13º Congresso de Inovação, Ciência e Tecnologia do IFSP - 2022

ANÁLISE COMPUTACIONAL DO PAPEL TOPOLÓGICO DE GENES SIGNIFICATIVOS PARA O CÂNCER EM REDES DE INTERAÇÃO GÊNICA

HELOISA ALVES DA SILVA[1], RODRIGO HENRIQUE RAMOS[2], JORGE FRANCISCO CUTIGI[3]

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

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

[3] 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 causada por mutações genéticas. O genes são segmentos de uma molécula de DNA que contém um código para produzir proteínas que desempenham funções vitais nas células. Os genes interagem fisicamente ou são funcionalmente relacionados, o que permite representar tais associações computacionalmente a partir do uso de redes complexas. Considerando que alguns genes são significativos para o câncer, e portanto relevantes para a compreensão da doença, e as redes de interação gênicas são amplamente utilizadas em métodos computacionais da Bioinformática do Câncer, é desejável identificar se genes significativos para o câncer possuem papel topológico distinto quando comparado com os demais genes em diferentes redes de interação. Para isso, extraiu-se medidas de centralidade de todos os genes. Tais medidas foram comparadas em diferentes redes e observou-se que algumas delas possuem valores mais altos para as medidas dos genes significativos para o câncer do que para os outros genes de cada respectiva rede.

PALAVRAS-CHAVE: bioinformática; câncer; redes complexas; medidas de centralidade; análise topológica.

COMPUTATIONAL ANALYSIS OF THE TOPOLOGICAL ROLE OF GENES SIGNIFICANCE FOR CANCER IN NETWORKS OF GENE INTERACTION

ABSTRACT: Cancer is a disease caused by genetic mutations. Genes are segments of a DNA molecule that contain code to produce proteins that perform vital functions in cells. The genes interact physically or are functionally related, which allows the representation of such associations computationally using complex networks. Considering that some genes are significant for cancer and therefore relevant to the understanding of the disease, and gene interaction networks are widely used in computational methods of Cancer Bioinformatics, it is desirable to identify whether genes significant for cancer have a distinct topological role when compared with the other genes in different interaction networks. For this, centrality measures were extracted from all genes. Such measures were compared in different networks, and it was observed that some of them have higher values for the measures of the genes significant for cancer than for the other genes of each respective network.

 ISSN: 2178-9959

KEYWORDS: bioinformatics; cancer; complex networks; centrality measures; topological analysis.

INTRODUÇÃO

O câncer está associado as mutações genéticas que acontecem ao longo da vida do indivíduo. Os genes codificam proteínas que atuam nas funções vitais das células do corpo humano. Entretanto não são todos os genes que sofrem mutações que estão relacionados ao câncer. Alguns genes, ao sofrerem mutações, se corrompem e codificam proteínas alteradas. Um conjunto de genes significativos para o câncer já conhecidos, fornecem informações importantes para a compreensão do câncer, esses genes são conhecidos como *drivers* (HABER; SETTLEMAN, 2007). Um gene não funciona sozinho, mas estabelece interações complexas com outros genes e com as proteínas que eles produzem.

Na Computação, as Redes Complexas fornecem uma representação natural de sistemas complexos, como é o caso do sistema celular (KIM; CHO; PRZYTYCKA, 2016). Nas redes de interação gênica, os genes são nós, e as arestas conectam genes que interagem fisicamente ou são funcionalmente relacionados (KIM; CHO; PRZYTYCKA, 2016).

Devido a importância dos genes significativos do câncer na compreensão da doença, foram desenvolvidas diversas abordagens para entender sua importância, a partir da extração de medidas de centralidade, tais como o grau, intermediação (*betweenness*), proximidade (*closeness*), entre outras. Essas medidas consideram aspectos distintos da estrutura e topologia da rede para caracterizar a importância de um nó, destacando assim seu papel central de acordo com cada uma delas (OLDHAM et al., 2019).

Com o intuito de caracterizar e comparar os genes significativos para o câncer com os demais, foram usados medidas de centralidade para estudar o papel topológico dos genes, e assim chegar a resultados quanto a sua importância ou centralidade em cada rede. Em geral, busca-se determinar a importância de cada vértice ou aresta em relação a sua posição na rede (MACHADO; BOERES, 2016). Com isso o objetivo geral desta pesquisa é analisar a topologia de genes significativos para o câncer em diferentes redes, por meio das medidas de centralidade, a fim de verificar se tais genes possuem papel topológico semelhante nessas redes.

MATERIAIS E MÉTODOS

A metodologia utilizada neste trabalho compreende uma análise da topologia dos genes de diferentes redes com uso de medidas de centralidade, a partir de programas computacionais. Como ilustra a Figura 1, o primeiro passo para o processo de pesquisa e análise foi coletar os dados das redes que posteriormente foram trabalhadas e a coleta dos genes *drivers*. O segundo passo foi a extração das medidas de centralidade dos genes em cada rede. O terceiro passo constituiu na análise dos resultados gerados em forma de gráficos.

Linguagem e bibliotecas: Para o desenvolvimento da pesquisa foi utilizada a linguagem de programação *Python*. Uma linguagem amplamente utilizada na Ciência de Dados e que possui diversas bibliotecas que auxiliam na pesquisa. Para o trabalho com redes utilizou-se o pacote *NetworkX* (HAGBERG; SWART; CHULT, 2008) que fornece estruturas de dados para representar diversos tipos de redes. O pacote pandas (MCKINNEY et al., 2010) também foi utilizado devido a sua eficiência com o processamento de dados. E para a criação de gráficos utilizou-se o pacote *matplotlib*(HUNTER, 2007).

Coleta de dados: Para a pesquisa, foi utilizada uma base de dados de proteínas Human Protein Reference Database (HPRD)(PERI et al., 2004), além dela também foi utilizada a rede de dados HINT

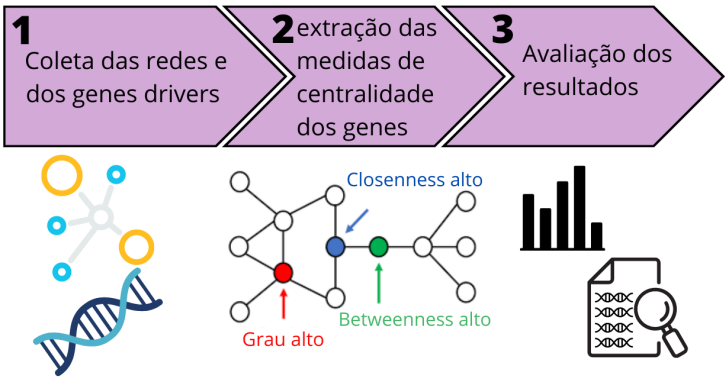

Figura 1: Processo geral.

(DAS; YU, 2012) e Reactome (FABREGAT et al., 2018). Utilizou-se também os *drivers* (HABER; SETTLEMAN, 2007) do câncer que são os atuais genes conhecidos e identificados como significativos para o câncer.

Medidas de centralidade: Para análise do papel topológico dos genes, foram utilizadas quatro medidas de rede:

Grau: Refere-se à quantidade de ligações de um nó. Em uma rede social por exemplo, uma medida que chama imediatamente a atenção é a quantidade de amigos que uma pessoa possui. A medida de centralidade grau reflete a ideia de que um nó importante está conectado com muitos nós (BORBA, 2013). No geral se interpreta que quanto mais conexões um nó possui, mais importante ele é, relativo a sua capacidade de mobilizar outros nós.

Closeness: Mede quanto cada vértice está próximo dos demais, ou seja, esta medida é dada pela distância geodésica total de um vértice a todos os outros da rede (DEL-VECCHIO et al., 2009). Em vários contextos, mais importante do que ter muitas conexões é não estar longe demais dos outros nós da rede, por isso, tem-se que um nó importante está próximo dos outros nós (BORBA, 2013).

Betweenness: Avalia quanto um vértice está no caminho geodésico entre dois outros vértices, isto é, analisa a importância do vértice na passagem de informação entre outros dois (DEL-VECCHIO et al., 2009). Está medida apresenta a ideia de que um nó importante faz parte de muitos caminhos (BORBA, 2013).

Clustering: Mede o quanto os nós se agrupam (BORBA, 2013). O coeficiente de *clustering* de um nó indica o percentual em que os vizinhos são interconectados.

RESULTADOS E DISCUSSÃO

A distribuição das medidas de centralidades é apresentada usando *Boxplots* (ou "diagramas de caixa"). As caixas em azul representam os *drivers* e as em laranja os não *drivers*. Por questões de visualização, os *outliers* foram removidos.

De acordo com a Figura 2, observa-se que nas três redes analisadas, HPRD, HINT e Reactome, Na medida centralidade grau, os resultados foram visualmente parecidos, observa-se que nas três redes os *drivers* tem destaque quanto ao seu valor. Nos três casos a mediana dos não *drivers* está significativamente abaixo da mediana dos *drivers*.

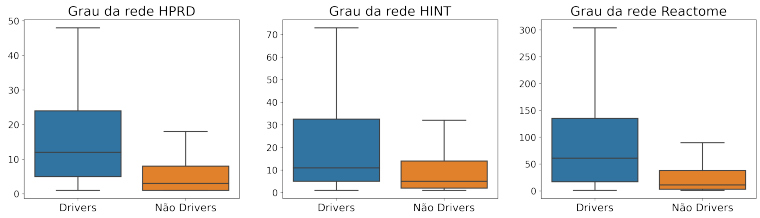

Figura 2: Grau das redes.

Na Figura 3 a mesma análise pode ser feita quanto a medida de centralidade *closeness*, onde os *drivers* tem maior valor de *closennes*.

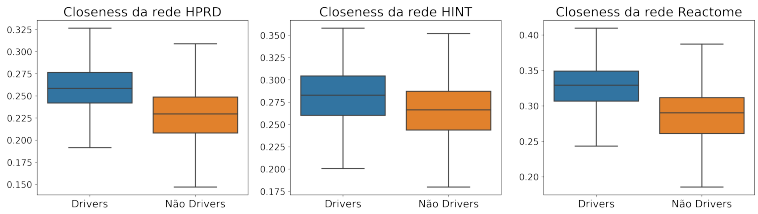

Figura 3: Closeness das redes.

O mesmo ocorre para a medida de centralidade *betweenness*, como pode ser observado na Figura 4, onde os não *drivers* tem um *betweenness* baixo, e no caso dos *drivers* observa-se que o terceiro quartil alcança valores altos e o limite superior também.

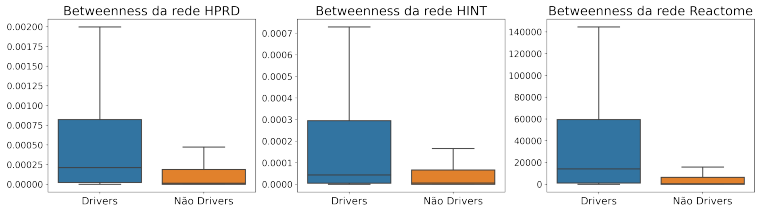

Figura 4: Betweennes das redes.

Já para a medida de centralidade *clustering*, que pode ser observada na Figura 5, o gráfico da rede HPRD é muito similar ao gráfico da rede Reactome, onde o valor da mediana dos não *drivers* é relativamente maior ao dos *drivers* das respectivas redes. No gráfico da rede HINT observa-se que ocorre o contrário em relação ao valor da mediana, que é mais alto no caso dos *drivers* e quase zero no caso dos não *drivers*.

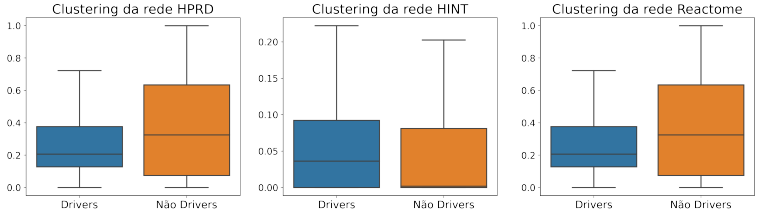

Figura 5: Clustering das redes.

CONCLUSÕES

Este trabalho apresentou uma analise do papel topológico de genes significativos para o câncer em três redes diferentes, utilizando de quatro medidas de centralidade. Com base no que foi apresentado,

pode-se concluir que os genes significativos para o câncer apresentam um papel topológico importante em relação aos outros genes. Nota-se que em três medidas diferentes e em três redes diferentes os *drivers* apresentaram valores superiores aos outros genes. Com isso conclui-se que os *drivers*, possuem maior grau, ou seja, são mais conectados que os não *drivers*, possuem também, maior *closeness*, ou seja, são genes que estão no centro da rede, e os resultados mostram também que os *drivers* tem maior *betweenness*, ou seja, possuem uma alta taxa de intermediação, indicando que nesses genes há um grande fluxo de informação sendo passada por eles. Em pesquisas futuras, pretende-se testar outras redes, além de testar outras medidas de centralidade e identificar o comportamento dos *drivers* nessas redes, afim de analisar os papeis topológicos dos genes com outras medidas e outras redes.

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 2022, Edital nº 39/2021.

REFERÊNCIAS

BORBA, E. M. Medidas de centralidade em grafos e aplicações em redes de dados. 2013.

DAS, J.; YU, H. Hint: High-quality protein interactomes and their applications in understanding human disease. *BMC systems biology*, BioMed Central, v. 6, n. 1, p. 1–12, 2012.

DEL-VECCHIO, R. R. et al. Medidas de centralidade da teoria dos grafos aplicada a fundos de ações no brasil. *XLI SBPO*, p. 1–4, 2009.

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

HABER, D. A.; SETTLEMAN, J. Drivers and passengers. *Nature*, Nature Publishing Group, v. 446, n. 7132, p. 145–146, 2007.

HAGBERG, A.; SWART, P.; CHULT, D. S. *Exploring network structure, dynamics, and function using NetworkX*. [S.l.], 2008.

HUNTER, J. D. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, IEEE COMPUTER SOC, v. 9, n. 3, p. 90–95, 2007.

KIM, Y.-A.; CHO, D.-Y.; PRZYTYCKA, T. M. Understanding genotype-phenotype effects in cancer via network approaches. *PLoS computational biology*, Public Library of Science San Francisco, CA USA, v. 12, n. 3, p. e1004747, 2016.

MACHADO, A. M.; BOERES, M. C. S. Aplicação de medidas de centralidade e análise da estrutura da rede brasileira de financiamento de campanha eleitoral de 2014. *XLVIII SBPO SIMPÓSIO BRASILEIRO DE PESQUISA OPERACIONAL. XLVII., Vitória. Anais... Vitória, ES*, 2016.

MCKINNEY, W. et al. Data structures for statistical computing in python. In: AUSTIN, TX. *Proceedings of the 9th Python in Science Conference*. [S.l.], 2010. v. 445, p. 51–56.

OLDHAM, S. et al. Consistency and differences between centrality measures across distinct classes of networks. *PloS one*, Public Library of Science San Francisco, CA USA, v. 14, n. 7, p. e0220061, 2019.

PERI, S. et al. Human protein reference database as a discovery resource for proteomics. *Nucleic acids research*, Oxford University Press, v. 32, n. suppl_1, p. D497–D501, 2004.