades de aplicações na área de bioinformática.

Palavras-chave: Análise de sequências biológicas; Filogenia; Técnicas livres de alinhamento; Algoritmo Genético; Shortest Path

ABSTRACT

This dissertation presents the rSWeeP and R3TO tools as innovations in the field of bioinformatics and network optimization. rSWeeP is an R implementation of the SWeeP method, developed to popularize this methodology in the field of bioinformatics. By using SWeeP to represent biological sequences in low-dimensional vectors, rSWeeP facilitates fast and accurate analysis of large volumes of genomic data in one of the most popular platforms in bioinformatics: the R language. R3TO is introduced as an efficient alternative to traditional distance calculation methods, such as Dijkstra's algorithm, demonstrating superiority when dealing with electrical distribution networks in rural areas. This research also proposes a synergistic use of the capabilities of rSWeeP and R3TO, which can result in significant gains in accuracy and efficiency. The study, therefore, highlights the potential of both tools, alone and together, to drive advances in phylogenetic analysis, expanding the possibilities of applications in the field of bioinformatics.

Keywords: Analysis of biological sequences; Phylogeny; Alignment-free techniques; Genetic Algorithm; Shortest Path

LISTA DE FIGURAS

SUMÁRIO

1. INTRODUÇÃO

A aplicação da Inteligência Artificial (IA) na bioinformática tem revolucionado o estudo da filogenia, permitindo análises mais precisas e rápidas das relações evolutivas entre espécies por meio de técnicas avançadas, como redes neurais e algoritmos de classificação, que superam os métodos tradicionais em termos de eficiência e acurácia (Karim et al. 2023). Além disso, a IA tem sido fundamental na análise de grandes volumes de dados genômicos e transcriptômicos, possibilitando a identificação de genes essenciais e a construção de redes de interação gênica complexas, contribuindo para um entendimento mais profundo da biologia molecular e da biodiversidade (Zou et al. 2024). Fora da filogenia, a IA também desempenha um papel crucial na predição de estruturas proteicas, essencial para o desenvolvimento de novos medicamentos e tratamentos personalizados, além de ser aplicada na agricultura de precisão para melhorar o rendimento das culturas por meio da análise de dados fenotípicos e genéticos (Sharma et al. 2022). Portanto, a IA demonstra seu vasto potencial não só no estudo da evolução das espécies, mas também em outras áreas da ciência, promovendo avanços significativos na medicina e na agricultura moderna (Karim et al. 2023).

Das ferramentas de IA com grandes vantagens para o estudo da filogenia, o SWeeP se destaca por sua capacidade de representar sequências biológicas em vetores numéricos de baixa dimensão, facilitando análises filogenéticas sem a necessidade de alinhamento, o que reduz significativamente o tempo de processamento e permite a análise de grandes volumes de dados biológicos (Perico et al. 2024). O método original, desenvolvido por De Pierri et al. (2020), utiliza uma projeção de k-mers em vetores de alta dimensão que são reduzidos a vetores de baixa dimensão, preservando as distâncias evolutivas entre as sequências (De Pierri et al. 2020). Isso possibilita estudos mais robustos e rápidos em comparação com técnicas tradicionais de construção de árvores filogenéticas, que muitas vezes exigem complexos alinhamentos múltiplos de sequências (Perico et al., 2024). Apesar dessas vantagens, a implementação do SWeeP ainda é limitada por estar disponível apenas em ambientes específicos, restringindo seu acesso e potencial de integração com outras plataformas de bioinformática.

A Implantação rSWeeP se apresenta como uma solução para essa lacuna de uso do conhecimento, sendo um pacote em linguagem R e disponível no repositório

Bioconductor. A linguagem de programação R e o repositório Bioconductor têm experimentado um crescimento significativo no campo da bioinformática nos últimos anos. Estudos recentes destacam a ampla adoção dessas ferramentas para análises genômicas e transcriptômicas. Por exemplo, uma revisão publicada em 2023 enfatiza o papel fundamental do Bioconductor na análise de dados genéticos, destacando sua capacidade de integrar e interpretar grandes volumes de informações biológicas (SILVA; Alves. 2023).

Neste contexto podemos inferir que a aplicação da inteligência artificial (IA) tem se mostrado fundamental na área de filogenética, com aplicações técnicas livres de alinhamento em combinação com algoritmos de cálculo de distância. Estes métodos permitem a análise de grandes volumes de dados genômicos, facilitando a construção de árvores evolutivas com maior precisão e eficiência (Trindade; Oliveira. 2024). No contexto algoritmos de cálculo de distância, os grafos de lattice são frequentemente empregados para resolver problemas complexos de roteamento e otimização, como o traçado de redes elétricas em áreas rurais, considerando obstáculos naturais e minimizando custos (Brandão. 2020). Além disto a aplicação da teoria dos grafos na filogenética tem se mostrado fundamental para a compreensão das relações evolutivas entre espécies. Estudos recentes destacam o uso de redes filogenéticas, que permitem representar eventos complexos como hibridização e transferência horizontal de genes, oferecendo uma visão mais abrangente da evolução biológica. Por exemplo, Huber et al. (2023) introduziram o conceito de "shared ancestry graphs" para modelar ancestralidades compartilhadas de forma mais precisa, contribuindo para uma representação mais detalhada das relações evolutivas.

Nesta dissertação buscamos apresentar duas contribuições relevantes para esse contexto: o rSWeeP, uma versão em software livre do SWeeP, que facilita a representação de sequências biológicas por meio de vetores, e o R3TO, uma solução para o problema de traçado ótimo em grafos de lattice, testada no mapeamento de fios de energia em áreas rurais. Essas ferramentas visam contribuir para o aprimoramento da acurácia das inferências filogenéticas e a eficiência dos sistemas de roteamento de redes, conforme discutido nos capítulos 2 e 3. O texto apresenta suas contribuições em dois capítulos principais: o capítulo RSWeeP: Um Pacote R/Bioconductor Para Representação De Sequências Sweep e o capítulo 3 R3TO: Algoritmo Para Roteamento De Redes Elétricas Com Alocação De Postes e Transposição De Obstáculos. Cada um destes capítulos contêm o artigo que a

apresenta as ferramentas desenvolvidas neste projeto. E no último capítulo é discutido as vantagens e aplicações já realizadas dessas ferramentas, além de sugerir trabalhos futuros.

1.1 OBJETIVOS

1.1.1 Objetivo geral

Apresentar a ferramenta rSWeeP como uma plataforma popularização de técnicas livres de alinhamento na bioinformática, principalmente o método SWeeP. E a ferramenta R3TO como uma técnica possível para auxiliar esse tipo de análise de sequencias biológicas

1.1.2 Objetivos específicos

- Demonstrar as contribuições da implementação rSWeeP para a popularização do método SWeeP.
- Demonstrar as vantagens do método R3TO para otimização de distancias quando comparado aos algo algoritmos tradicionais.
- Aplicar o R3TO em cenários práticos para validar sua eficiência em termos de economia de recursos e melhorias no planejamento de redes elétricas em áreas rurais.
- Comparar o desempenho e os resultados do rSWeeP e do R3TO com métodos tradicionais e disponíveis, demonstrando as vantagens das abordagens de inteligência artificial desenvolvidas em ambos os contexto.

2 RSWEEP: UM PACOTE R/BIOCONDUCTOR PARA REPRESENTAÇÃO DE

rSWeeP: Um pacote R/Bioconductor para representação de sequências SWeeP

Danrley Fernandes[1,2], Mariane G. Kulik[1,2], Diogo J. S. Machado[1,2], Jeroniza N. Marchaukoski[1,2], Fabio O. Pedrosa[2,3], Camilla R. De Pierri[1,3] e Roberto T. Raittz[1,2*]

[1]Laboratório de Inteligência Artificial aplicada à Bioinformática, Universidade Federal do Paraná, Curitiba, Paraná, Brasil. [2]Programa de Pós-Graduação em Bioinformática, Universidade Federal do Paraná, Curitiba, Paraná, Brasil. 3Departamento de Bioquímica e Biologia Molecular, Universidade Federal do Paraná, Curitiba, Paraná, Brasil

Resumo

O pacote rSWeeP é uma implementação R do modelo SWeeP, projetado para lidar com Big Data. O rSweeP atende à crescente demanda por métodos eficientes de representação heurística na área de Bioinformática, em plataformas acessíveis a toda a comunidade científica. Exploramos a implementação do rSWeeP usando um conjunto de dados contendo 31.386 proteomas virais, realizando análises filogenéticas e de componentes principais. Como estudo de caso analisamos as cepas virais mais próximas do SARS-CoV, responsável pela atual pandemia de COVID-19, confirmando que o rSWeeP pode classificar com precisão os organismos taxonomicamente. O pacote rSWeeP está disponível gratuitamente em https://bioconductor.org/packages/release/bioc/html/rSWeeP.html.

Palavras-chave: Inteligência Artificial, Bioinformática, Mineração de Dados

1 Introdução

Na era do Big Data, abordagens livres de vetores e alinhamento para comparar e representar sequências biológicas se destacam por serem mais eficientes do que a maioria dos métodos heurísticos baseados em alinhamento [1]. Estudos mostram que a representação vetorial de dados biológicos apresenta uma solução eficaz para a área de Bioinformática [2–5].

Recentemente, o modelo SWeeP apresentou resultados expressivos na análise de proteomas completos. SWeeP é um modelo implementado em MatLab, que permite a representação de informação biológica através da projeção de sequências de DNA em vetores com dimensionalidade reduzida [5]. Aqui apresentamos o rSWeeP, uma implementação do modelo SWeeP usando a linguagem de programação R/Bioconductor. Para testar a eficácia da implementação, executamos um teste de desempenho, análise filogenética e análise de componentes principais (PCA) usando proteomas virais como conjunto de dados. Nossos resultados mostraram que o rSWeeP mantém a eficácia do modelo SWeeP na comparação de um grande número de sequências. No estudo de caso, identificamos a proximidade entre o vírus responsável pela atual pandemia de SARS (Síndrome Respiratória Aguda Grave) (COVID19) e o vírus responsável pela pandemia de SARS em 2003, sugerindo que se trata do mesmo coronavírus, conforme já relatado em outro estudo [6].

2 Método

Todas as etapas de implementação do rSWeeP foram realizadas em um processador Intel core I5 320Gz com 12 GB de RAM.

Implementação. A entrada rSWeeP é um arquivo multiFASTA ou um objeto de classe "AAStringSet", contendo sequências de aminoácidos. rSWeeP consiste em duas funções principais:

(1) "OrthBase": para gerar uma matriz ortonormal em tamanho especificado;

(2) "SWeeP": gerar e projetar vetores SWeeP para comparar as informações das sequências.

Conjunto de dados. Usamos todas as sequências de vírus na categoria "genoma completo" do NCBI na montagem, disponível em https://ftp.ncbi.nlm.nih.gov/genomes/Viruses/ (baixado em 6 de fevereiro de 2020). Um tutorial para executar o rSWeeP e as árvores está disponível em https://github.com/DanrleyRF/Suplementar. Para todas as análises, aplicamos o mesmo protocolo adotado no estudo de De Pierri e colegas (2020) [5].

3 Resultado e Discussão

Construção de Árvores Filogenéticas. Para explorar o potencial do pacote rSWeeP, criamos vetores SWeeP a partir das anotações proteicas de todos os genomas virais completos disponíveis no NCBI, no total, 31.386 sequências de proteomas (31k). Uma árvore filogenética viral foi construída quase dez vezes maior que a referência encontrada na literatura [3].

Existem muitos espécimes para cada espécie de vírus. Assim, para uma melhor compreensão da distribuição taxonômica viral, utilizando os mesmos vetores, filtramos os dados, selecionando espécies de vírus classificadas como "exemplos do ICTV (International Committee on Viral Taxonomy)" [7]. uma amostra de vírus por espécie, totalizando 4.833 proteomas virais (4k).

Teste de performance. A árvore de 31k foi gerada em 2 horas e 17 minutos, e a árvore de 4k levou 13 minutos. O rSWeeP apresentou uma curva de crescimento linear no tempo de processamento, como já era esperado, de cerca de 500 sequências comparadas por minuto. Assim, as funções apresentadas pelo pacote rSWeeP podem dar poder computacional aos seus usuários para gerenciar variáveis biológicas de big data em um tempo prático.

Análise de conjunto de dados global. Traçamos os dois componentes principais para os dois conjuntos de dados (31k e 4k). Observamos que os dados apresentaram o mesmo padrão de agrupamento em relação ao tipo de organização dos ácidos nucléicos (fig. 1). É

evidente a separação entre os três principais clados de RNA de fita simples (ssRNA), DNA de fita simples (ssDNA) e DNA de fita dupla (dsDNA) (de acordo com ICTV). Observamos também alguns casos de incompatibilidade nos genomas, principalmente no cluster ssDNA. Essas divergências ocorrem por vírus com genótipos que não estão inseridos em nenhum clado. Além disso, existem alguns casos de erros de taxonomia e anotação na base de dados NCBI, também identificados no estudo de Calisher e colegas (2006) [8], e muitas dessas anotações não foram corrigidas até à data. Isso mostra que o rSWeeP apresenta uma solução para classificar com precisão os organismos taxonomicamente.

Análise de SARS-CoV. Segundo Gorbalenya e colegas (2020) [6], os critérios adotados para definir o coronavírus do surto atual, como um "novo" coronavírus, não seriam o ideal, uma vez que se trata do mesmo coronavírus relatado por Drostren C. e colegas (2003) [9]. Segundo o ICTV, as novas nomenclaturas são SARS-CoV-1 – para o vírus isolado em 2003 – e SARS-CoV-2 – para vírus isolado em 2019.

Para que um vírus seja considerado um novo espécime, ele não deve estar incluído em grupos conhecidos, mas sim distante desses grupos [6]. Nossas análises filogenéticas corroboram esta afirmação. Na árvore 4k (Figura 1c) identificamos que a cepa de pneumonia isolada do mercado de frutos do mar de Wuhan (GCF_009858895.2) - responsável pelo atual surto de SARS - está posicionada notavelmente perto das cepas de coronavírus relacionadas à síndrome respiratória aguda grave (GCF_000864885. 1) e coronavírus beta-morcego (GCF_000926915.1). Isto reforça a nossa afirmação anterior de que o rSWeeP é eficaz para análise e classificação taxonómica de organismos virais.

4 Conclusões

Mostramos que a implementação do modelo SWeeP na linguagem R, pacote rSWeeP, é igualmente precisa. O teste de desempenho, análises filogenéticas e PCA demonstram a eficiência do rSWeeP para análise de proteomas virais. O pacote rSWeeP apresenta uma solução para classificação taxonômica de organismos, disponível gratuitamente para toda a comunidade científica.

Agradecimentos

Os autores agradecem à Fundação Araucária e ao grupo de Inteligência Artificial aplicada à Bioinformática da Universidade Federal do Paraná.

Os autores declararam não haver conflito de interesses

Referencias

[1] Zielezinski, A., Vinga, S., Almeida, J. *et al.* Alignment-free sequence comparison: benefits, applications, and tools. *Genome Biol* **18,** 186 (2017). https://doi.org/10.1186/s13059-017-1319-7

[2] Asgari E., Mofrad M. R. K. Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics. PLoS ONE 10(11): e0141287 (2015). https://doi.org/10.1371/journal.pone.0141287

[3] Zhang, Q., Jun, S., Leuze, M. *et al.* Viral Phylogenomics Using an Alignment-Free Method: A Three-Step Approach to Determine Optimal Length of *k-mer*. *Sci Rep* **7,** 40712 (2017). https://doi.org/10.1038/srep40712

[4] Leimeister, C., Schellhorn, J., Dörrer, *et al*. *Prot-SpaM*: fast alignment-free phylogeny reconstruction based on whole-proteome sequences, *GigaScience* **8**, 3 (2019), giy148,

https://doi.org/10.1093/gigascience/giy148

[5] De Pierri, C.R., Voyceik, R., Santos de Mattos, L.G.C. *et al.* SWeeP: representing large biological sequences datasets in compact vectors. *Sci Rep* **10,** 91 (2020). https://doi.org/10.1038/s41598-019-55627-4.

[6] Gorbalenya, A.E., Baker, S.C., Baric, R.S. *et al.* The species *Severe acute respiratory syndrome-related coronavirus*: classifying 2019-nCoV and naming it SARS-CoV-2. *Nat Microbiol* **5,** 536–544 (2020). https://doi.org/10.1038/s41564-020-0695-z

[7] Lefkowitz, E. J., Dempsey, D. M., Hendrickson, R. C., et al. Virus taxonomy: the database of the International Committee on Taxonomy of Viruses (ICTV). *Nucleic acids research*, *46* (D1), D708–D717 (2018). https://doi.org/10.1093/nar/gkx932

[8] Calisher, C. H., Childs, J. E., Field, H. E., et al. .Bats: important reservoir hosts of emerging viruses. *Clinical microbiology reviews*, *19*(3), 531–545. (2006). https://doi.org/10.1128/CMR.00017-06

[9] Drosten C, Günther S, Preiser W, et al. Identification of a novel coronavirus in patients with severe acute respiratory syndrome. *N Engl J Med*. 348(20):1967-1976 (2003). https://doi:10.1056/NEJMoa030747

3 R3TO: ALGORITMO PARA ROTEAMENTO DE REDES ELÉTRICAS COM ALOCAÇÃO DE PÓLOS E TRANSPOSIÇÃO DE OBSTÁCULOS

International Journal of Advanced Engineering Research and Science (IJAERS)
Peer-Reviewed Journal
ISSN: 2349-6495(P) | 2456-1908(O)
Vol-9, Issue-10; Oct 2022
Journal Home Page Available: https://ijaers.com/
Article DOI: https://dx.doi.org/10.22161/ijaers.910.49

R3TO: Algoritmo para Roteamento de Redes Elétricas com Alocação de Postes e Transposição de Obstáculos

Danrley Rafael Fernandes, Géssica Michelle Dos Santos Pereira, Maricler Toigo, Roberto Tadeu Raittz

Instituto de Tecnologia para o Desenvolvimento (LACTEC), Curitiba-PR, Brasil

Recebido: 28 de setembro de 2022,

Recebido na forma revisada: 20 de outubro de 2022, Aceito: 25 de outubro de 2022,

Disponível on-line: 31 de outubro de 2022

***Palavras-chave*— Linhas de distribuição, caminho mínimo, algoritmos genéticos, Dijkstra.**

***Resumo*—*O aumento do consumo de energia e a necessidade de manter a continuidade do fornecimento, tornam essencial planear a expansão das redes eléctricas. Embora exista extensa literatura a respeito da geração automática de roteamento de redes de distribuição, os custos para alocação de postes ainda merecem atenção. Este estudo apresenta a proposta R3TO, que combina Algoritmos Evolutivos, com o algoritmo determinístico de Dijkstra, para resolver problemas de roteamento de redes considerando a adequação da alocação de postes. O R3TO é capaz de determinar a postagem em uma rede, com minimização no número de postes e transposição de obstáculos, como rios ou estradas. Para avaliar o método foram realizados testes em uma área teórica e no mapeamento de uma região de teste piloto em uma área real. Os resultados obtidos com o R3TO mostraram redução de custos quando comparado ao algoritmo determinístico tradicional, e obteve sucesso na alocação de postes em áreas restritas, encurtando o caminho total. Assim, ficou comprovado que R3TO gera um caminho minimizado, e com posições sugeridas para colocação de postes apresentando boas perspectivas para melhorar as abordagens atuais do problema.*

I. INTRODUCTION

The O livro Sobre a Origem das Espécies por Meio da Seleção Natural [1] inspirou revoluções científicas em diversas áreas. No campo da computação isso se expressou no surgimento dos Algoritmos Evolutivos e suas diversas vertentes [2]. Dentre estes, podem-se destacar os Algoritmos Genéticos (AG) [3], que utilizam conceitos evolutivos como mutação e recombinação genética, para gerar soluções adequadas a um ambiente de pressão evolutiva. Esta seletividade é modelada por uma função de custo (função de aptidão – FF), que visa obter soluções que tendem a ser ótimas. A otimização busca minimizar ou maximizar o valor retornado pelo FF de acordo com o algoritmo adotado. Assim um algoritmo pode ser aplicado – como neste estudo – para obter um traço que minimize a função que determina o custo de um determinado roteamento de uma rede. As características do AG tornaram-no uma abordagem comum a vários estudos recentes, que buscam resolver problemas de otimização de rastreamento [4][5][6]. Particularmente, uma das aplicações que pode se beneficiar deste tipo de metodologia é o roteamento de cabos de energia.

As áreas não urbanas oferecem frequentemente vários desafios para a distribuição de redes elétricas. Destes, podemos destacar as longas distâncias entre estradas e casas, terrenos acidentados e obstáculos naturais, como rios e falésias [7]. Para superar esses problemas, os engenheiros apresentaram diversos projetos, técnicas e sistemas para soluções de distribuição de redes elétricas, mas é preciso considerar que grande parte dos percursos da fiação elétrica ainda são planejados manualmente, através do método de tentativa e erro.

Um dos principais ganhos da aplicação de algoritmos de otimização de traços é o econômico. Ao reduzir o comprimento dos fios, a complexidade das redes e

diminuindo as quedas de tensão, é possível maximizar a rentabilidade de uma rota elétrica. Além disso, a redução das quedas de tensão pode garantir maior segurança e durabilidade dos equipamentos elétricos [8].

Neste artigo apresentamos um algoritmo para Roteamento de Redes Rurais com Alocação de Polos – R3TO, cujo principal objetivo é otimizar rotas de fiação elétrica, de forma a reduzir o comprimento dos circuitos e evitar obstáculos através de AG, explorando e melhorando soluções iniciais. Buscamos ampliar a capacidade oferecida pelo algoritmo Dijkstra reduzindo os custos de rastreamentos e adicionando novas possibilidades na modelagem de restrições.

A solução apresentada, quando possível, sobrepõe os obstáculos, ao invés de desviá-los, inovando o planejamento automático de expansão dos sistemas de distribuição de redes rurais com alocação de postes, que apresenta caminhos capazes de vencer obstáculos terrestres, como corpos d'água, terrenos irregulares.

O modelo R3TO tem a capacidade de explorar gráficos latissimus, além dos pontos imediatamente próximos. Este recurso permite que o R3TO utilize linhas entre pontos mapeados para passar por caminhos que não precisam ficar restritos às limitações das transições ponto a ponto de um gráfico. Isto faz com que a rota corresponda melhor ao relevo real de um mapa analisado. A Figura 1 ilustra uma situação em que o algoritmo R3TO supera Dijkstra em uma situação simulada.

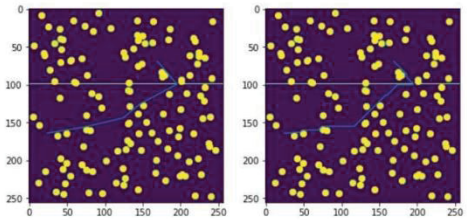


Fig.1 – Comparação entre R3TO (esquerda) e o método Dijkstra (direita).

II. ALGORITIMO R3TO

Para R3TO, o espaço de busca inicial é dado por um gráfico latisse determinado em um raster com um valor de custo em cada ponto. Ao identificar o ponto de origem e destino, determina-se o objetivo, que é obter a rota de menor custo possível que una esses pontos. Um conjunto de pontos, em uma ordem específica, define uma solução candidata. Cada solução possível pode ser representada por um conjunto de bits – cromossomo. O algoritmo genético constrói a evolução por meio de substituições sucessivas de populações de soluções. A estratégia desenvolvida para este trabalho é alimentar a população inicial adicionando cromossomos com alto potencial evolutivo ao conjunto de soluções geradas aleatoriamente. Para os primeiros cromossomos foram modelados contendo:

a) Pontos intermediários da reta contendo os pontos inicial e final;

b) Pontos intermediários da reta contendo os pontos inicial e final;

Estabelecemos também uma estratégia de feedback que inclui cromossomas adaptados às melhores soluções em cada geração.

Cada cromossomo consiste em uma sequência de bits cujo tamanho depende das dimensões do raster e do número máximo de pontos intermediários a serem considerados, que são determinados pelo analista. Aos bits que representam as coordenadas dos pontos é adicionado um bit lógico que habilita (1) ou não (0) a utilização da coordenada na solução representada. O número de bits por cromossomo é dado por:

$$BS = PM^* [bool + DR1 + DR2]$$

Onde:

BS: comprimento da cadeia de bits (em bits)

PM: número máximo de pontos intermediários (parâmetro definido pelo usuário);

Bool: 1 bit que habilita a coordenada;

DR1: número de bits necessários para representar o número de linhas da matriz (raster);

DR2: número de bits necessários para representar o número de colunas da matriz (raster).

A decodificação de cada cromossomo fornece um traço que é avaliado pela função de aptidão. O custo de cada solução das populações é calculado até que o número máximo de iterações seja atingido. A avaliação se dá calculando um custo médio para percorrer o latisse pelos pontos indicados pela trajetória, multiplicado pela distância euclidiana total para percorrer o caminho pelos pontos intermediários contidos na solução – cromossomo correspondente decodificado.

Além de buscar o menor valor de custo possível para a rede, a função fitness visa otimizar a alocação de polos. Devido a isso, o algoritmo é capaz de sugerir pontos estratégicos no gráfico para posicionamento dos postes, permitindo que o layout "pule" obstáculos. É importante ressaltar que o algoritmo não tem como objetivo definir o posicionamento preciso de cada um dos polos, mas sim obter um percurso onde não existam regiões intransponíveis por possível postagem. A solução identifica dois tipos distintos de obstáculos: intransponíveis e transponíveis. Os obstáculos intransponíveis são marcados como um nó muito caro no raster, o que, portanto, obriga o layout a desviá-los

Os obstáculos transponíveis são ultrapassados em formato raster, com pontos específicos do gráfico. Isso é usado para penalizar apenas as postagens sugeridas nesta área, e não o rastreamento em si. Na parametrização o analista define dois valores:

- Peso 1, (Ω1), que pressiona pela diminuição do número de postes, evitando sugestões desnecessárias.
- Peso 2 (Ω2) para evitar que postes sejam sugeridos em áreas de obstáculos intransponíveis.

O valor de Ω1, é definido dividindo o custo do rastreamento pelo peso 1. O Ω1 é multiplicado pelo número de postes no rastreamento gerando a penalidade 1. O valor de Ω2, definido pela multiplicação do custo do rastreamento pelo peso 2. O Ω2 é multiplicado pelo número de postes colocados em áreas restritas (obstáculos), gerando a penalidade 2. Ao final, as penalidades 1 e 2 são somadas ao custo do traçado, retornando o custo final do traçado por:

$$FO = MIN\ CM + (\Omega 1 \div x1) + (\Omega 2\ y2)$$

Onde:

FO: função objetivo;

CM: custo médio de roteamento sem postes;

Ω1: peso que penaliza o número de postes.

X1: número de pólos;

Ω2: peso que penaliza postes alocados na transposição áreas;

X2: número de postes alocados nas áreas de transposição.

O modelo cumpre as seguintes etapas principais em seu processo: geração da solução inicial, decodificação cromossômica e avaliação cromossômica. O fluxograma geral do método é apresentado na Figura 2.

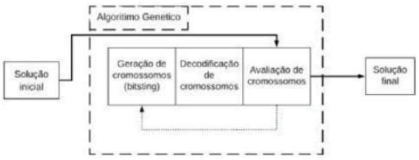


Fig.2 – Fluxograma geral do método

O desenvolvimento do algoritmo foi feito na linguagem Python, a partir de um protótipo construído em MatLab®. Para resolver a solução inicial de Dijkstra foi utilizada a biblioteca Skimage, com o módulo Graph e a função Route_Through_Array. Para o algoritmo genético foi utilizada a biblioteca SciPy, utilizando o método Brusque Scanning considerando os parâmetros:

- Estrategia: best1bin
- Crossover: 0.35
- Seed: 10
- Maxiter: 1000
- Mutação: 0.8

III. RESULTADOS

As figuras a seguir demonstram a aplicação deste método em uma área piloto, na região rural da cidade de Piên/PR.

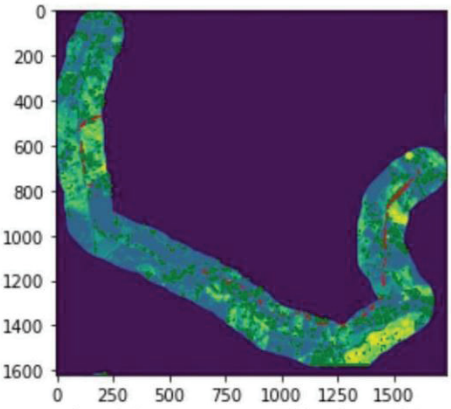


Fig.3 - The red dots represent possible poles, and the pixels in green the restricted area. Lines between the indicated poles do not contain insurmountable regions.

A Figura 3 é um bitmap gerado a partir da matriz de custos da área piloto de Piên. Nesta imagem os pixels em verde marcam as áreas que não permitem a colocação de postes, conforme determinação de um analista. Já os pontos vermelhos são uma marcação amplificada dos pixels que representam a área onde pode ser interessante colocar postes. É possível notar que as sugestões para aluguel de postes transpõem as áreas restritas em verde. Também é possível verificar a qualidade do resultado através da saída final do R3TO, que é um array onde cada linha representa as coordenadas YX de cada post dentro da imagem bitmap. A Figura 4 mostra um trecho de uma imagem georreferenciada do mapeamento da região de Piên onde é notável a transposição de um obstáculo (estrada) pelo R3TO.

Fig.4 –Destaque mostrando como o algoritmo permite soluções atravessando uma estrada.

Neste teste, o algoritmo obteve sucesso em encontrar um caminho de rede otimizado, com transposição de obstáculos e minimização de pólos em uma região real. Porém, foi observado um tempo de processamento de 65 horas e 24 minutos para processá-lo. O custo computacional está compensando, no entanto, os ganhos econômicos.

IV. CONCLUSÃO

Este artigo procurou destacar a necessidade de estudos para solucionar os problemas de expansão da rede no meio rural. Foi apresentada uma proposta de solução utilizando a linguagem de programação Python, incluindo um algoritmo que reduz o número de postos alocados e a transposição de obstáculos, ao contrário de outros estudos anteriores que apenas os desviaram, desconsiderando que as companhias aéreas podem superar obstáculos em vez de apenas desviá-los. Isto foi evidenciado pela apresentação de testes em um mapa sintético e em uma área de teste real. Mostramos uma solução promissora que pode contribuir de diversas maneiras para a geração automática de roteamento de rede. Porém, outros estudos visando mais testes e melhoria no tempo de processamento para obtenção das soluções são necessários.

AGRADECIMENTOS

Os autores agradecem à Agência Nacional de Energia Elétrica (ANEEL) e à Copel Distribuição S.A. (COPEL-DIS) pelo financiamento do projeto de P&D ANEEL PD-02866- 0509/2019 "UAV e IA COMO SUBSÍDIO PARA GERAÇÃO AUTOMÁTICA DE LD ÓTIMOLAYOUT", do qual este trabalho fez parte. Agradecemos também a contribuição dos demais pesquisadores do Lactec envolvidos neste projeto.

REFERÊNCIA

[1] Darwin, C. 1869. On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life. London: John Murray,

[2] Eiben, A. E., Smith, J. E.2003.Introduction to evolutionary computing (Vol. 53, p. 18). Berlin: springer.

[3] Holland, J.H. 1962.Outline for a logical theory of adaptive systems. J. Assoc. Comput. Mach., vol. 3, pp. 297–314.

[4] Ha, Q. M., Deville, Y., Pham, Q. D., & Hà, M. H.2020.A hybrid genetic algorithm for the traveling salesman problem with drone. Journal of Heuristics, 26(2), 219-247

[5] Nurdiawan, O., Pratama, F. A., Kurnia, D. A., Rahaningsih, N.2020.Optimization of Traveling Salesman Problem on Scheduling Tour Packages using Genetic Algorithms. In Journal of Physics. Johannesburg, pp. 42–46. doi:10.1109/67.560861

[6] Boardman JT, Meckiff CC.1985.A branch and bound formulation of an electricity distribution planning problem. Ieee Trans. Power App. Syst. 104:2112–2118

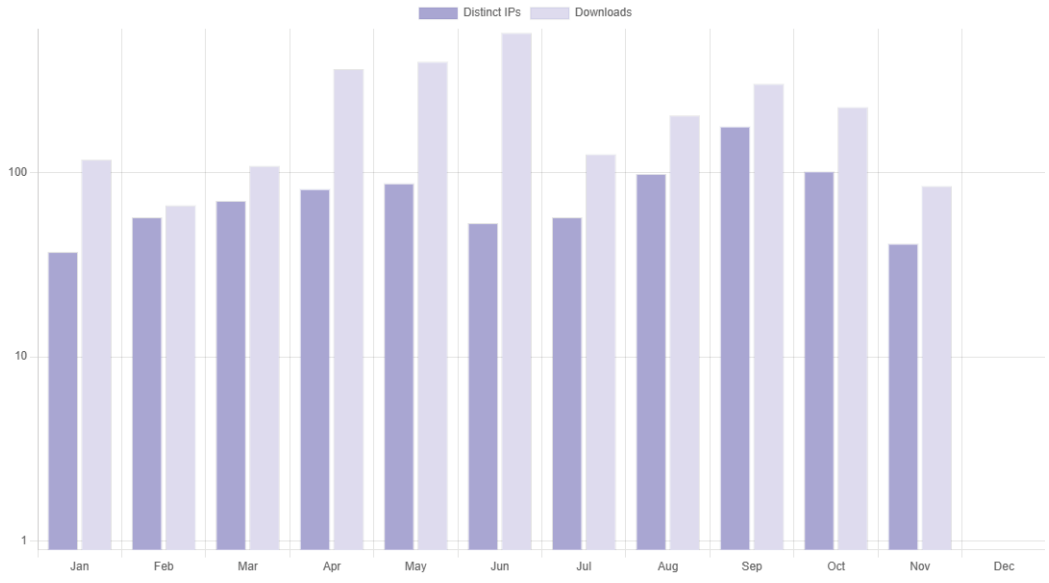
[7] Bouchard, D.E. Salama, M.M.A. Chikhani, A.Y.1994.Optimal distribution feeder routing and optimum substation sizing and placement using evolutionary strategies. In: Canadian Conference On Electrical And Computer Engineering, p. 661-664.

[8] Karaboga, N. Cetinkaya, B.(2003).Performance Comparison of Genetic Algorithm based Design Methods of Digital Filters with Optimal Magnitude Response and Minimum Phase. In: The 46th IEEE Midwest Symposium on Circuits and Systems.

4 CONCLUSÃO

Nos últimos anos, o método SWeeP, foi amplamente difundido através do pacote rSWeeP no ambiente R/Bioconductor e tem se tornado uma ferramenta popular para a bioinformática. Demonstrando eficaz para facilitar análises filogenéticas de grandes volumes de dados com alta eficiência, ele é capaz de representar sequências biológicas em vetores de baixa dimensão (Perico et al, 2022; Fernandes et al, 2022; de Pierri et al, 2020). E implementação rSWeeP vem se mostrando com uma plataforma auxiliar eficaz na expansão da base de usuários do método, sendo utilizada em publicação recente de grande relevância cientifica (Perico et al, 2022). A FIGURA 1 **- DOWLOADS DA FERRAMENTA RSWEEP NA PLATAFORMA BIOCONDUCTOR** ilustra a frequência que os usuários procuraram a ferramenta no ano de 2024.

FIGURA 1 - DOWLOADS DA FERRAMENTA RSWEEP NA PLATAFORMA BIOCONDUCTOR



Month	Nb of distinct IPs	Nb of downloads
Jan/2024	37	117
Feb/2024	57	66
Mar/2024	70	108
Apr/2024	81	362
May/2024	87	396
Jun/2024	53	569
Jul/2024	57	125
Aug/2024	98	203
Sep/2024	177	301
Oct/2024	101	225
Nov/2024	41	84
Dec/2024	0	0
all/2024	763	2556

FONTE: Bioconductor (2024).

Além disso, o estudo "*Genomic landscape of the SARS-CoV-2 pandemic in Brazil suggests an external P.1 variant origin*"**,** que utilizou o rSWeeP, destacou-se ao analisar a pandemia no Brasil e identificar uma possível origem externa para a variante

P.1. Esse trabalho aponta que, durante a segunda onda de COVID-19 no Brasil, a diversidade genética do vírus aumentou drasticamente, especialmente com o surgimento da variante P.1, cujas origens podem estar associadas a um evento recombinante com a variante B.1.1.28 (Perico et al., 2024). Esse tipo de análise foi fundamental para entender a dispersão das variantes de preocupação e as implicações na saúde pública, ilustrando que a ferramenta rSWeeP, pode ser uma ferramenta poderosa para estudos de análises de sequencias biológicas.

No campo da otimização de redes elétricas, o algoritmo R3TO oferece uma abordagem inovadora em comparação com os métodos determinísticos tradicionais, como o algoritmo de Dijkstra. Combinando algoritmos evolutivos, o R3TO supera obstáculos no terreno (rios, montanhas) ao otimizar a alocação de postes e reduzir custos no traçado de redes de distribuição elétrica em áreas rurais. Essa vantagem foi observada em diversos testes realizados, demonstrando que o R3TO consegue propor soluções eficientes em locais onde métodos convencionais seriam limitados. A importância dessa inovação foi reconhecida em workshops organizados pela empresa Lactec para a COPEL (Companhia Paranaense de Energia), onde engenheiros e técnicos puderam conhecer as aplicações práticas do R3TO e sua contribuição para um planejamento de rede mais eficiente e economicamente viável.

4.1 Recomendações para trabalhos futuros

Desta forma podemos afirmar que dissertação aqui apresentada buscou explorou o potencial das ferramentas rSWeeP e R3TO em análises filogenéticas e otimização de redes, propondo caminhos para futuras investigações que podem ampliar suas capacidades, com base nos recentes avanços nos métodos livres de alinhamento e algoritmos determinísticos para medidas de distância.

O método SWeeP, em sua implementação R (rSWeeP), oferece uma plataforma promissora para popularizar a utilização de modelos livres de alinhamento para análises filogenéticas, especialmente ao apresentar um modelo de frequências de k-mers. Sarmashghi et al. (2019) introduziram o Skmer, uma ferramenta que utiliza perfis de k-mers para estimar distâncias genômicas de forma eficiente e precisa, demonstrando a viabilidade de métodos livres de alinhamento em análises filogenéticas. Além disso, Höhl et al. (2006) desenvolveram um método de estimativa de distância filogenética baseado em padrões, sem a necessidade de alinhamento,

destacando a eficácia de abordagens determinísticas na reconstrução de árvores filogenéticas. Assim, a incorporação de técnicas de k-mers na rSWeeP pode potencializar análises comparativas de linhagens virais, viabilizando estudos filogenéticos de larga escala com alta precisão.

O algoritmo R3TO, inicialmente aplicado para otimizar redes de distribuição elétrica, mostrou-se promissor na transposição de obstáculos e na criação de rotas otimizadas. Fernandes et al. (2022) demonstraram, em testes de campo, que a aplicação do R3TO reduziu custos e tempo quando comparado com algoritmos determinísticos. No contexto da bioinformática, sua aplicação na otimização de rotas de transmissão de dados genéticos permitiria a integração de redes filogenéticas e a distribuição eficiente de grandes conjuntos de dados entre diferentes centros de pesquisa. Esta abordagem otimizada seria particularmente benéfica em sistemas de vigilância epidemiológica, onde a velocidade na análise e no compartilhamento de dados pode ter impacto direto em decisões de saúde pública (SILVA & Alves; 2023).

Por fim, a implementação rSWeeP possibilita uma gama maior de uso para o método SWeeP, que já se mostrou em diversos estudos como uma ferramenta útil para análises de dados biológicos (Perico et al, 2022; Fernandes et al, 2022; de Pierri et al, 2020). E o método R3TO se mostra como uma poderosa possível ferramenta complementar para análises distancia de sequencias biológicas, tanto em complementariedade ao método SWeeP quanto em outros métodos.

REFERÊNCIAS

1. BRANDÃO, R. INTELIGÊNCIA ARTIFICIAL, TRABALHO E PRODUTIVIDADE. **Revista de Administração de Empresas**, v. 60, n. 5, p. 378–379, 2020.
2. CALHAU, F. G.; MARTINS, J. S. B. A Electric Network Reconfiguration Strategy with Case-Based Reasoning for the Smart Grid., 2019. Disponível em: <http://arxiv.org/abs/1907.05885>.
3. CALHAU, F. G.; PEZZUTTI, A.; MARTINS, J. S. B. On Evaluating Power Loss with HATSGA Algorithm for Power Network Reconfiguration in the Smart Grid., 2022. Disponível em: <http://arxiv.org/abs/2205.10126>.
4. HÖHL, Michael; RIGOUTSOS, Isidore; RAGAN, Mark A. Pattern-based phylogenetic distance estimation and tree reconstruction. **Evolutionary Bioinformatics**, v. 2, p. 117693430600200016, 2006.
5. HUBER, Katharina T.; MOULTON, Vincent; SCHOLZ, Guillaume E. Shared ancestry graphs and symbolic arboreal maps. **SIAM Journal on Discrete Mathematics**, v. 38, n. 4, p. 2553-2577, 2024.
6. KARIM, M. R.; ISLAM, T.; SHAJALAL, M.; et al. Explainable AI for Bioinformatics: Methods, Tools, and Applications. **Briefings in Bioinformatics**, v. 24, n. 5, 2023. Oxford University Press.
7. OWERKO, D.; GAMA, F.; RIBEIRO, A. Optimal Power Flow Using Graph Neural Networks., 2019. Disponível em: <http://arxiv.org/abs/1910.09658>.
8. PERICO, C. P.; FERNANDES, D. R.; VARASCHIN, J. F.; et al. rSWeeP: Alignment-free method for vectorising biological sequences.
9. DE PIERRI, C. R.; VOYCEIK, R.; SANTOS DE MATTOS, L. G. C.; et al. SWeeP: representing large biological sequences datasets in compact vectors. **Scientific Reports**, v. 10, n. 1, p. 91, 2020.
10. SARMASHGHI, S., BOHMANN, K., P. GILBERT, M.T. *et al.* Skmer: assembly-free and alignment-free sample identification using genome skims. *Genome Biol* **20**, 34 (2019). https://doi.org/10.1186/s13059-019-1632-4
11. SHARMA, S.; PARTAP, A.; BALAGUER, M. A. DE L.; MALVAR, S.; CHANDRA, R. DeepG2P: Fusing Multi-Modal Data to Improve Crop Production., 2022. Disponível em: <http://arxiv.org/abs/2211.05986>.
12. SILVA, Ruana Carolina Cabral da; ALVES, Maria Cidinaria Silva. O uso de ferramentas de bioinformática para análise de dados genéticos: uma revisão. **Scientific Electronic Archives**, *[S. l.]*, v. 17, n. 1, 2023. DOI: 10.36560/17120241872.
13. TRINDADE, A. S. C. E. DA; OLIVEIRA, H. P. C. DE. INTELIGÊNCIA ARTIFICIAL (IA) GENERATIVA E COMPETÊNCIA EM INFORMAÇÃO: HABILIDADES INFORMACIONAIS NECESSÁRIAS AO USO DE FERRAMENTAS DE IA GENERATIVA EM DEMANDAS INFORMACIONAIS DE NATUREZA ACADÊMICA-CIENTÍFICA. **Perspectivas em Ciência da Informação**, v. 29, 2024.
14. ZOU, Y.; ZHANG, Z.; ZENG, Y.; et al. Common Methods for Phylogenetic Tree Construction and Their Implementation in R. **Bioengineering**, 1. maio 2024. Multidisciplinary Digital Publishing Institute (MDPI).