15º Congresso de Inovação, Ciência e Tecnologia do IFSP - 2024

Algoritmo para a identificação de conjuntos de genes mutuamente exclusivos em dados de mutação de câncer

SARA CASATTI QUEIROZ RAMOS ARIAS[1], JORGE FRANCISCO CUTIGI[2]

[1]Estudante do Técnico em Informática para Internet Integrado ao Ensino Médio, Bolsista PIBIFSP, IFSP, Câmpus São Carlos, c.sara@aluno.ifsp.edu.br.

[2]Professor de Computação, IFSP, Câmpus São Carlos, cutigi@ifsp.edu.br.

RESUMO: O câncer é uma doença complexa causada por mutações genéticas que se acumulam durante toda a vida do indivíduo. Essas mutações são alterações no DNA das células, podendo ocorrer de forma mutuamente exclusiva, que é quando a mutação de um gene exclui a de outro. Esse padrão de exclusividade mútua é um dos mais importantes da genômica do câncer. Por isso, a detecção desses genes mutuamente exclusivos é crucial para o melhor entendimento da doença e para a identificação de possíveis alvos terapêuticos, já que a ocorrência de dois ou mais genes mutados mutuamente exclusivos em uma célula impacta sua capacidade de sobrevivência. Atualmente, existem diversos métodos que identificam essas mutações, porém, são abordagens frequentemente baseadas em métodos probabilísticos complexos e de difícil interpretação. Nesse contexto, esse projeto teve como motivação investigar e propor alternativas para a identificação de conjuntos de genes mutuamente exclusivos, de forma a simplificar a análise dos dados genômicos de câncer. O método proposto utiliza grafos para representar a conexão entre os genes, e contribuir para a identificação de conjuntos de genes mutuamente exclusivos. Foi realizada uma validação por meio da comparação com diferentes abordagens, a qual mostrou resultados promissores apesar das limitações encontradas no método.

PALAVRAS-CHAVE: Bioinformática; Câncer; Exclusividade Mútua.

Investigation and proposal of an algorithm for identifying mutually exclusive gene sets in cancer mutation data

ABSTRACT: Cancer is a complex disease caused by genetic mutations that accumulate throughout an individual's life. These mutations are alterations in the DNA of cells and can occur in a mutually exclusive way, which is when the mutation of one gene excludes that of another. This pattern of mutual exclusivity is one of the most important in cancer genomics. Therefore, the detection of these mutually exclusive genes is crucial for a better understanding of the disease and for the identification of possible therapeutic targets, since the occurrence of two or more mutually exclusive mutated genes in a cell affects its ability to survive. There are currently several methods for identifying these mutations, but they are often based on complex probabilistic methods that are difficult to interpret. In this context, the motivation behind this project was to investigate and propose alternatives for identifying sets of mutually exclusive genes in order to simplify the analysis of cancer genomic data. The proposed method uses graphs to represent the connection between genes and to contribute to the identification of mutually exclusive gene sets. Validation was carried out by comparing different approaches, which showed promising results despite the limitations found in the method.

KEYWORDS: Bioinformatics; Cancer; Mutual Exclusivity.

INTRODUÇÃO

O padrão de exclusividade mútua está relacionado a um grupo de dois ou mais genes, em que os genes desse grupo raramente sofrem mutação no mesmo paciente, ou seja, mutações simultâneas desses genes são menos frequentes do

que o comumente esperado em um mesmo paciente. Isso acontece pois tais mutações podem ser consideradas "fortes". Ao ter mais de uma mutação nesse grupo de genes, a capacidade de sobrevivência da célula é diretamente impactada. Dessa forma, a identificação desses genes mutuamente exclusivos é de extrema importância, pois trata-se, também, da descoberta de genes que são potencialmente significativos para o câncer.

Existem diversos métodos que têm como objetivo identificar a exclusividade mútua em genes relacionados com o câncer. De acordo com Deng et. al (2019), esses métodos são divididos em dois grupos: métodos *De novo*, que utilizam apenas dados gnômicos para explorar padrões de exclusividade mútua não aleatória; métodos baseados em conhecimento, que integram outras informações, como redes de interação gênicas (DENG et al., 2019). Os métodos *De novo* são particularmente úteis quando se deseja descobrir padrões, associações ou informações que não são previamente conhecidas. Dessa forma, neste projeto, busca-se a investigação de métodos dessa categoria. Boa parte das abordagens *De novo* são comumente baseadas em métodos probabilísticos complexos e de difícil entendimento. Isso se deve à gigantesca quantidade de dados disponíveis, tornando o número total de combinações de genes também elevado. Além disso, algoritmos para exclusividade mútua precisam encontrar genes que, ao mesmo tempo, tenham alta exclusividade e ampla cobertura/alcance.

Nesse contexto, este projeto busca contribuir para a simplificação da análise de dados genômicos de câncer, facilitando a identificação de conjuntos de genes mutuamente exclusivos de maneira mais eficaz e eficiente. Para isso, foi utilizada uma abordagem que utiliza grafos para representar a conexão entre os genes, o que torna possível investigar a relação entre esses genes por meio de pontuações calculadas no algoritmo. Com esse grafo é possível encontrar conjuntos mutuamente exclusivos de formas mais simples contribuindo para uma melhor compreensão do método.

MATERIAIS E MÉTODOS

Na Figura 1 é apresentada uma visão geral do método proposto neste trabalho, o qual é composto por quatro passos, conforme detalhado a seguir.

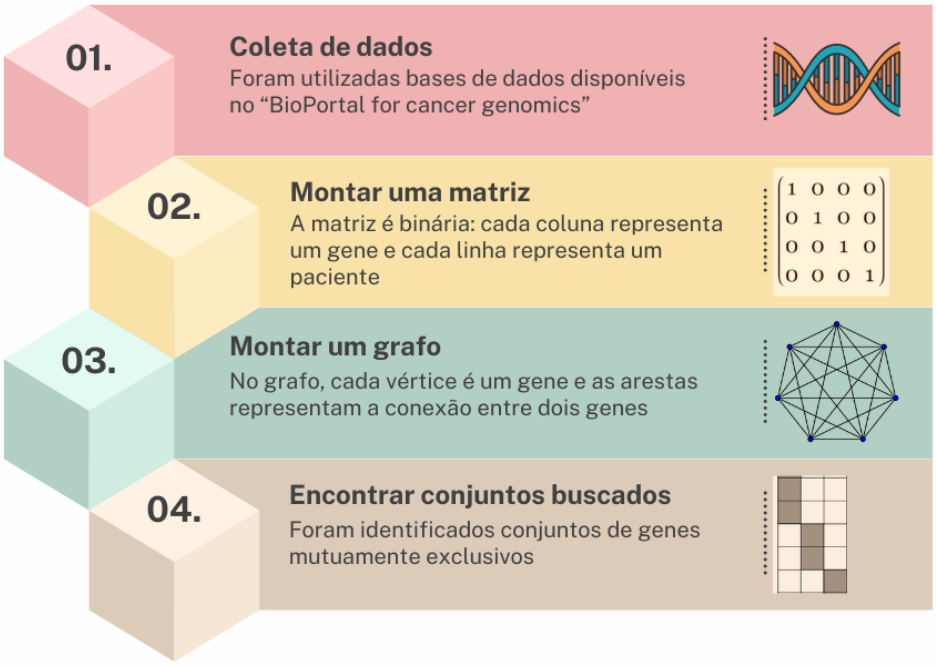

Figura 1: Ilustração dos principais passos empregados neste trabalho.

Passo 01 - Coleta de dados: Para a pesquisa foram utilizadas bases de dados disponíveis no *cBioPortal for cancer genomics* (CERAMI et al., 2012). No interior dessas bases de dados existe um arquivo MAF (Mutation Annotation Format), que contém as informações de todas as mutações observadas no estudo. Nesse arquivo, é possível localizar diver-

sas informações, mas para este projeto são necessárias apenas duas: Hugo_Symbol (genes) e Tumor_Sample_Barcode (pacientes), as quais mostram quais genes estão mutados em cada paciente.

Passo 02 - Montar uma matriz binária: Os dados selecionados formam uma tabela de duas colunas, que representam os pacientes e os genes. Essa tabela é transformada em uma matriz binária, ou seja, uma matriz composta por números um (1) e zero (0) como ilustrado na Figura 2, em que cada coluna representa um gene e cada linha representa um paciente. Se o valor de uma célula for zero (0) significa que o gene não está mutado no paciente; já se o valor for um (1), o gene está mutado naquele paciente.

Passo 03 - Montar um grafo: Para achar conjuntos de genes mutuamente exclusivos, a abordagem escolhida foi montar um grafo, em que cada nó é um gene e as arestas representam a conexão entre dois genes, como ilustrado na Figura 2. Considerando que, no padrão estudado, uma mutação exclui a outra, é preciso encontrar um conjunto de genes que não ocorrem simultaneamente em um mesmo paciente, ou seja, quanto menos conectados os genes forem (quando um gene é zero, o outro é um), melhor o conjunto é.

g1	g2	g3	g4	g5	g6	g7
1	0	1	0	0	0	1
1	0	0	1	0	1	1
0	1	1	0	0	1	0
0	1	1	0	0	1	1
0	1	0	0	1	0	0
1	0	0	0	1	1	0

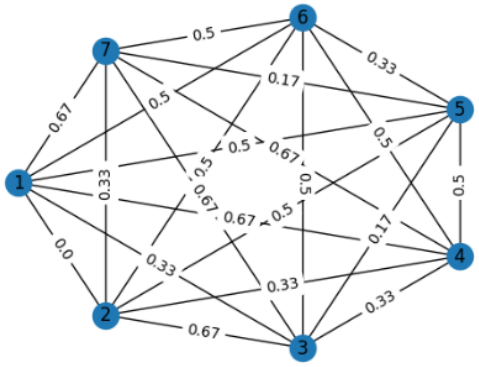

Figura 2: Matriz binária ao lado do grafo formado, com as respectivas pontuações em cada aresta.

Passo 04 - Encontrar conjuntos mutuamente exclusivos: Com a elaboração do grafo, foi possível tornar o algoritmo mais simples e de mais fácil interpretação, já que os dois principais problemas encontrados pelo método Dendrix (VANDIN; UPFAL; RAPHAEL, 2012) podem ser contornados pelo algoritmo proposto. Em relação ao problema de cobertura máxima em uma submatriz, ou seja, como achar conjuntos de genes com exclusividade alta e amplo alcance, propõe-se que seja resolvido por meio do cálculo de uma pontuação para cada dupla de gene.

Essa pontuação leva em conta a exclusividade mútua e a cobertura de cada conjunto, e ela é calculada ao comparar dois genes, quando os dois genes não se excluem, ou seja, são zero e zero ou um e um na mesma pessoa, é somado um a pontuação, ao iterar por todas as pessoas repetindo esse processo a pontuação final é dividida pela quantidade de pessoas, por exemplo na Figura 2 os genes quatro e cinco não se excluem 3 vezes, ao dividir esse valor pela quantidade de pacientes (6 pacientes) temos a pontuação 0,5.

Já em relação ao problema para achar subconjuntos com maior número de genes mutuamente exclusivos, com o algoritmo proposto é possível encontrar um caminho no grafo em que a soma das pontuações seja relevante para localizar um conjunto de vértices mutuamente exclusivo. Então, para encontrar esses conjuntos de genes foi usado o algoritmo de Dijkstra, que identifica o menor caminho entre um gene e todos os outros,ou seja, vão ser calculados os menores caminhos do gene um para todos os outros genes e então do gene dois para todos os outros e assim por diante até o último gene. Após o cálculo de todos os menores caminhos, foram escolhidos os conjuntos com as pontuações relevantes, possibilitando localizar os conjuntos mutualmente exclusivos, como os ilustrados na Figura 3.

RESULTADOS E DISCUSSÃO

Para analisar os resultados obtidos pelo algoritmo proposto, foram realizadas diversas comparações com outros métodos da literatura da área. Como mostrado na Tabela 1, o método proposto teve alguns resultados similares aos métodos Dendrix (VANDIN; UPFAL; RAPHAEL, 2012), Mult-Dendrix (LEISERSON et al., 2013), além de ter encontrado conjuntos de genes mutuamente exclusivos citados em (DENG et al., 2019). O método Dendrix (VANDIN; UPFAL; RAPHAEL, 2012) é uma abordagem *De novo* que utiliza um algoritmo guloso para tentar encontrar os melhores genes e um algoritmo probabilístico chamado Markov chain Monte Carlo (MCMC) que é utilizado para ajudar a achar conjuntos

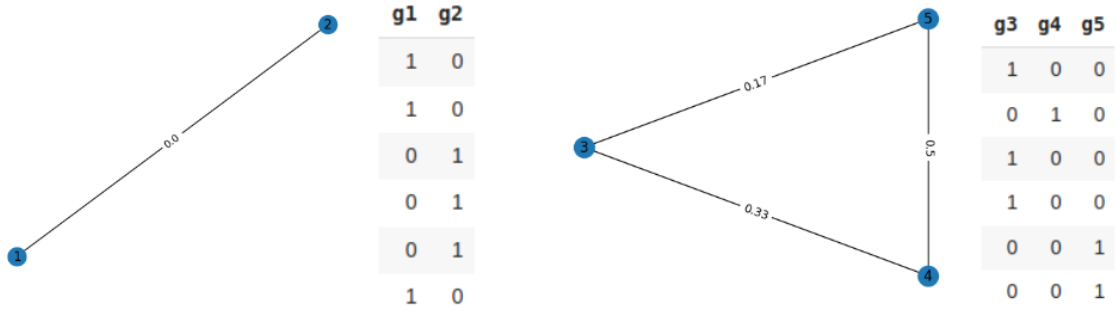

Figura 3: Conjuntos de genes mutuamente exclusivos encontrado pelo método, com as respectivas pontuações em cada aresta.

Tabela 1: Resultados encontrados com o novo algoritmo comparados com os obtidos por outros métodos da literatura.

Dados da literatura		Resultado		Dados coletados durante o projeto
Fonte id*	Tipo de câncer	Gene 1	Gene 2	Tipo de câncer
1	Câncer de pulmão	KRAS	TP53	Adenocarcinoma pulmonar
1	Adenocarcinoma de cólon	ATM	TP53	Adenocarcinoma pulmonar
2	Adenocarcinoma pulmonar	EGFR	KRAS	Adenocarcinoma pulmonar
2	Adenocarcinoma pulmonar	ATM	KRAS	Adenocarcinoma pulmonar
2	Diversos tipos de câncer	KRAS	NRAS	Diversos tipos de câncer
3	Glioblastoma	PTEN	PIK3R1	Glioblastoma
3	Glioblastoma	TP53	MDMA	Glioblastoma
3	Câncer de mama	PIK3CA	PIK3R1	Câncer de mama

1 - Identyfing mutual exclusivity accross cancer genomes : computacional approaches to discover genetic interaction and reveal tumor vulnerability. (DENG et al., 2019)
2 - (Dendrix) De novo discovery of mutated driver pathways in cancer. (VANDIN; UPFAL; RAPHAEL, 2012)
3 - (Mult-Dendrix) Simultaneous Identification of Multiple Driver Pathways in Cancer.

de genes significantes. Já o método Mult-Dendrix (LEISERSON et al., 2013) aprimora o Dendrix (VANDIN; UPFAL; RAPHAEL, 2012) ao possibilitar achar diversos conjuntos mutuamente exclusivos simultaneamente e encontrar maiores conjuntos mutuamente exclusivos.

Os resultados encontrados durante as comparações feitas entre esses métodos e o algoritmo proposto foram apenas duplas. Na Tabela 1 a coluna com os dados da literatura mostra a fonte dos resultados analisados e em que tipo de câncer esses resultados foram encontrados, já na coluna de dados coletados durante o projeto é informado em que tipo de câncer o algoritmo proposto encontrou os resultados, e a coluna com os resultados mostra a dupla de genes mutuamente exclusivos encontrados nos métodos da literatura e no algoritmo proposto.

Diante dos testes realizados, também foram identificadas duas principais limitações no algoritmo proposto: a primeira é que, quando os *data sets* são muito grandes, existe uma lentidão no processo de pontuação e na montagem do grafo, o que pode demorar alguns minutos; a segunda se refere à dificuldade em encontrar conjuntos mutuamente exclusivos com mais de dois genes. Essa segunda limitação se deve à maneira como a pontuação é calculada. Nota-se que, em dados reais, como as matrizes mutacionais possuem muitos algarismos zeros (0), as pontuações dos conjuntos ficam extremamente baixas, o que dificulta a localização de conjuntos de genes com uma pontuação relevante. Para lidar com essa segunda limitação, será necessário criar uma fórmula matemática que possibilite calcular a pontuação de forma que o alcance tenha uma prioridade menor. Isso porque a situação em que nenhum dos genes do conjunto é mutado (0, 0) é muito recorrente, o que reduz as pontuações dos conjuntos, acarretando em pontuações extremamente baixas. Apesar dessas limitações, foi possível obter resultados satisfatórios com o algoritmo, como mostrado na Tabela 1. Na maioria dos casos, foram utilizados *data sets* dos mesmos tipos de câncer, porém diferentes dos usados nas pesquisas referenciadas, o que também pode ter causado certa discrepância nos resultados obtidos. Além disso, o algoritmo proposto também encontrou outros conjuntos que não foram mencionados nos estudos citados, mas que também podem ser significativos para uma melhor compreensão da genômica do câncer.

CONCLUSÕES

Este trabalho tem como objetivo propor um novo método de identificação de genes mutuamente exclusivos, o qual possa ser mais eficiente, eficaz, mas também simples e interpretável. Com base no estudo realizado até o momento e aqui

apresentado, pode-se concluir que o método proposto consegue obter resultados satisfatórios, apesar de suas limitações.

Quando comparados com outros métodos mais complexos, os resultados adquiridos com o algoritmo criado neste trabalho possibilitaram descobrir conjuntos similares aos identificados nesses outros métodos. Além disso, o algoritmo desenvolvido não demanda cálculos complexos, utilizando algoritmos e estruturas já conhecidas, o que contribui para facilitar a compreensão do método.

Vale ressaltar que este trabalho ainda está em andamento e pretende-se realizar mais estudos a fim de reduzir as limitações identificadas.

CONTRIBUIÇÕES DOS AUTORES

S.C.Q.R.A desenvolveu e escreveu o trabalho. J.F.C orientou, revisou e contribuiu para a escrita do trabalho. Todos os autores contribuíram com a revisão do trabalho e aprovaram a versão submetida.

AGRADECIMENTOS

Ao Instituto Federal de Educação Ciência e Tecnologia de São Paulo (IFSP) pelo apoio financeiro por meio do Programa Institucional de Bolsas de Iniciação Científica e Tecnológica do IFSP (PIBIFSP), Edição 2024, Edital nº 34/2023.

REFERÊNCIAS

CERAMI, E. et al. The cbio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer discovery*, AACR, v. 2, n. 5, p. 401–404, 2012.

DENG, Y. et al. Identifying mutual exclusivity across cancer genomes: computational approaches to discover genetic interaction and reveal tumor vulnerability. *Briefings in Bioinformatics*, Oxford University Press, v. 20, n. 1, p. 254–266, 2019.

LEISERSON, M. D. et al. Simultaneous identification of multiple driver pathways in cancer. *PLoS computational biology*, Public Library of Science San Francisco, USA, v. 9, n. 5, p. e1003054, 2013.

VANDIN, F.; UPFAL, E.; RAPHAEL, B. J. De novo discovery of mutated driver pathways in cancer. *Genome research*, Cold Spring Harbor Lab, v. 22, n. 2, p. 375–385, 2012.