Abordagens computacionais para a descoberta de genes significativos para o câncer

Jorge Francisco Cutigi[1,2], **Adriane Feijó Evangelista**[3], **Adenilso da Silva Simão**[2]

[1]Instituto Federal de Educação, Ciência e Tecnologia de São Paulo – IFSP São Carlos
São Carlos – SP – Brasil

[2]Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo
São Carlos – SP – Brasil

[3]Centro de Pesquisa em Oncologia Molecular – Hospital de Câncer de Barretos
Barretos – SP – Brasil

cutigi@ifsp.edu.br, adriane.feijo@gmail.com, adenilso@icmc.usp.br

Resumo. *O câncer é uma doença complexa provocada por alterações genéticas que se acumulam por toda a vida do indivíduo. A essas alterações dá-se o nome de mutação genética. Células de câncer possuem um elevado número de mutações, das quais um pequeno número delas é significativo para o câncer. A identificação de genes significativamente mutados, isto é, genes com mutações significativas, é essencial para a compreensão dos mecanismos de iniciação e progressão do câncer. Essa tarefa é um desafio chave na genômica do câncer, uma vez que estudos mostram que genes significativos podem sofrer mutação em uma frequência muito baixa. Com o sequenciamento de nova geração, uma extensa quantidade de conjuntos de dados genômicos foram gerados, criando o desafio de analisar e interpretar esses dados. Para identificar genes significativos ao câncer, redes de interação gênica combinadas com dados de mutação têm sido exploradas. Neste contexto, esta pesquisa de doutorado buscou a descoberta de genes significativos para o câncer por meio da proposição de abordagens computacionais para tal fim. O genes são identificados por um método baseado em redes que combina frequência de mutação ponderada e influência de vizinhos na rede, e possíveis falso-positivos são detectados por método baseado em aprendizado de máquina, o qual utiliza-se de dados de mutação e redes de interação gênica para induzir modelos preditivos. Um estudo experimental conduzido com seis tipos de câncer revelou o potencial das abordagens na descoberta de genes já conhecidos e de possíveis novos genes significativos para o câncer.*

1. Introdução

O câncer é uma das principais causas de morte no mundo. Trata-se de uma doença causada pelo acúmulo de alterações genéticas que acontecem durante a vida de um indivíduo. Tais alterações são chamadas mutações genéticas, e são causadas por diversos fatores, que podem ser internos ao organismo (por exemplo, falha na divisão celular) ou externos (por exemplo, exposição excessiva ao sol). As mutações podem resultar em um crescimento desordenado de células, que invadem tecidos e órgãos, causando o câncer [Stratton 2009].

Mutações genéticas no câncer têm sido estudadas há muito tempo por meio do sequenciamento de DNA/RNA, e um grande número de mutações recorrentes foi identificado [Vogelstein et al. 2013]. Novas tecnologias de sequenciamento de genoma, chamadas *Next-Generation Sequencing* (NGS), permitem sequenciamento genômico rápido e econômico, bem como a geração de um grande volume de dados biológicos em curto espaço de tempo. Tais dados auxiliam no estudo e análise de alterações genéticas em muitas doenças, incluindo câncer. No entanto, com a abundância de dados genômicos tem-se a dificuldade de se processar e a extração de informações clinicamente úteis. Nesse sentido, a Bioinformática desenvolve e utiliza métodos e técnicas computacionais para a interpretação de dados, obtendo informações e fornecendo subsídios para profissionais de saúde e pesquisadores.

Uma das categorias de métodos computacionais inclui aqueles que visam identificar mutações significativas (ou mutações *drivers*) e seus genes associados (ou genes *drivers*) para o desenvolvimento do câncer. Uma célula com câncer pode apresentar dois tipos de mutações: 1) mutações *passengers*, que não alteram o comportamento da célula, e 2) mutações *drivers*, que alteram o comportamento celular e são responsáveis pelo desenvolvimento do câncer, ou seja, fornecem às células vantagem seletiva em relação às demais. A identificação de mutações *drivers* e seus genes associados é um dos desafios mais significativos na área de Genômica do Câncer [Hou and Ma 2013, Raphael et al. 2014].

Nos últimos anos, vários métodos computacionais para identificação de genes significativamente mutados em câncer foram foram propostos [Hou and Ma 2013, Raphael et al. 2014, Cheng et al. 2015, Dimitrakopoulos and Beerenwinkel 2017, Cutigi et al. 2020a]. Esses métodos adotam estratégias diversas para descobrir genes *drivers*, como por exemplo, a análise de redes de interação gênica. Tal análise é uma fonte de informação muito importante, pois genes afetados por mutações *drivers* tendem a participar de atividades biológicas em comum [Ozturk et al. 2018]. Além disso, mutações significativas podem alterar o gene mutado e as vias que o gene participa [Vogelstein et al. 2013].

2. Motivação

O conhecimento sobre os genes que causam o início e a progressão do câncer é um ponto chave na Genômica do Câncer. Tanto o diagnóstico quanto o tratamento do câncer poderiam mais assertivos se os genes mutados de uma célula com câncer fossem conhecidos, no sentido de personalizar o tratamento de um determinado paciente. Nesse sentido, os tumores seriam identificados e caracterizados, possibilitando o tratamento mais adequado e personalizado [Vincent 2017]. A identificação de genes significativos para o câncer pode ser suportada por métodos computacionais baseados em diversos tipos de dados atualmente disponíveis, como dados de mutação e redes de interação gênica. Embora os métodos computacionais sejam utilizados para a identificação de genes significativos para o câncer, eles podem classificar erroneamente alguns genes como significativos, exigindo, portanto, curadoria especializada para filtrar seus achados [Bailey et al. 2018]. Tal erro de classificação é devido a alguns genes (referidos como *drivers* falso-positivos, ou falso *drivers*) exibindo características de serem significativos para o câncer, apesar de não estarem realmente envolvidos em sua iniciação e progressão.

3. Objetivos

Objetivo geral da tese foi descobrir genes significativos para o câncer com o uso de duas abordagens computacionais. O objetivo é baseado nas hipóteses de que genes significativamente mutados em câncer podem ser descobertos por meio da combinação de frequência ponderada de mutação e influência de vizinhos em redes, e possíveis falso-positivos podem ser detectados com o uso de dados de mutação e redes de interação gênica. A mutação ponderada, extraída dos dados de mutação, é baseada no impacto funcional de cada tipo de mutação na célula. A influência dos vizinhos, extraída das redes de interação gênica, é obtida a partir de medidas de força de espalhamento assimétrica entre todos os pares de nós, que levam em consideração vizinhos diretos e indiretos na rede. Tal proposta é baseada na conhecida e clássica hipótese local [Barabási et al. 2011] de que genes envolvidos no câncer tendem a interagir entre si.

4. Abordagens computacionais propostas

Para atingir os objetivos da pequisa, duas aborgagens computacionais foram propostas: 1) DiSCaGe (**Di**scovering **S**significant **Ca**ncer **Ge**nes), que prioriza genes significativos para o câncer diretamente relacionados ao impacto de diferentes tipos de mutação e interações gênicas; e 2) DFDriver (**D**etecting **F**alse **Driver**), que classifica genes supostamente *drivers* como *drivers* reais ou *drivers* falso-positivos com base em dados de mutação e redes de interação gênica. Tais abordagens são descritas nas seções a seguir.

4.1. Abordagem computacional para a descoberta de genes significativos para o câncer

DiSCaGe usa dados de mutação de câncer (SNVs e InDels) e um conjunto de redes de interação de genes não direcionadas e não ponderadas como entrada, enquanto a saída é uma classificação de possíveis genes *drivers*. O método é composto por seis etapas bem definidas, conforme ilustrado na Figura 1. Na Etapa 1, uma matriz de mutação ponderada (WMM) é construída e um valor real é atribuído para cada par paciente-gene, de acordo com o peso definido para a classificação de cada tipo de mutação e número de pacientes mutados. A Etapa 2 utiliza a WMM para obter uma pontuação de mutação para cada gene, chamada frequência de mutação ponderada. Em seguida, na Etapa 3, uma operação de união é realizada nas redes de interação gênica, resultando em uma rede de consenso não direcionada e ponderada. Com base em tal rede, na Etapa 4, uma rede de força de espalhamento de genes (GSSN) é obtida, de acordo com a força de espalhamento de um gene para seus vizinhos diretos e indiretos. A Etapa 5 uma influência de mutação exercida em todos os genes por seus vizinhos é extraída, com base na GSSN e as frequências ponderadas obtidas. Finalmente, na Etapa 6, a pontuação das mutações nos genes é enriquecida com a influência dos vizinhos e uma lista ordenada de genes priorizados é obtida.

O método foi aplicado e avaliado em seis tipos de câncer com o uso de seus conjuntos de dados de mutação e duas redes de interação gênica. Os dados de mutação foram submetidos a uma rotina de pré-processamento e as redes passaram por um processo de previsão de links. As listas de genes priorizados foram avaliadas por meio das medidas de precisão e ganho cumulativo descontado (DCG) e por seis *benchmarks* de genes conhecidamente relacionados ao câncer. Além disso, também foi realizada uma revisão

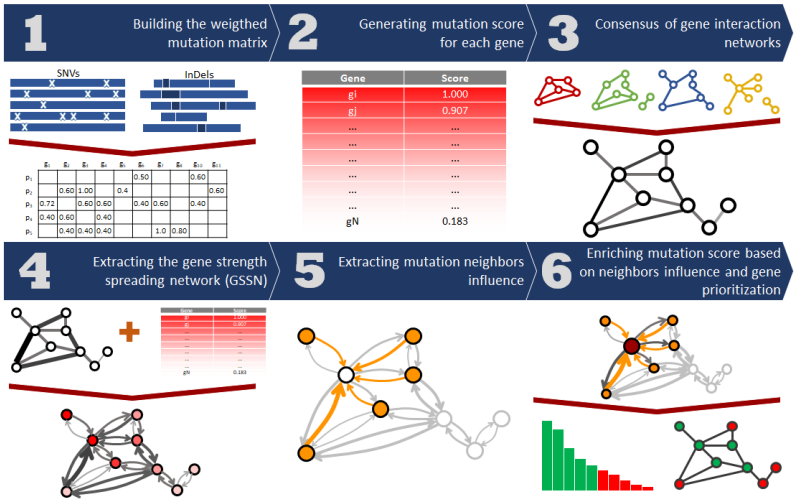

Figura 1. Uma visão geral do método DiSCaGe.

automatizada da literatura dos genes descobertos. Os resultados mostraram o potencial do método DiSCaGe para descobrir genes relacionados ao câncer já conhecidos e sugerir novos genes possivelmente relacionados, incluindo genes mutantes de muito baixa frequência.

4.2. Abordagem computacional para a detecção de genes significativos falso-positivos para o câncer

DFDriver se trata de uma abordagem baseada em aprendizado de máquina supervisionado que detecta possíveis falso-positivos em um conjunto de genes candidatos a serem relacionados ao câncer. Na Figura 2 é apresentada uma visão geral do método. Na Etapa 1, os dados de mutação do câncer, redes de interação gênica e genes rotulados são selecionados de fontes confiáveis e amplamente utilizadas. Na Etapa 2, os dados são pré-processados e os atributos são extraídos para compor um conjunto de dados rotulados criado a partir da combinação de dados de mutação de 33 tipos de câncer e medidas de centralidade a união de redes de interação gênica. Por fim, na Etapa 3, um ajuste de hiperparâmetros é realizado para que modelos otimizados possam ser induzidos e avaliados por meio de validação cruzada estratificada, de acordo com um conjunto de métricas de avaliação.

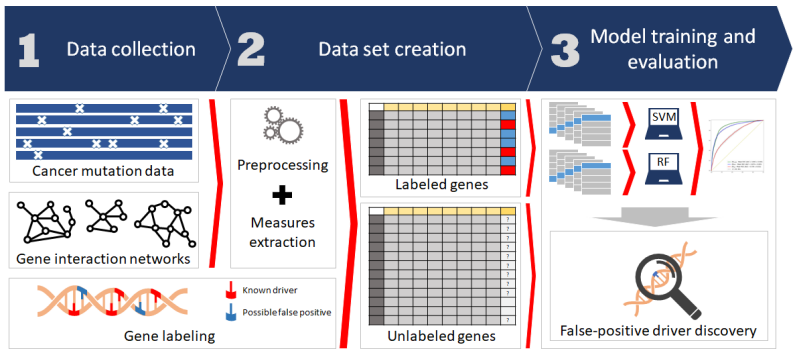

Figura 2. Uma visão geral do método DFDriver.

A avaliação foi realizada usando métricas clássicas de aprendizado de máquina,

e mostrou o potencial preditivo dos modelos e os benefícios na combinação de dados de mutação e redes gênica, melhorando a capacidade preditiva dos modelos.

5. Contribuições e discussão

A contribuição central da tese foi a descoberta de genes significativos para o câncer. Tais genes foram identificados e uma detecção de possíveis falso-positivos pôde ser realizada, a fim de obter resultados mais confiáveis. A contribuição foi alcançada por meio da proposta de duas abordagens computacionais. Com as abordagens propostas foi possível: 1) Priorizar genes conhecidamente relacionados ao câncer; 2) Priorizar genes relacionados ao câncer com baixa frequência de mutação; 3) Sugerir genes que não estão em *benchmarks* de genes conhecidos, mas são citados em trabalhos de pesquisa como relacionados ao câncer; 4) Sugerir possíveis novos genes relacionados ao câncer; e 5) Sugerir possíveis candidatos a genes falso-positivos.

As contribuições das abordagens propostas podem ser resumidas em três perspectivas: 1) Perspectiva computacional: redes complexas e seus algoritmos foram usados em redes de interação gênica combinados com dados de mutação. Especialmente, uma força de espalhamento assimétrica adaptada foi empregada para quantificar como uma mutação pode influenciar a vizinhança dos genes na rede; 2) Perspectiva biológica: A hipótese local definida por [Barabási et al. 2011] foi utilizada como base do método, juntamente com as mutações ponderadas, baseadas no impacto funcional de diferentes tipos de mutação nos genes. Além disso, uma abordagem de previsão de links foi realizada nas redes para tratar com o problema de interações gênicas incompletas; 3) Perspectiva do usuário: as abordagens foram construídas para serem facilmente utilizadas pelos usuários finais, exigindo dados de entrada em formatos padrão. Além disso, o usuário deve definir pesos de mutação, sem definição de hiperparâmetros pouco claros e de difícil interpretação.

Como decorrência da pesquisa de doutorado, vários trabalhos de pesquisa foram submetidos e publicados em conferências e revistas: 1) Um resumo expandido no Simpósio Brasileiro de Computação Aplicada a Saúde de 2019 [Cutigi et al. 2019b]; 2) Um artigo completo no *Brazilian Symposium on Bioinformatics* de 2019 [Cutigi et al. 2019a]; 3) Um artigo completo, como segundo autor, no Simpósio Brasileiro de Computação Aplicada a Saúde de 2020 [Ramos et al. 2020]; 4) Um artigo completo na revista J*ournal of Bioinformatics and Computational Biology* em 2020 [Cutigi et al. 2020a]; 5) Um artigo completo no B*razilian Symposium on Bioinformatics* de 2021 [Cutigi et al. 2020b]; e 6) Um artigo completo na revista *Nature Scientific Reports* em 2021 [Cutigi et al. 2021].

Referências

Bailey, M. H. et al. (2018). Comprehensive characterization of cancer driver genes and mutations. *Cell*, 173(2):371 – 385.e18.

Barabási, A.-L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. *Nature reviews genetics*, 12(1):56–68.

Cheng, F., Zhao, J., and Zhao, Z. (2015). Advances in computational approaches for prioritizing driver mutations and significantly mutated genes in cancer genomes. *Briefings in Bioinformatics*, 17(4):642.

Cutigi, J. F., Evangelista, A. F., Reis, R. M., and Simao, A. (2021). A computational approach for the discovery of significant cancer genes by weighted mutation and asymmetric spreading strength in networks. *Scientific reports*, 11(1):1–10.

Cutigi, J. F., Evangelista, A. F., and Simao, A. (2019a). GeNWeMME: A network-based computational method for prioritizing groups of significant related genes in cancer. In *Advances in Bioinformatics and Computational Biology*, pages 29–40. Springer.

Cutigi, J. F., Evangelista, A. F., and Simao, A. (2019b). A proposal of a graph-based computational method for ranking significant set of related genes in cancer. In *Anais Principais do XIX Simpósio Brasileiro de Computação Aplicada à Saúde*, pages 300–305, Porto Alegre, RS, Brasil. SBC.

Cutigi, J. F., Evangelista, A. F., and Simao, A. (2020a). Approaches for the identification of driver mutations in cancer: A tutorial from a computational perspective. *Journal of Bioinformatics and Computational Biology*, 18(03):2050016. PMID: 32698724.

Cutigi, J. F., Evangelista, R. F., Ramos, R. H., Ferreira, C. d. O. L., Evangelista, A. F., de Carvalho, A. C., and Simao, A. (2020b). Combining mutation and gene network data in a machine learning approach for false-positive cancer driver gene discovery. In *Brazilian Symposium on Bioinformatics*, pages 81–92. Springer.

Dimitrakopoulos, C. M. and Beerenwinkel, N. (2017). Computational approaches for the identification of cancer genes and pathways. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 9(1):e1364.

Hou, J. P. and Ma, J. (2013). *Identifying Driver Mutations in Cancer*, pages 33–56. Springer Netherlands.

Ozturk, K., Dow, M., Carlin, D. E., Bejar, R., and Carter, H. (2018). The emerging potential for network analysis to inform precision cancer medicine. *Journal of molecular biology*, 430(18):2875–2899.

Ramos, R. H., Cutigi, J. F., de Oliveira Lage Ferreira, C., Evangelista, A. F., and Simao, A. (2020). Analyzing different cancer mutation data sets from breast invasive carcinoma (brca), lung adenocarcinoma (luad), and prostate adenocarcinoma (prad). In *Anais Principais do XX Simpósio Brasileiro de Computação Aplicada à Saúde*, pages 37–48, Porto Alegre, RS, Brasil. SBC.

Raphael, B. J., Dobson, J. R., Oesper, L., and Vandin, F. (2014). Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine. *Genome Medicine*, 6(1):5.

Stratton, M. R. (2009). The cancer genome. *Nature*, 458(7239):719–724.

Vincent, J.-L. (2017). The coming era of precision medicine for intensive care. *Critical Care*, 21(3):314.

Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A., and Kinzler, K. W. (2013). Cancer genome landscapes. *Science*, 339(6127):1546–1558.