57º SBPO
SIMPÓSIO BRASILEIRO DE PESQUISA OPERACIONAL
Gramado • RS
DE 5 A 9 DE OUTUBRO DE 2025

Problemas de consensos para árvores de mutação

Thiago Lopes Nascimento
Universidade Federal Fluminense
Departamento de Ciência da Computação - IC - Niterói - RJ
thiago_nascimento@id.uff.br

Luís Felipe Ignácio Cunha
Universidade Federal Fluminense
Departamento de Ciência da Computação - IC - Niterói - RJ
lfignacio@ic.uff.br

RESUMO

A heterogeneidade mutacional dos tumores pode ser descrita com uma árvore para representar a história evolutiva do tumor. Com dados de sequenciamento ruidosos, pode haver incerteza na estrutura da árvore inferida. O ruído inerente aos métodos estatísticos atuais para a construção de árvores de evolução de mutações de células de câncer apresenta um desafio significativo: lidar com esse conjunto de árvores para determinar uma árvore de consenso que represente com precisão o conjunto e avaliar a extensão de sua variabilidade ou dispersão. Dado um conjunto de árvores de mutação e a noção de distância, há pelo menos duas maneiras naturais de definir a árvore "alvo", como uma min-sum (*árvore Mediana*) ou um min-max (*árvore Closest*) de um conjunto de árvores. Assim, considerando um conjunto de árvores como entrada e lidando com os problemas MEDIANA e CLOSEST, provamos que ambos os problemas são NP-completos, mesmo com apenas três árvores de entrada.

PALAVRAS CHAVE. Problema da Mediana. Problema de Closest. Distância entre árvores.

Teoria dos Grafos e Algoritmos Relacionados, Otimização Contínua

ABSTRACT

The mutational heterogeneity of tumours can be described with a tree representing the evolutionary history of the tumour. With noisy sequencing data there may be uncertainty in the inferred tree structure. The inherent noise in current statistical methods for constructing mutation evolution trees of cancer cells presents a significant challenge: handling such collections of trees to determine a consensus tree that accurately represents the set and evaluating the extent of their variability or dispersion. Given a set of mutation trees and the notion of distance, there are at least two natural ways to define the "target" tree, such as a min-sum (*median tree*) or a min-max (*closest tree*) of a set of trees. Thus, considering a set of trees as input and dealing with the MEDIAN and CLOSEST problems, we prove that both problems are NP-complete, even with only three input trees.

KEYWORDS. Median problem. Closest problem. Tree distances.

Graph Theory and Related Algorithms, Continuous Optimization

DOI: 10.59254/sbpo-2025-212286
http://live-proceedings.pantheonsite.io/p/212286

1. Introdução

Inferência de sequência por meio de uma única célula Árvores filogenéticas são ferramentas importantes usadas no estudo da evolução de espécies e vírus [Gorbalenya e Lauber, 2017]. De forma similar, elas também apresentam um papel importante na evolução de tumores. Conforme o câncer progride, as células do tumor acumulam uma diversa e complexa quantidade de mutações genéticas, que por sua vez permitem ao mesmo tumor se proliferar e evoluir mais. O sequenciamento de tumores no nível de células individuais pode prover uma reconstrução com alta resolução do histórico evolucionário dos cânceres [Kuipers et al., 2017]. Entre grupos de pacientes [Morita et al., 2020], padrões evolucionários comuns podem ser aprendidos para começar a modelar e predizer futuras evoluções [Luo et al., 2023].

Diferente da filogenética clássica, que infere uma árvore de descendentes [Gorbalenya e Lauber, 2017], em oncologia computacional existe um foco maior na reconstrução da árvore de eventos de mutação. Nessas árvores, os nós são as próprias deformidades genômicas, ao contrário de serem as células do tumor [Kuipers et al., 2017]. A árvore de eventos representa diretamente o conjunto de genótipos presentes em diferentes clones de um tumor, assim como os estados ancestrais. Portanto, descreve diretamente os conjuntos de populações de células que podem desenvolver resistência a tratamentos e causar relapso [McGranahan e Swanton, 2015].

Sequenciamento baseado em uma única célula é tecnicamente desafiador, apresenta alta quantidade de ruído e erros no processamento que podem gerar incertezas na estrutura da árvore filogenética, que pode ser caracterizada pela coleta de amostras de árvores da sua distribuição posterior [Jahn et al., 2016]. Um desafio é, portanto, trabalhar com esse conjunto de objetos combinatórios, para, por exemplo, encontrar uma árvore consenso que resuma o conjunto e entender o quão amplamente espalhadas essas árvores estão. Quando se considera um conjunto de pacientes, mais desafios surgem, como em entender a distribuição e distância entre as árvores de diferentes realidades da evolução dos tumores e extrair padrões comuns de árvores entre esses pacientes.

Figura 1 ilustra a evolução de um tumor e a sua matriz de mutações associada, que mostra quais mutações estão presentes em cada célula. Tendo em vista os dados presentes em Figura 1(c), é feita a transformação para a árvore Figura 1(a), onde os nós internos representam as mutações e as folhas as células. Dessa forma, é feita uma associação com problemas envolvendo árvores que são estudados em teoria dos grafos.

Os dados obtidos pelos experimentos de sequenciamento baseados em uma única célula são propensos a conter erros. Esses erros estão presentes na forma de falsos negativos, falsos positivos e dados faltantes. Árvores geradas com esses dados são denotadas como *árvores imperfeitas*, como visto em Figura 1(b). A matriz Figura 1(c) representa um cenário ideal onde todos os elementos estão presentes e são válidos.

Problemas de consenso considerando permutações Como proposto por [Watterson et al., 1982], um problema de *Rearranjo de genoma* é interpretado como a transformação de uma permutação em outra através do menor número de operações, dependendo dos possíveis rearranjos permitidos. Se uma entrada tiver mais de dois genomas, existem diversas abordagens para encontrar genomas ancestrais [Bader, 2011; Caprara, 2003; Cunha et al., 2020; Cunha e Protti, 2019; Haghighi e Sankoff, 2012; Pe'er e Shamir, 1998]. Um problema relevante é o problema da MEDIANA, em que, para uma métrica M, o objetivo é encontrar uma solução que minimize a soma das distâncias entre a solução e todos os genomas de entrada.

[Haghighi e Sankoff, 2012] observaram que, no que diz respeito à métrica de breakpoint, a tendência para a mediana é estar sobre ou perto de um dos genomas de entrada, o que não representa uma informação útil para a reconstrução filogenética. Eles também conjecturaram o

DOI: 10.59254/sbpo-2025-212286
http://live-proceedings.pantheonsite.io/p/212286

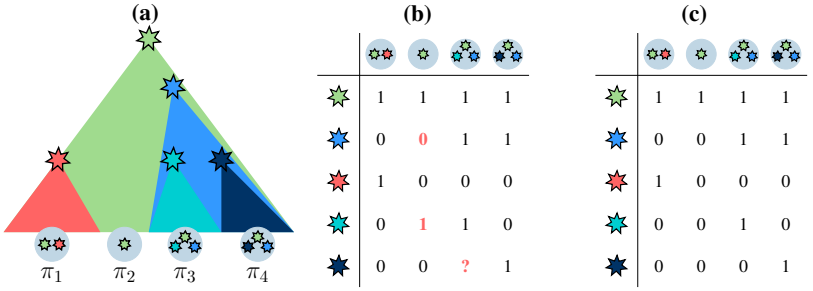


Figura 1: (a) Representação da evolução de um tumor. Estrela representa uma nova mutação e a expansão de um subclone. Círculos representam células sequenciadas após a remoção de um tumor e as estrelas dentro delas mostram quais mutações estão presentes em cada célula. (b) Matriz de mutação real representando os estados de células de tumor sequenciadas. Uma entrada zerada representa a ausência de uma mutação na respectiva célula enquanto uma entrada contendo um representa a presença. 0, 1 e **?** denotam falso negativo, falso positivo e dado faltante, respectivamente. (c) Matriz de mutação ideal representando os estados de células de tumor sequenciadas.

mesmo comportamento para outras métricas. Dessa forma, uma abordagem alternativa é considerar o problema de CLOSEST para uma métrica fixa M, que visa encontrar um genoma que minimize a distância máxima a qualquer genoma na entrada, o que pode ser visto como encontrar um genoma no centro de todos os outros, ou seja, um genoma correspondente ao raio do conjunto de entrada.

[Lanctot et al., 2003] estudaram o problema de CLOSEST para strings considerando a distância de Hamming, e concluíram que este problema é NP-difícil mesmo para strings binárias. [Popov, 2007] estudou o problema de CLOSEST considerando permutações utilizando a operação de swap, e mostrou que é NP-difícil. [Cunha et al., 2020] mostrou que o problema de CLOSEST é NP-difícil para algumas distâncias de rearranjo do genoma bem conhecidas. Também mostrou resultados CLOSEST (*fixed parameterized tractable*) para os problemas de CLOSEST e MEDIANA considerando diversas métricas em permutações [Cunha et al., 2024a,b].

Resultados Esse trabalho visa continuar o estudo iniciado em [Nascimento e Cunha, 2024], na qual são demonstradas heurísticas para determinar a distância entre árvores, além de desenvolver algoritmos para encontrar o consenso entre um conjunto de árvores. Nesse trabalho, se considera problemas mais gerais onde a entrada consiste em mais de duas árvores e o objetivo é determinar o min-max (CLOSEST) e a min-sum MEDIANA árvores soluções, no final provando que ambos são NP-completo mesmo quando se tem somente três árvores como entrada.

2. Preliminares

É necessário estabelecer como é feito o cálculo da distância entre duas árvores, e para isso deve-se entender a estrutura das árvores trabalhadas. No contexto das árvores de filogenia do câncer, os nós internos representam populações que sofreram a mesma mutação. Todo nó interno deve ter ao menos dois descendentes. Isso se deve pois um único descendente representaria que a população inteira passou por duas mutações, de modo que essas mutações afetaram todo o grupo sem que ocorresse uma ramificação. Em virtude desses eventos afetarem o grupo como um todo, ambas mutações podem ser colapsadas como um único evento. A existência de nós com um único descendente é uma característica biologicamente redundante.

Determinar a distância entre T_p e T_q, definida como $d(T_p, T_q)$, é encontrar a menor quantidade de reordenamentos de vértices em T_p de modo a transformar T_p em T_q. Um *reordenamento*

DOI: 10.59254/sbpo-2025-212286
http://live-proceedings.pantheonsite.io/p/212286

de vértice em T é definido como a movimentação de uma subárvore do seu nível atual para um nível acima ou abaixo em T. Sejam dois vértices u e v, na qual v é pai de u, o reordenamento do vértice u em T é feito pela remoção da aresta u, v e pela adição de uma aresta entre u e j, sendo que j pode ser tanto o pai de v quanto um outro filho de v. Uma subárvore é tanto um nó interno junto a todos os seus descendentes quanto uma folha única de T.

A altura de uma árvore T, denotada como $h(T)$, é definida como a quantidade de arestas do maior caminho entre a raiz da árvore e uma folha.

A distância entre árvores pode ser analisada de duas formas distintas: baseada na altura das árvores ou por meio da *maior subárvore comum* [Nascimento e Cunha, 2024].

Uma permutação pode ser representada como um elemento seguido da sua imagem. Seja $\pi : \Sigma \to \Sigma$ permutações, onde $\Sigma = \{1, 2, \ldots, n\}$. Para um elemento $x \in \Sigma$, define-se $\pi^k(x)$ como o resultado de aplicar π repetidamente k vezes em x. Formalmente, é demonstrado de forma recursiva: $\pi^0(x) = x,\ \pi^k(x) = \pi(\pi^{k-1}(x))$ para $k > 0$. A notação $\pi^k(x)$ representa o elemento Σ encontrado após k aplicações de π iniciando em x.

Um *ciclo algébrico* de uma permutação π é um subconjunto de elementos em Σ que são permutados de modo cíclico por repetidas aplicações de π. Especificamente, começando por qualquer $x \in \Sigma$, o ciclo contendo x é definido como: $C = \{x, \pi(x), \pi^2(x), \ldots, \pi^{k-1}(x)\}$, onde k é o menor inteiro positivo de tal modo que $\pi^k(x) = x$. Todo ciclo C satisfaz as seguintes condições:

1. *Fechamento cíclico:* Para cada $x \in C$, $\pi^k(x) = x$ para o menor $k > 0$, e $\pi^j(x) \neq x$ for $0 < j < k$.

2. *Elementos distintos:* Cada elemento de C é distinto sobre repetidas aplicações de π, ou seja, $\pi^i(x) \neq \pi^j(x)$ para $0 \leq i < j < k$.

3. *Ciclos disjuntos:* Os ciclos π são disjuntos, ou seja, não existe elemento que pertença a mais de um ciclo. Além disso, todo $x \in \Sigma$ pertence a exatamente um ciclo.

Uma operação de *swap* envolvendo os elementos a e b, onde a e b pertencem ao mesmo ciclo, é chamada de swap que *quebra* o ciclo em dois. Em contrapartida, se a e b pertencem a ciclos diferente, trocar esses elementos irá *unir* ambos os ciclos Fertin et al. [2009]. A distância de swap entre duas permutações π e σ é definida por $d_{\mathsf{swap}}(\pi, \sigma) = n - c(\pi, \sigma)$.

3. Problemas de consenso

A relação entre a distância de *swap* e árvores de altura dois implica que o conhecimento dos problemas de consenso pode ser diretamente aplicado ao trabalhar com árvores. Popov [Popov, 2007] provou que o problema de CLOSET com *swap* é NP-completo. Como cada operação de *swap* quebra ou merge ciclos, ela pode ser vista como uma operação em uma árvore de altura dois. Desse modo, uma permutação é uma solução para o problema de CLOSET com *swap* se e somente se a sua representação em ciclo é uma solução do problema de CLOSEST com o formato de Newick [Nascimento e Cunha, 2024]. Portanto, isto implica na Proposição 1.

Proposição 1. *A solução para o problema de* CLOSEST *tendo um conjunto de árvores é* NP*-completo, mesmo que a altura das árvores da entrada seja igual a 2.*

Em relação ao problema da mediana, provamos o seguinte: o problema da mediana para um conjunto de árvores é NP-completo, mesmo que a entrada seja somente três árvores. Essa restrição do problema da mediana será chamada de TREE-MEDIAN$_3$. A NP-dificuldade do TREE-MEDIAN$_3$ (Teorema 2) é obtida por uma redução do BREAKPOINT-MEDIAN$_3$ considerando instâncias que não contenham dois breakpoints consecutivos. Essa restrição para o problema do

DOI: 10.59254/sbpo-2025-212286
http://live-proceedings.pantheonsite.io/p/212286

BREAKPOINT-MEDIAN$_3$ tem sua prova de NP-dificuldade em Teorema 1, vindo do problema NP-completo BREAKPOINT-MEDIAN$_3$ genérico [Pe'er e Shamir, 1998]. Baseado nisso, provamos em Teorema 3 que o problema de CLOSEST para um conjunto de árvores é NP-completo, mesmo para três árvores. Este problema é denotado como de TREE-CLOSEST$_3$.

Uma *adjacência de uma permutação π com relação a uma permutação σ* é um par $(\pi[i], \pi[i+1])$, sendo dois elementos consecutivos em π de modo que eles estão sejam consecutivos em σ, i.e., $(\pi[i] = \sigma[j])$ e $(\pi[i+1] = \sigma[j+1])$. Se um par de elementos consecutivos não formar uma adjacência, então esse par é chamado de $breakpoint$, e denotamos por $d_{BP}(\pi, \sigma)$ como o número de breakpoint de π com relação a σ. O conjunto $Adj(\pi)$ é o conjunto de *adjacências de* π, dado por $Adj(\pi) = \{\{\pi[i], \pi[i+1]\} | i = 1, \cdots, n-1\}$. Dessa forma, a distância de breakpoint entre π e σ é $d_{BP}(\pi, \sigma) = |Adj(\pi) - Adj(\sigma)|$.

Teorema 1. BREAKPOINT-MEDIAN$_3$ *é* NP*-completo, mesmo para instâncias que não tenham dois breakpoints consecutivos.*

Demonstração. BREAKPOINT-MEDIAN$_3$ é NP-completo, provado por [Bryant, 1998] e [Pe'er e Shamir, 1998]. Agora, provamos que esse problema se mantém NP-completo mesmo que a entrada consista de três permutações que não tenham dois breakpoints consecutivos. Essa prova é feita por uma redução de BREAKPOINT-MEDIAN$_3$.

Seja π_1, π_2, π_3 permutações de tamanho n entradas para o BREAKPOINT-MEDIAN$_3$. Para cada permutação, definimos uma operação de extensão transformando-a em outra pelo seguinte procedimento: entre dois elementos consecutivos $\pi_x[i], \pi_x[i+1]$, adicione um par de novos elementos $n+1, n+i+1$, para $i = 1, \cdots, n-1$ e $x \in \{1, 2, 3\}$.

Cada uma das três permutações possui tamanho $3n-2$ e foram adicionadas $n-1$ novas adjacências da forma $n+1, n+i+1$, para $i = 1, \cdots, n-1$. Como provado em [Bryant, 1998], se uma adjacência ocorre em todas as permutações de entrada então também ocorrerá em uma solução BREAKPOINT-MEDIAN. Portanto, uma solução do BREAKPOINT-MEDIAN$_3$ terá todas as novas adjacências adicionadas. Além disso, um par $\pi_x[i], \pi_x[i+1]$ é um breakpoint se e somente se $\pi_x[i], n+1$ e $\pi_x[i+1], n+i+1$ também forem breakpoints na versão estendida da permutação π_x. Caso contrário, $\pi_x[i], n+1$ for uma adjacência, isso implica que existe um par $\pi_y[i], n+1$ para $y \neq x$, e que ele também deve ter $\pi_y[i+1], n+i+1$, visto que $n+1$ e $n+i+1$ são colocados juntos na construção das permutações da entrada. Portanto, duas permutações π_x e π_y possuem $d_{BP}(\pi_x, \pi_y) = k$ se e somente se as versões estendidas π'_x e π'_y possuírem $d_{BP}(\pi'_x, \pi'_y) = 2k$.

Baseado nisso, temos que σ é uma solução do BREAKPOINT-MEDIAN$_3$ com entradas π_1, π_2, π_3 se e somente se a versão estendida σ' for uma solução BREAKPOINT-MEDIAN$_3$ com as entradas sendo obtidas pela extensão de π_1, π_2, π_3.

□

Teorema 2. TREE-MEDIAN$_3$ *é* NP*-completo, mesmo para árvores com altura no máximo 2.*

Demonstração. Essa prova é feita por meio de uma redução do BREAKPOINT-MEDIAN$_3$ com restrição das instâncias não conterem dois breakpoints consecutivos, provado ser NP-completo em Teorema 1.

Seja π_1, π_2, π_3 entradas do BREAKPOINT-MEDIAN$_3$ que satisfaçam a restrição de não possuírem dois breakpoints consecutivos. Assuma, sem perda de generalidade, que T_1 seja uma árvore estrela associada com π_1, i.e., existe um nó raiz no qual todos os seus filhos são folhas e representam os elementos de π_1. Considerando os breakpoints de π_2 com relação a π_1, construa a seguinte árvore (veja Figura 2):

DOI: 10.59254/sbpo-2025-212286
http://live-proceedings.pantheonsite.io/p/212286

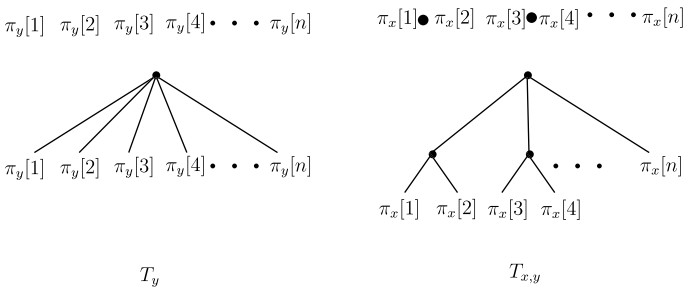


Figura 2: Na esquerda é apresentada a permutação π_y e sua árvore correspondenteT_y. Na direita é apresentada a permutação π_x com os seus breakpoints em relação a π_y, assim como a árvore $T_{x,y}$. Os pontos entre $\pi_x[1]$ e $\pi_x[2]$, e também $\pi_x[3]$ e $\pi_x[4]$ representam breakpoints entre π_x e π_y.

Crie uma árvore estrela enraizada em um nó v; Para cada breakpoint a b de π_2 com relação a π_1, crie um filho u de v de modo que u seja pai das folhas a e b; Chame essa árvore de $T_{2,1}$. Em seguida, aplique a mesma transformação entre os outros pares de permutações, obtendo as árvores $T_{3,1}$ e $T_{2,3}$. Por construção, cada árvore possui altura no máximo dois e representa os breakpoints de uma permutação com respeito a outra permutação e tem-se como resultado a criação de três árvores distintas.

Cada breakpoint de uma permutação implica em uma ramificação na árvore estrela que contém o par e, além disso, não existem dois breakpoints consecutivos entre as permutações da entrada. Isso implica que $d_{BP}(\pi_x, \pi_y) = d(T_{x,y}, T_y)$, visto que cada breakpoint a b deve subir na árvore para se tornarem filhos da raiz v como em T_y. Como a distância de breakpoint entre permutações continua igual ao observar as árvores obtidas, temos que TREE-MEDIAN$_3$ também é NP-completo.

□

Dessa forma, é demonstrada a relação entre problemas envolvendo MEDIANA e CLOSEST com a distância entre árvores, que não pode ser obtida considerando outras operações de reordenamento de genoma [Cunha et al., 2024a, 2020].

Teorema 3. TREE-CLOSEST$_3$ *é* NP*-completo.*

Demonstração. Essa prova é feita por meio de uma redução polinomial do TREE-MEDIAN$_3$.

Seja T_1, T_2, T_3 árvores para o TREE-MEDIAN$_3$ com n folhas em cada uma. Portanto, todos os elementos de 1 a n aparecem em todas as árvores. Primeiro, é apresentado um limite superior para a distância entre duas árvores.

Fato 3.1. *Seja h a altura de duas árvores T_x e T_y com n folhas. Logo, $d(T_x, T_y) \leq 2nh$.*

Demonstração. Para transformar T_x em T_y, para cada folha de T_x, deve no máximo subir a folha para ser filho da raiz r_x e, em seguida, descer para a posição correta como colocada em T_y. Portanto, no máximo, serão realizados $2h$ movimentos para cada uma das n folhas. □

Primeiramente, se obtém a árvore T'_x (veja Figura 3(i)) como apresentado a seguir. Seja r_x a raiz da árvore T_x com altura h_x, então:

1. Crie um grafo caminho com $2nh_x + 1$ nós, $p_x^1, \ldots, p_x^{2nh_x+1}$.

DOI: 10.59254/sbpo-2025-212286
http://live-proceedings.pantheonsite.io/p/212286

2. Para cada nó u do grafo caminho criado anteriormente, com exceção do último nó $p_x^{2nh_x+1}$, adicione um nó folha u' que seja filho do nó u.

3. Para cada folha $p_x^{i\,\prime}$, filho do nó p_x^i criado anteriormente, mude o seu rótulo como $i+n$, para $i = 1, \ldots, 2nh_x$.

4. Identifique os nós $p_x^{2nh_x+1}$ e r_x.

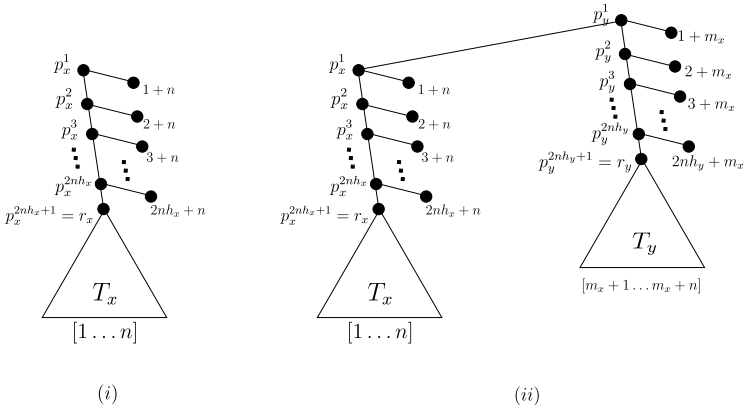


Figura 3: (i) Construção de T'_x da árvore T_x com n folhas e altura h_x. (ii) Árvore obtida pela operação de merge entre T'_x e T'_y, onde p_y^1 é a raiz da árvore $T'_{x,y}$.

Seja T'_x e T'_y árvores obtidas pelo procedimento anterior, crie a árvore $T'_{x,y}$ aplicando uma *operação de merge* entre T'_x e T'_y como definida a seguir (veja Figura 3(ii)).

Seja p_x^1 e p_y^1 as raízes de T'_x e T'_y, respectivamente, então:

1. Para cada folha i de T'_y, for $i = 1, \ldots, 2nh_y + 1$, mude seu rótulo para $i + m_x$, onde m_x é o número de folhas de T'_x, i.e. $m_x = 2nh_x + 1$.

2. Adicione uma aresta entre p_x^1 e p_y^1 de tal modo que p_x^1 seja filho de p_y^1.

A árvore resultante $T'_{x,y}$ contém todos os elementos de 1 até $2n(h_X + h_y) + 2$.

Baseado em uma instância geral do TREE-MEDIAN$_3$ construímos uma instância particular do TREE-MEDIAN$_3$ por uma série de operações de merge, e provamos uma equivalência entre as correspondentes soluções, como descrito em Lema 1.

Lema 1. *Seja T_1, T_2 e T_3 três árvores, a árvore T_σ é solução do* TREE-MEDIAN$_3$ *se e somente se a árvore $T'_{\sigma,\sigma,\sigma}$ for solução do* TREE-CLOSEST$_3$ *para as árvores $T'_{1,2,3}, T'_{2,3,1}$, e $T'_{3,1,2}$.*

Demonstração. Visto que as árvores $T'_{1,2,3}, T'_{2,3,1}$, e $T'_{3,1,2}$ são compostas por três partes da operação de merge e, além disso, que para cada árvore T_x foram adicionados mais de $2nh$ elementos, baseado no Fato 3.1, é menos custoso trabalhar com cada parte separadamente do que mover a árvore como um todo para outra posição. Portanto, cada parte pode ser tratada separadamente sem perda de otimalidade, e então existe uma solução do TREE-CLOSEST$_3$ onde todas as partes possuem a mesma solução T'_δ. Concluindo, existe uma árvore $x \in \{T'_{1,2,3}, T'_{2,3,1}, T'_{3,1,2}\}$ de tal modo que $d(T'_{\delta,\delta,\delta}, x) = d(T'_\delta, T'_1) + d(T'_\delta, T'_2) + d(T'_\delta, T'_3)$. Considerando que desejamos T'_δ de tal forma que $d(T'_{\delta,\delta,\delta}, x)$ seja mínimo. também desejamos δ de tal forma que $d(T'_\delta, T'_1) + d(T'_\delta, T'_2) + d(T'_\delta, T'_3)$ seja minimizado. Logo, isso só acontece se e somente se $T_\delta = T_\sigma$, onde T_σ é uma solução do TREE-MEDIAN$_3$. □

DOI: 10.59254/sbpo-2025-212286
http://live-proceedings.pantheonsite.io/p/212286

Portanto, Lema 1 conclui a prova.

□

4. Conclusão

Sequenciamento baseado em uma única célula é tecnicamente desafiador, apresenta alta quantidade de ruído e erros no processamento que podem gerar incertezas na estrutura da árvore filogenética. Soluções para esse problema geralmente envolvem certa forma de probabilidade e estatística. Nesse trabalho focamos em provar a complexidade dos problemas estudados, mostrando que todos eles são NP-completos.

Como trabalhos futuros, queremos explorar soluções envolvendo o uso de técnicas de inteligência artificial como *Hyperdimensuional Computing (HDC)* para a análise de padrões e sumarização das árvores de forma mais efetiva.

O potencial de *HDC* na análise de dados genômicos já foi demonstrado em trabalhos recentes. Por exemplo, GenieHD introduziu um design de hardware-software que melhora o *HDC* para paralelizar tarefas de busca de padrões, resultando em significativo speedups [Kim et al., 2020]. De forma similar, uma plataforma de análise de genoma hiperdimensional é proposta em [Chen et al., 2023], que mapeia as sequências do genoma em um espaço hiperdimensional, possibilitando uma eficiente correspondência das sequências por meio de pesquisas paralelas de similaridade.

Referências

Bader, M. (2011). The transposition median problem is NP-complete. *Theor. Comput. Sci.*, 412 (12-14):1099–1110.

Bryant, D. (1998). The complexity of the breakpoint median problem. *Centre de recherches mathematiques, Technical Repert*.

Caprara, A. (2003). The reversal median problem. *INFORMS J. Comput.*, 15(1):93–113.

Chen, H., Kim, Y., Sadredini, E., Gupta, S., Latapie, H., e Imani, M. (2023). Sparsity controllable hyperdimensional computing for genome sequence matching acceleration. In *Proceedings of the IFIP/IEEE International Conference on Very Large Scale Integration (VLSI-SoC)*, p. 56–61. IEEE.

Cunha, L., Sau, I., e Souza, U. (2024a). On the Complexity of the Median and Closest Permutation Problems. In *24th International Workshop on Algorithms in Bioinformatics (WABI 2024)*, volume 312, p. 2:1–2:23, Dagstuhl, Germany. ISBN 978-3-95977-340-9.

Cunha, L., Sau, I., e Souza, U. (2024b). On the parameterized complexity of the median and closest problems under some permutation metrics. *Algorithms for Molecular Biology*, 19(1):24.

Cunha, L. F. I., Feijão, P., dos Santos, V. F., Kowada, L. A. B., e de Figueiredo, C. M. (2020). On the computational complexity of closest genome problems. *Discrete Applied Mathematics*, 274: 26–34.

Cunha, L. F. I. e Protti, F. (2019). Genome rearrangements on multigenomic models: Applications of graph convexity problems. *J. Comput. Biol.*, 26(11):1214–1222.

Fertin, G., Labarre, A., Rusu, I., Vialette, S., e Tannier, E. (2009). *Combinatorics of Genome Rearrangements*. MIT press.

DOI: 10.59254/sbpo-2025-212286
http://live-proceedings.pantheonsite.io/p/212286

Gorbalenya, A. E. e Lauber, C. (2017). Phylogeny of viruses. *Reference Module in Biomedical Sciences*.

Haghighi, M. e Sankoff, D. (2012). Medians seek the corners, and other conjectures. In *BMC bioinformatics*, volume 13, p. 1–7. Springer.

Jahn, K., Kuipers, J., e Beerenwinkel, N. (2016). Tree inference for single-cell data. *Genome Biology*, 17(1):86.

Kim, Y., Imani, M., Moshiri, N., e Rosing, T. (2020). Geniehd: Efficient dna pattern matching accelerator using hyperdimensional computing. In *Proceedings of the IEEE/ACM Design Automation and Test in Europe Conference (DATE)*, p. 320–325. IEEE.

Kuipers, J., Jahn, K., e Beerenwinkel, N. (2017). Advances in understanding tumour evolution through single-cell sequencing. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, 1867 (2):127–138.

Lanctot, J. K., Li, M., Ma, B., Wang, S., e Zhang, L. (2003). Distinguishing string selection problems. *Inf. Comput.*, 185(1):41–55.

Luo, X. G., Kuipers, J., e Beerenwinkel, N. (2023). Joint inference of exclusivity patterns and recurrent trajectories from tumor mutation trees. *Nature Communications*, 14(1):3676.

McGranahan, N. e Swanton, C. (2015). Biological and therapeutic impact of intratumor heterogeneity in cancer evolution. *Cancer Cell*, 27(1):15–26.

Morita, K., Wang, F., Jahn, K., Hu, T., Tanaka, T., Sasaki, Y., Kuipers, J., Loghavi, S., Wang, S. A., Yan, Y., et al. (2020). Clonal evolution of acute myeloid leukemia revealed by high-throughput single-cell genomics. *Nature Communications*, 11(1):5327.

Nascimento, T. L. e Cunha, L. F. I. (2024). Determinando distâncias e consensos entre árvores. In *Anais do LVI SBPO*. Galoá.

Pe'er, I. e Shamir, R. (1998). The median problems for breakpoints are NP-complete. In *Elec. Colloq. on Comput. Complexity*, volume 71.

Popov, V. Y. (2007). Multiple genome rearrangement by swaps and by element duplications. *Theor. Comput. Sci.*, 385(1-3):115–126.

Watterson, G. A., Ewens, W. J., Hall, T. E., e Morgan, A. (1982). The chromosome inversion problem. *J. Theor. Biol.*, 99(1):1–7.

DOI: 10.59254/sbpo-2025-212286
http://live-proceedings.pantheonsite.io/p/212286