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

PROPOSTA DE MÉTODO COMPUTACIONAL PARA IDENTIFICAÇÃO DE CONJUNTO DE GENES RELACIONADOS E SIGNIFICATIVOS PARA O CÂNCER

THIAGO DE ALMEIDA MACIEL[1], JORGE FRANCISCO CUTIGI[2]

[1]Graduando em Bacharelado em Engenharia de Software, Bolsista PIBIFSP, IFSP, Câmpus São Carlos, maciel.thiago@aluno.ifsp.edu.br.

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

Área de conhecimento (Tabela CNPq): 1.03.03.00-6 Metodologia e Técnicas da Computação.

RESUMO: O câncer é uma doença complexa impulsionada por mutações genéticas que resultam na proliferação descontrolada de células. Nem todas as mutações gênicas desempenham um papel significativo no desenvolvimento do câncer, e identificar estes genes é um desafio, pois muitas mutações são apenas "passageiras", isto é, não têm impacto direto na progressão da doença. Além disso, nem sempre apenas um gene mutado causa o câncer, mas sim a interação de múltiplos genes e mutações que juntos alteram o comportamento celular. Por isso, a identificação de conjuntos de genes que atuam de forma coordenada é essencial para uma compreensão mais profunda dos mecanismos moleculares envolvidos no câncer. Neste trabalho foi proposto e desenvolvido um método computacional baseado em redes para identificar comunidades de genes significativos para o câncer, indo além da análise de genes individuais e considerando a interação entre múltiplos genes. O método foi aplicado a conjuntos de dados reais de câncer, demonstrando a capacidade em identificar comunidades gênicas potencialmente relevantes.

PALAVRAS-CHAVE: mutações genéticas; comportamento celular; método computacional; comunidades gênicas.

PROPOSAL OF A COMPUTATIONAL METHOD FOR IDENTIFYING RELATED AND SIGNIFICANT GENE COMMUNITIES IN CANCER

ABSTRACT: Cancer is a complex disease driven by genetic mutations that lead to uncontrolled cell proliferation. Not all genetic mutations play a significant role in cancer development, and identifying these critical genes is a challenge, as many mutations are merely "passengers,"meaning they do not have a direct impact on the disease's progression. Moreover, cancer is not always caused by a single mutated gene, but by the interaction of multiple genes and mutations that collectively alter cellular behavior. Therefore, identifying sets of genes that act in a coordinated manner is essential for a deeper understanding of the molecular mechanisms involved in cancer. In this work, a network-based computational method was proposed and developed to identify significant gene communities for cancer, going beyond the analysis of individual genes and considering the interaction between multiple genes. The method was applied to real cancer datasets, demonstrating its ability to identify potentially relevant gene communities.

KEYWORDS: genetic mutations; cellular behavior; computational method; gene communities.

INTRODUÇÃO

O câncer é uma doença multifatorial caracterizada por alterações genéticas que levam à proliferação descontrolada de células. As mutações gênicas desempenham um papel central na oncogênese e são geralmente classificadas em duas categorias: mutações *driver*, que conferem vantagem seletiva e impulsionam o desenvolvimento do tumor, e mutações *passenger*, que são secundárias e não afetam diretamente a progressão da doença. Um dos grandes desafios na pesquisa oncológica é identificar quais dessas mutações têm impacto significativo no câncer, visto que a maioria das mutações encontradas em células tumorais são *passenger*. Além disso, a análise isolada de genes individuais limita a compreensão da complexidade e das interações entre as mutações, dificultando a identificação de padrões funcionais que poderiam revelar novos *drivers* e caminhos moleculares relevantes para o câncer.

Entre os métodos disponíveis para identificar genes significativos, majoritariamente concentrados na análise de genes individuais, o *DiSCaGe (Discovering Significant Cancer Genes*) foi escolhido para este estudo devido à sua capacidade de gerar uma série de artefatos ao final de sua execução, como pontuações de mutações e análises de influência dentro de redes de interação gênica (CUTIGI et al., 2021). Porém, o *DiSCaGe* não fornece como resultado informações diretas sobre conjuntos de genes relacionados, o que representa uma limitação para a compreensão integrada da oncogênese.

Reconhecendo essa lacuna, o presente trabalho propõe o desenvolvimento de um método computacional que vai além das abordagens focadas em genes isolados, buscando identificar comunidades gênicas que atuam em conjunto e têm relevância para o câncer. A validação do método foi realizada utilizando dados genéticos reais de câncer de próstata e de ovário, demonstrando seu potencial em identificar agrupamentos de genes.

MATERIAIS E MÉTODOS

Neste estudo, busca-se identificar conjuntos de genes significativos para o câncer, utilizando dados provenientes de arquivos de mutações gênicas, redes gênicas e *drivers* canônicos. Para alcançar esse objetivo, foi desenvolvido o método *NRGC* (*Normalized Random Gene Clustering*). Essa abordagem é aplicada à Rede de Espalhamento de Força Gênica (*GSSN*), gerada como parte fundamental do processo de análise. A seguir, são detalhados os passos metodológicos e as estratégias empregadas. Na Figura 1, é apresentada uma visão geral do método utilizado no estudo, destacando as etapas de processamento dos dados e as abordagens específicas para a identificação de comunidades gênicas.

1. Coleta e Preparação dos Dados

Foram utilizados arquivos de múltiplas origens, sendo eles: 1) Arquivos com dados de mutações gênicas (*Mutation Annotation Format*): Contêm dados de mutações somáticas de amostras de tumores *Prostate Mutation* (HUANG; HE; MO, 2018); e *Ovarian Mutation* (NETWORK et al., 2011); 2) Redes gênicas: Arquivos que representam interações entre genes (FABREGAT et al., 2018); e 3) *Drivers* canônicos: Conjunto de genes conhecidos por serem relacionados ao câncer (REPANA et al., 2019), os quais são utilizados para validação do método proposto.

2. Geração da rede de espalhamento de força gênica (*GSSN*)

O passo subsequente envolve a criação da Rede de Espalhamento de Força Gênica (*GSSN*), que desempenha um papel central nos métodos desenvolvidos. A *GSSN* é uma rede que reflete a propagação

Figura 1: Ilustração dos principais passos do Método.

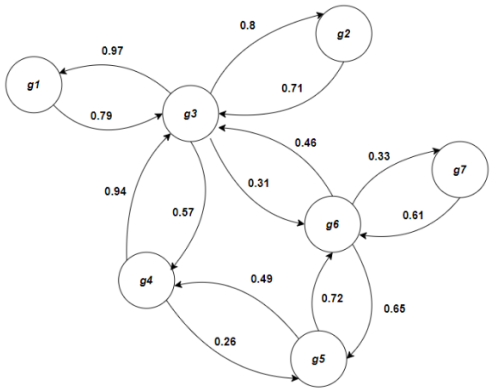

Figura 2: Representação da Rede de Espalhamento de Força Gênica (*GSSN*).

da influência das mutações entre genes em uma rede de consenso. Esta rede é construída a partir de diferentes matrizes e redes geradas durante o processo, como a Matriz de Mutação Ponderada (*WMM*), que quantifica as mutações em genes específicos de pacientes, e a Matriz de Frequência Ponderada e Normalizada (*NWF*), que avalia a contribuição relativa de cada gene. A *GSSN* resulta da consolidação da Rede de Genes Não Direcionada (*UGN*), na qual são atribuídos pesos às arestas baseados em valores normalizados, criando uma rede que reflete as interações gênicas mais significativas (CUTIGI et al., 2021). Essa rede é usada como base para a detecção de comunidades gênicas pelo método porposto. Na Figura 2, é apresentado um exemplo da *GSSN*. Os nós representam os genes e as arestas a influência entre eles, as quais possuem um valor atrelado que indica a força dessa influência.

3. Abordagens do método *NRGC*

O *Normalized Random Gene Clustering* (*NRGC*) foi o método desenvolvido para a detecção de comunidades de genes na GSSN. A abordagem do *NRGC* é baseada em um processo de expansão aleatória normalizada, descrito a seguir:

3.1. Seleção Aleatória do Nó Semente

No *NRGC*, o nó semente é escolhido aleatoriamente. A seleção aleatória do nó semente permite explorar múltiplos *clusters* potenciais, além de capturar diferentes estruturas de *clusters* que refletem a complexidade das interações gênicas. A partir da rede de espalhamento presente na Figura 2, a

seleção do nó semente feita pelo *NRGC* seria aleatória. Por exemplo, após a aleatorização realizada pelo método, qualquer gene poderia ser selecionado.

3.2. Expansão Aleatória Normalizada

A expansão do *cluster* é realizada com base em uma aleatoriedade calculada a partir dos pesos das arestas conectando o nó semente aos seus vizinhos. Esses pesos são normalizados para garantir que a aleatoriedade de inclusão de cada vizinho seja proporcional ao contexto local da rede. Isso significa que, se um gene tem várias conexões, o método garante que essas conexões sejam consideradas em relação umas às outras, equilibrando a contribuição de cada aresta. A normalização evita que arestas com pesos excepcionalmente altos dominem o processo de expansão, o que poderia resultar na formação de *clusters* artificiais e menos significativos biologicamente.

A escolha por uma abordagem aleatória se dá pela natureza estocástica dos processos biológicos, onde as interações entre genes e proteínas não seguem um comportamento determinístico simples. Em vez disso, essas interações são influenciadas por múltiplos fatores, incluindo a expressão gênica variável, o ambiente celular e a presença de mutações. Um método aleatório permite capturar essa variabilidade intrínseca, refletindo mais precisamente as dinâmicas reais das redes gênicas.

No caso da *GSSN* utilizada como exemplo na Figura 2, a começar de *g6*, o método analisa os vizinhos a partir da normalização dos pesos das arestas. Primeiro, *g5* pode ser adicionado ao *cluster*, já que o valor associado ao peso da aresta pode ser maior comparado aos outros vizinhos.

3.3. Condição de Parada da Expansão

A expansão do *cluster* é interrompida quando não há mais vizinhos não visitados que possam ser adicionados com base na aleatoriedade calculada. Essa abordagem aleatória garante que apenas as conexões mais fortes e significativas sejam consideradas, evitando a expansão excessiva e a criação de *clusters* irrelevantes. Além disso, essa estratégia simula a incerteza das interações biológicas, melhora a eficiência computacional e permite a exploração de diferentes configurações de *clusters*.

3.4. Heurísticas de Filtragem

Componente Principal da Rede: Foi utilizada a componente principal do grafo gerado para garantir que apenas a parte mais conectada e relevante da rede fosse considerada no processo de detecção de comunidades.

Filtragem por Coeficiente de *Clustering* e Precisão: O coeficiente de *clustering* é uma métrica utilizada para medir o grau de coesão em uma rede, ele avalia a densidade das conexões dentro de uma comunidade de genes, refletindo quão fortemente os genes estão interconectados. Um coeficiente de *clustering* alto sugere que os genes dentro da comunidade formam um grupo coeso, onde as interações são mais prováveis de acontecer. A precisão, por outro lado, refere-se à proporção de genes *driver* canônicos presentes em uma comunidade, em relação ao número total de genes nessa comunidade. Para melhorar a eficiência e relevância das comunidades identificadas, são aplicadas heurísticas que descartam *clusters* com coeficiente de *clustering* e precisão abaixo de um limiar definido (no caso, 0.20), assegurando a relevância biológica das comunidades preservadas.

3.5. Iteração e Randomização

O processo é repetido várias vezes, com a aleatorização da ordem dos nós, para garantir que diferentes configurações de *clusters* possam ser exploradas. Isso ajuda a capturar a diversidade da rede gênica e a identificar diferentes agrupamentos de relevância biológica.

3.6. Processamento dos Resultados

A implementação do método foi realizada em *Python* utilizando a biblioteca *NetworkX* para construção e manipulação de grafos, e *Pandas* para manipulação de dados. O algoritmo percorre a rede gênica, gerando comunidades de genes de acordo com o método *NRGC*. Cada comunidade gerada foi avaliada utilizando o coeficiente de *clustering*, que mede a densidade das conexões dentro da comunidade. Enquanto que a precisão foi calculada como a proporção de genes canônicos dentro de cada comunidade em relação ao total de genes na comunidade. Os resultados foram armazenados em arquivos de saída, contendo a listagem das comunidades ranqueadas de acordo com o coeficiente de *clustering* e a precisão em relação aos genes *driver* canônicos.

RESULTADOS E DISCUSSÃO

Os resultados obtidos a partir dos exemplos de câncer de próstata e câncer de ovário destacam a eficácia do método *NRGC* em identificar comunidades de genes potencialmente associados a esses tipos de câncer. Nas Tabelas 1 e 2, são apresentadas as 10 principais comunidades identificadas para os tipos de câncer de próstata e de ovário, respectivamente. Os resultados estão ordenados pelo coeficiente de *clustering* e precisão, e os genes já conhecidos como *drivers* na literatura estão destacados em negrito.

Tabela 1: Câncer de Próstata - Top 10 Comunidades

Posição	Comunidade	Coeficiente	Precisão (%)
1	KITLG **SOCS1 STAT5B**	0.66	0.67
2	CDC42 FYN GNAI2 **PIK3R2 RAC1** TAS2R4	0.53	0.33
3	**EGFR** EGR1 HSP90AA1 **MAPK1** PLA2G2D	0.51	0.4
4	DAP3 **MAPK1** MAPK3 MRPS23	0.51	0.25
5	ADTRP **RAC1** RELA **STAT3**	0.47	0.5
6	**CDH1** CTNNA1 EXOC4 EXOC6B	0.47	0.25
7	**AKT1** NFKB1 RAC2 RHEBL1	0.46	0.25
8	**FOXA1** HNF4G NR0B2	0.45	0.33
9	DUSP19 MAPK8 RAC2 **RHOA**	0.44	0.25
10	ACOX1 APOB **CEBPA FOXA1** RXRA	0.44	0.4

Tabela 2: Câncer de Ovário - Top 10 Comunidades

Posição	Comunidade	Coeficiente	Precisão (%)
1	**KRAS NRAS** RASAL1	0.71	0.67
2	JUND **PRKCB** TCF12 ZFP36L2	0.52	0.25
3	CD164 **CXCR4** GNG3 GPSM1	0.50	0.25
4	**HRAS** NTRK2 RAC3 **VAV1**	0.47	0.5
5	FOS **STAT3 SYK** TTPA	0.44	0.5
6	FOS **HRAS MAPK1** MAPK8 NFYB NUCKS1 SP1 **TP53**	0.43	0.38
7	CEBPB FOS NQO1 **SMAD3**	0.42	0.25
8	**JUN** MAPK14 **NRAS** SRA1	0.42	0.5
9	**MAPK1 RAD21 RUNX1T1** SPI1	0.39	0.75
10	FYN IBSP ITGA9 **SRC**	0.38	0.25

Os resultados mostram que o método *NRGC* foi capaz de identificar comunidades de genes altamente relevantes para os tipos de câncer de próstata e ovário. No caso do câncer de próstata, as

comunidades identificadas exibem coeficientes de *clustering* e precisão significativos, refletindo a capacidade do método em capturar relações biológicas importantes. Por exemplo, a comunidade composta por KITLG, SOCS1, STAT5B apresentou um coeficiente de 0.66 e precisão de 67%, destacando a relevância desses genes na oncogênese.

Para o câncer de ovário, as comunidades identificadas também mostram uma precisão relevante, embora os coeficientes de *clustering* variem mais amplamente, o que pode ser reflexo da complexidade biológica associada a este tipo de câncer. A comunidade KRAS, NRAS, RASAL1 apresentou um coeficiente de 0.71 e precisão de 67%, sugerindo a importância desses genes na patogênese do câncer de ovário.

Desse modo, é possível perceber que esses resultados reforçam a utilidade do método *NRGC* na identificação de comunidades gênicas significativas, podendo auxiliar na compreensão dos mecanismos moleculares do câncer e na identificação de potenciais alvos terapêuticos. A implementação completa do método *NRGC*, conforme descrito neste estudo, está disponível publicamente para a comunidade científica e demais interessados. O código fonte, juntamente com a documentação detalhada, pode ser acessado no repositório dedicado ao projeto: <https://github.com/ThiagoDeAM/GeCoDiM/>.

CONCLUSÕES

Este estudo propôs, desenvolveu, implementou e aplicou o método *NRGC* (*Normalized Random Gene Clustering*) para identificar comunidades gênicas em dados de câncer de próstata e ovário. Os resultados mostraram que o método foi eficaz em detectar grupos gênicos, com coeficientes de *clustering* e precisão relevantes, alinhando-se aos objetivos estabelecidos. Além disso, os resultados sugerem que o método pode ser aplicável a outros tipos de câncer ou doenças, desde que os dados estejam disponíveis.

Para o futuro, propõe-se o aprimoramento do método *NRGC*, incluindo novas heurísticas e a aplicação a diferentes tipos de dados genéticos e doenças. A integração de outras fontes de dados biológicos também é uma perspectiva importante para ampliar a aplicabilidade e o impacto do método.

CONTRIBUIÇÕES DOS AUTORES

T.A.M. contribuiu para o desenvolvimento e a escrita do projeto. J.F.C. contribuiu com a concepção e o escopo do estudo. Todos fizeram a revisão do trabalho e aprovaram a versão submetida.

REFERÊNCIAS

CUTIGI, J. F. et al. A computational approach for the discovery of significant cancer genes by weighted mutation and asymmetric spreading strength in networks. *Scientific reports*, Nature Publishing Group UK London, v. 11, n. 1, p. 23551, 2021.

FABREGAT, A. et al. The reactome pathway knowledgebase. *Nucleic acids research*, Oxford University Press, v. 46, n. D1, p. D649–D655, 2018.

HUANG, Z.-G.; HE, R.-Q.; MO, Z.-N. Prognostic value and potential function of splicing events in prostate adenocarcinoma. *International journal of oncology*, Spandidos Publications, v. 53, n. 6, p. 2473–2487, 2018.

NETWORK, C. G. A. R. et al. Integrated genomic analyses of ovarian carcinoma. *Nature*, NIH Public Access, v. 474, n. 7353, p. 609, 2011.

REPANA, D. et al. The network of cancer genes (ncg): a comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens. *Genome biology*, Springer, v. 20, p. 1–12, 2019.