14º Congresso de Inovação, Ciência e Tecnologia do IFSP - 2023

Análise comparativa de métodos computacionais para identificação de genes significativos para o câncer em relação à combinação de redes de interação gênica

JHENNIFER STEFANY DA CRUZ MIRARCHI[1], JORGE FRANCISCO CUTIGI[2]

[1] Estudante do Técnico em Informática para Internet Integrado ao Ensino Médio, PIVICT, IFSP, Câmpus São Carlos, jhennifer.mirarchi@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 está relacionado a mutações genéticas que se desenvolvem ao longo da vida de um indivíduo. Essas mutações podem ocorrer nos genes, que são regiões específicas do DNA com um papel crucial na produção de proteínas essenciais para as funções celulares. Esses genes interagem fisicamente ou são funcionalmente relacionados. Essas interações podem ser representadas de maneira computacional usando redes complexas, conhecidas como redes de interação gênica. Essas redes desempenham um papel importante em métodos computacionais na área da Bioinformática do Câncer. Um exemplo é a investigação de genes relevantes para o câncer. A literatura científica contém uma série de redes disponíveis para serem exploradas em pesquisas. Cada uma dessas redes possui suas características próprias, e quando aplicadas em métodos para identificação de genes significativos para o câncer, podem resultar em resultados diferentes. Neste contexto, devido às informações únicas presentes em cada rede, há a possibilidade de combinar essas redes previamente, com o objetivo de alcançar um resultado de consenso. Além disso, uma abordagem alternativa seria a combinação dos resultados produzidos por cada rede. O objetivo deste trabalho é comparar o resultado de métodos de identificação de genes significativos para o câncer considerando dois cenários: a combinação de duas redes de interação gênica com a combinação de resultados individuais utilizando um método de agregação de ranking. Com isso espera-se ter um entendimento do impacto da combinação dos resultados e comparar as abordagens utilizadas.

PALAVRAS-CHAVE: Bioinformática; Câncer; Redes Complexas.

A comparative analysis of computational methods for identifying significant cancer-related genes regarding the combination of gene interaction networks.

Cancer is associated with genetic mutations that develop over an individual's lifetime. These mutations can occur in genes, which are specific regions of DNA with a crucial role in producing proteins essential for cellular functions. It's important to note that not all genes are directly linked to cancer. These genes physically interact or are functionally related through interactions. These interactions can be computationally represented using complex networks, known as gene interaction networks. These networks play a crucial role in computational methods in the field of Cancer Bioinformatics. An example is the investigation of genes relevant to cancer. Scientific literature contains a range of networks

available for exploration in research. Each of these networks has its own characteristics, and when applied in methods to identify significant cancer-related genes, they can yield different results. In this context, due to the unique information present in each network, there's the possibility of pre-combining these networks with the aim of achieving a consensus result. Additionally, an alternative approach would involve combining the results produced by each network. The objective of this work is to compare the outcome of identifying significant cancer-related genes considering the combination of a set of gene interaction networks. The goal is to compare the result considering combined networks prior to method execution or aggregation of results post-execution. This is aimed at gaining an understanding of the impact of result combination and comparing the used approaches.

KEYWORDS: Bioinformatics; Cancer; Complex Networks.

INTRODUÇÃO

O câncer é uma doença que está associada a mutações genéticas que acontecem ao longo da vida do indivíduo. Devido a sua complexidade, o câncer é um dos maiores desafios enfrentados pela medicina e pela sociedade como um todo. Ele é causado por combinações de mutações genéticas que afetam genes envolvidos no controle de diversas funções celulares.

Nessa perspectiva, atualmente a computação desempenha um papel essencial no entendimento do câncer, como a identificação de genes significativos para a iniciação e progressão da doença. Os genes não atuam de forma isolada, mas estabelecem interações complexas com outros genes e as proteínas que eles produzem. Na Bioinformática, essas interações podem ser representadas em redes complexas, também conhecidas como redes de interação gênica, que refletem a complexidade dos sistemas biológicos.

Nesse contexto, muitos bancos de dados são fontes de informação sobre redes de interação gênica, por exemplo, Human Protein Reference Database (HPRD) (PERI et al., 2004) e Reactome Functional Interactions (Reactome)(JASSAL et al., 2020). Cada rede tem suas próprias características. Quando usadas em métodos computacionais para identificar genes significativos para o câncer, os resultados podem variar. Um desses métodos é o nCOP (Network-Constrained Pathway Optimization)(HRISTOV; SINGH, 2017), que utiliza redes biológicas e dados de mutação para identificar caminhos biológicos relevantes ou processos específicos que podem estar envolvidos em fenômenos como doenças ou respostas celulares. Neste cenário, a combinação prévia de redes oferece a oportunidade de alcançar um resultado de consenso, enquanto outra alternativa seria a agregação dos resultados das redes individuais após a execução no método.

Neste estudo, é apresentada uma abordagem destinada a compreensão entre os resultados por meio da aplicação do método de identificação de genes significativos para o câncer. O propósito é abordar a seguinte questão de pesquisa: Qual será a similaridade entre os resultados obtidos ao unir as redes individuais para formar uma rede combinada e submetê-la ao método, em comparação com uma estratégia de agregação de ranking?

MATERIAIS E MÉTODOS

A metodologia utilizada neste trabalho abrange a seleção e a coleta de redes de interação gênica e dados de mutação. Além disso, inclui a escolha de métodos para identificar genes significativos relacionados ao câncer e métodos de agregação de ranking. Na Figura 1 é apresentado a visão geral do método empregado neste trabalho, o qual é composto por cinco passos bem definidos: No Passo 1 são coletadas bases de dados públicas contendo informações de redes de interação gênica e dados de mutação de

pacientes com um tipo específico de câncer. O Passo 2 consiste na combinação das redes selecionadas anteriormente por meio do desenvolvimento de algoritmos específicos com o propósito de realizar essa integração. No Passo 3, é utilizado um método computacional nas redes de interação gênica previamente escolhidas. Adicionalmente, o método foi reexecutado utilizando a rede combinada, resultado da união das redes previamente selecionadas, juntamente com os dados de mutação provenientes dos pacientes que possuem determinado tipo de câncer. A partir desse ponto, conforme ilustrado no Passo 4, os resultados gerados pela execução das redes individuais foram combinados usando o método de ranking. No quinto passo, foi realizada uma análise comparativa entre a *Agregação dos resultados* das redes individuais após a execução do método de identificação de genes significativos para o câncer e o resultado da *Rede combinada*, que foi submetida ao mesmo método após ser combinada.

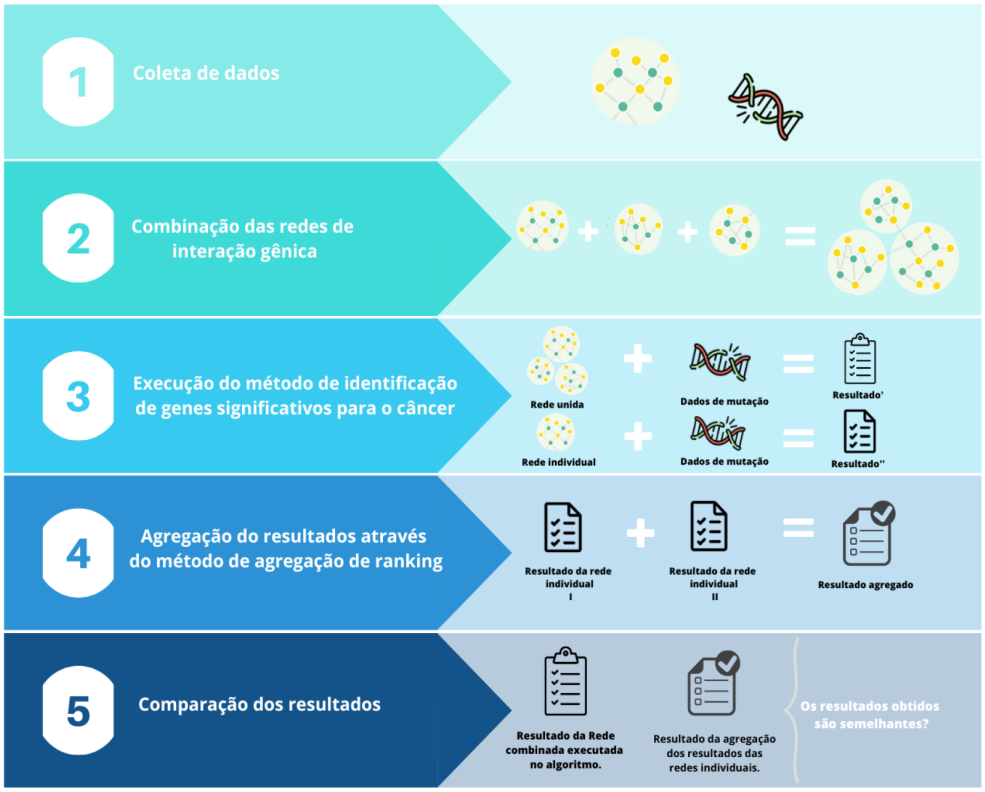

Figura 1: Ilustração dos principais passos do projeto.

As subseções seguintes descreve detalhadamente cada passo do método:

Passo 1 – Coleta de Dados: Para a pesquisa foram utilizadas base de dados de redes de interação proteína-proteína. As redes usadas foram: Human Protein Reference Database (HPRD)(PERI et al., 2004) e Reactome (FABREGAT et al., 2018). Adicionalmente, foram utilizados dados de mutações de pacientes portadores de um subtipo específico de câncer renal, conhecido como carcinoma de células renais (KIRC) (GRAY; HARRIS, 2019). Esses dados foram disponibilizados em conjunto com os arquivos do método utilizado neste projeto, o nCOP (HRISTOV; SINGH, 2017).

Passo 2 – Combinação das redes: Nesta fase, a combinação das redes foi conduzida utilizando a linguagem de programação Python, acompanhada por bibliotecas amplamente utilizadas nos campos de Bioinformática do Câncer e Redes Complexas, tais como o Pandas (MCKINNEY et al., 2010) e o NetworkX(HAGBERG; SWART; CHULT, 2008). O algoritmo desenvolvido para unir as redes tem como objetivo combinar duas redes de interação proteína-proteína em uma rede única, garantindo que todos os nós (proteínas) e todas as interações (arestas) das redes originais estejam presentes na rede

resultante. Nessa ideia, o algoritmo percorre as redes e se uma proteína ou uma interação não estiverem contidas em ambas as redes, elas são acrescentadas ao resultado final. Esse processo é repetido para todas as proteínas e interações. Ao final, como arquivo de saída, é gerada uma *Rede Combinada*, que engloba todas as interações das duas redes originais. Por exemplo: Considerando que a primeira rede possua as proteínas e interações ('A', 'B') e a segunda rede ('B', 'C'). Ao final, os nós presentes na rede combinada será: ['A', 'B', 'C'] e as interações da rede combinada será: [('A', 'B'), ('B', 'C')]. Com o objetivo de atingir um resultado de consenso, as redes *HPRD* (PERI et al., 2004) (com 9465 genes e 37080 interações) e *Reactome* (FABREGAT et al., 2018) (com 14058 genes e 268856 interações) foram unidas. O resultado da execução desse processo é uma *Rede combinada* que apresenta 15432 genes e 289159 interações.

Passo 3 – Execução do método de identificação de genes significativos para o câncer: Neste passo, procedeu-se à execução da rede combinada, bem como das redes individuais HPRD(PERI et al., 2004) e Reactome(FABREGAT et al., 2018), por meio do método computacional que envolve a identificação de genes significativos para o câncer. Nesse sentido, o método empregado para a realização da pesquisa foi o nCOP (Network-Constrained Pathway Optimization)(HRISTOV; SINGH, 2017). Se trata de um método que utiliza uma rede biológica e dados de mutações como arquivos de entrada para gerar caminhos de mutações, retornando uma lista como arquivo de saída com a classificação dos genes. É uma ferramenta que considera perfis mutacionais por indivíduo dentro do contexto de redes de interação proteína-proteína, a fim de identificar pequenas sub-redes conectadas de genes que, embora não sofram mutações individuais, compreendem vias que são alteradas através (isto é, "cobrem") uma grande fração de indivíduos(HRISTOV; SINGH, 2017). Nesse contexto, após o processo da execução da rede combinada no método, o algoritmo retornou uma lista como arquivo de saída, aquilo que denominamos de "Rede combinada", a união das duas redes.

Passo 4 – Agregação dos resultados: Após concluir a execução do método nas redes individuais e armazenar os resultados, avançou-se para a união desses resultados. Para isso, utilizou-se o Método de Borda (PIHUR; DATTA; DATTA, 2009), o qual foi desenvolvido utilizando a linguagem de programação Python com a biblioteca Pandas (MCKINNEY et al., 2010), após a execução, o resultado obtido foi denominado de "Agregação dos resultados". O método de agregação de Borda atribui pontos a cada posição em que determinado gene se encontra e soma todos os pontos de todas as listas. O gene com maior pontuação é colocado no topo da lista agregada, e assim, sucessivamente. De maneira geral, a lista agregada final é baseada na maior quantidade de vezes entre pares de genes encontrados em listas individuais. Por exemplo, se o gene "A" for classificado acima do gene "B" com maior frequência, então o gene "A" também deve ser classificado acima do gene "B" na lista final.

Passo 5 – Comparação dos resultados: No quinto passo, os resultados provenientes da agregação dos resultados usando o método de agregação de Borda(PIHUR; DATTA; DATTA, 2009), juntamente com os resultados obtidos após a execução da rede combinada no método, foram armazenados para a comparação. Na realização desta comparação, empregou-se representações visuais como o Diagrama de Venn e o Mapa de calor*(Heatmap)*, os quais foram analisados na seção intitulada "Resultados e Discussão".

RESULTADOS E DISCUSSÃO

A comparação entre os resultados da *Rede Combinada* e da *Agregação dos Resultados* é exibida utilizando uma série de diagramas. Para garantir uma análise eficaz, a observação dos gráficos foram

limitadas aos primeiros quarenta genes presentes nas listas de resultados. Essa abordagem facilita a visualização e a análise dos resultados nos gráficos.

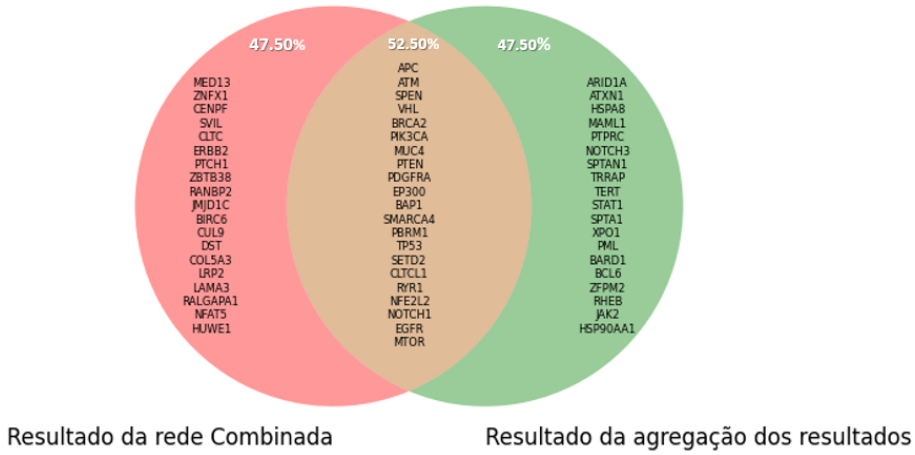

Figura 2: Diagrama de Venn comparativo que ilustra a porcentagem e os genes encontrados.

Na Figura 2, o Diagrama de Venn exibe informações tanto visuais quanto numéricas, com o objetivo de comparar os dois conjuntos de resultados, destacando a presença dos genes em ambas as situações. Ao comparar as duas abordagens, é evidente que o resultado da *Rede Combinada*, conforme ilustrado na Figura 2, engloba 47.50% dos genes, assim como a *Agregação dos Resultados*. Já a sobreposição das redes, ou seja, os genes que são compartilhados pelas duas abordagens, representa 52.50% dos genes, sinalizando uma proporção significativa de genes compartilhados entre as duas listas de resultados. Além disso, é possível obter uma visão abrangente da comparação pela visualização dos genes que são encontrados na intersecção. Nesse contexto, observa-se que existem 21 genes presentes em ambos os conjuntos de resultados, os quais fazem parte do total de quarenta genes presentes tanto na *Rede Combinada* quanto na *Agregação dos Resultados*.

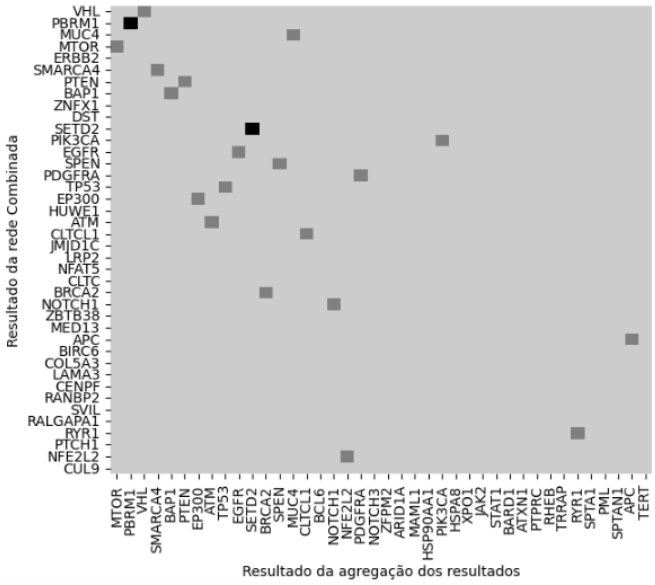

Figura 3: Mapa de calor (Heatmap) que ilustra a sobreposição de genes entre as duas abordagens.

Na Figura 3, é empregado um gráfico do tipo *Heatmap* para comparar os resultados sob uma perspectiva da ordem classificação dos genes no ranking. Esse gráfico ilustra a ocorrência dos genes,

sendo as células pintadas de acordo com a presença do gene em ambas as listas. Essa abordagem oferece uma representação visual da distribuição dos genes ao considerar as duas listas de resultados. Os pontos de interseção no gráfico denotam os genes que ocupam a mesma posição, como é o caso de PBRM1 e SETD2. Dessa maneira, em um cenário hipotético no qual as duas listas fossem idênticas, o gráfico apresentaria uma linha diagonal pintada. Neste cenário, há outros pontos que não se superpõem, correspondentes a genes que se afastam da diagonal. Isso sugere que, embora esses genes estejam presentes em ambas as listas, suas posições diferem entre as duas abordagens.

CONCLUSÕES

Este trabalho propôs responder a seguinte questão de pesquisa: "Qual será a similaridade entre os resultados obtidos ao unir as redes individuais para formar uma rede combinada e submetê-la ao método, em comparação com a estratégia de agregação de ranking?" No âmbito desta análise, ao analisar os gráficos destacados na seção "Resultados e Discussão", fica evidente que os resultados obtidos ao sujeitar a rede combinada ao método de identificação de genes significativos para o câncer, em comparação com a estratégia de agregação dos resultados das redes individuais usando o método de agregação de Borda, não apresentam semelhanças consideráveis. Quando se trata da presença de genes em ambas as abordagens, foi observada uma concordância de 52,50%, o que corresponde a 21 genes dos 40 analisados em cada conjunto de resultados, assim como demonstra a figura 2. Em outras palavras, mais da metade dos genes gerados em ambas as abordagens foram identificados. No entanto, uma observação adicional, conforme ilustrado na figura 3, revela que ao analisar a diagonal da figura, apenas dois genes ocuparam a mesma posição. Isso sugere que apesar de 52,50% dos genes estarem presentes em ambas as listas, somente dois deles coincidem na mesma posição. Esses resultados indicam que, dentro deste contexto, não há semelhanças significativas ou diferenças marcantes entre as duas abordagens.

CONTRIBUIÇÕES DOS AUTORES

J.S.C.M contribuiu para o desenvolvimento e a escrita do projeto. J.F.C orientou o trabalho, contribuiu com a concepção e escopo do estudo. Todos os autores contribuíram com a revisão do trabalho e aprovaram a versão submetida.

REFERÊNCIAS

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

GRAY, R. E.; HARRIS, G. T. Renal cell carcinoma: diagnosis and management. *American family physician*, v. 99, n. 3, p. 179–184, 2019.

HAGBERG, A.; SWART, P.; CHULT, D. S. Exploring network structure, dynamics, and function using networkx. 2008.

HRISTOV, B. H.; SINGH, M. Network-based coverage of mutational profiles reveals cancer genes. *Cell systems*, Elsevier, v. 5, n. 3, p. 221–229, 2017.

JASSAL, B. et al. The reactome pathway knowledgebase. *Nucleic acids research*, Oxford University Press, v. 48, n. D1, p. D498–D503, 2020.

MCKINNEY, W. et al. Data structures for statistical computing in python. v. 445, n. 1, p. 51–56, 2010.

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.

PIHUR, V.; DATTA, S.; DATTA, S. Rankaggreg, an r package for weighted rank aggregation. *BMC bioinformatics*, Springer, v. 10, p. 1–10, 2009.