10 Bioinformática na luta contra o câncer: os bancos de dados na pesquisa oncológica

Autores 10.1

Thayanne Thyssyanne de Souza Soares Costa, Lara Vitoria da Costa Bezerra

Revisão: Aline de Paula Dias da Silva, Thiago M. N. de Camargo

Cite este artigo 10.1

Costa, TTSS; Bezerra, LVC. **Bioinformática na luta contra o câncer: os bancos de dados na pesquisa oncológica.** BIOINFO. ISSN: 2764-8273. Vol. 3. p.10 (2023). doi: 10.51780/bioinfo-03-10

Resumo 10.1

A BIOINFORMÁTICA desempenha um papel fundamental em diversas áreas, principalmente, na área da saúde com destaque no ramo da oncologia. Análises de grande quantidades de dados genômicos gerados através de sequenciamento, estudo de interações de proteínas, auxiliando com suas ferramentas no desenvolvimento de novas terapias na área oncológica e na medicina identificando perfis genéticos. A bioinformática ainda pode auxiliar no monitoramento da resposta ao tratamento por meio de modelos de mineração de dados, no qual, vão encontrar padrões para determinado tipo de câncer, além de desenvolver muitas outras ferramentas nessa área. Atualmente existem diversos bancos de dados com grande potencial de mineração de dados na área da oncologia, sendo um aliado nos estudos genéticos e oncológicos.

10.1 Introdução

A tão conhecida bioinformática, é a área que utiliza conhecimentos biológicos, estatísticos e matemáticos. A importância e destaque da Bioinformática começou em virtude do Projeto Genoma, que foi um projeto com o objetivo de sequenciar todo o genoma humano, mapeando e identificando todos os genes presentes no DNA, fornecendo um mapa detalhado do código genético humano, permitindo grandes avanços na ciência como entender melhor a estrutura e funcionamento dos genes e suas interações com o corpo humano, e principalmente pelo grande volume de dados que começou a existir e pela possibilidade de lidar com o armazenamento dessas informações, auxiliando nas estatísticas, análises e identificações [5].

Interdisciplinaridade e multifuncionalidade são palavras-chave na bioinformática. Podemos ver as aplicações da bioinformática em diversas áreas, como no agronegócio, junto a produção de alimentos e descobertas em relação ao melhoramento genético vegetal [9]. Além disso, pode desempenhar papel fundamental na indústria farmacêutica, auxiliando na descoberta de novas vias

metabólicas e farmacêuticas, junto da modelagem computacional [4]. O mundo das possibilidades na bioinformática é gigantesco (Figura 10.1), desempenhando um papel fundamental na área da saúde, principalmente, quando falamos da área oncológica. Tendo isso em vista, a bioinformática é um campo vasto de oportunidades dentre de seus ramos, auxiliando no desenvolvimento de novas alternativas de ferramentas, tratamentos e padrões dentro da área da oncologia.

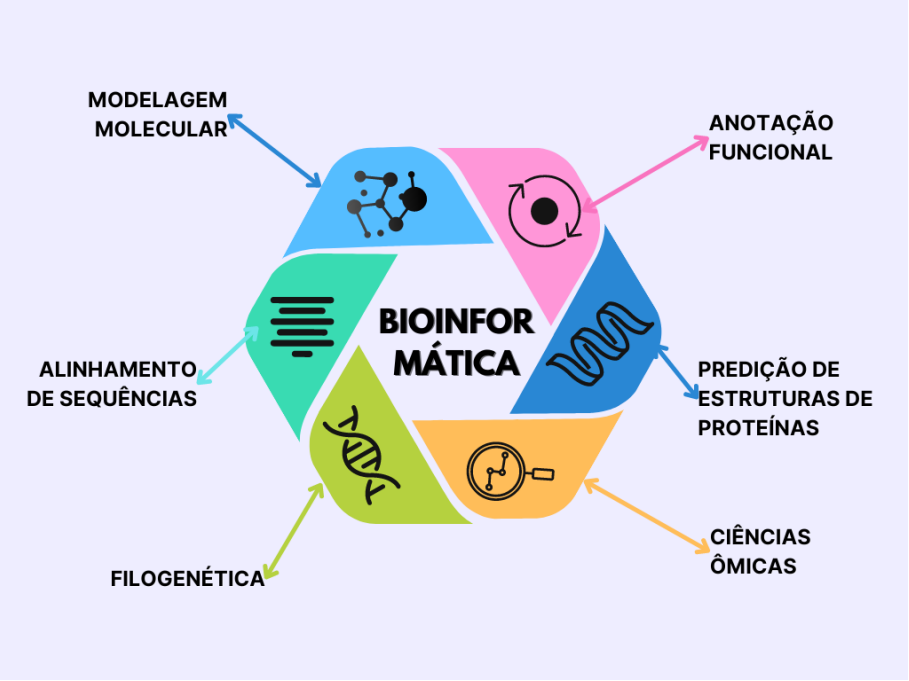

Figura 10.1: Aplicações da bioinformática e suas áreas de forma geral, ciências ômicas, filogenética, alinhamento de sequências, modelagem molecular, anotação funcional e predição de estruturas de proteínas. Fonte: adaptado de [14].

10.2 A Bioinformática na Oncologia

O câncer abrange mais de 100 doenças diferentes que têm em comum o crescimento desordenado de células cancerígenas que vai ocasionalmente proporcionar a formação de tumores malignos. Segundo estimativas do Instituto Nacional do Câncer (INCA) é esperado para o próximo triênio 2023-2025, mais de 700 mil novos casos de câncer no Brasil [15].

As diversas ferramentas da bioinformática têm auxiliado consideravelmente em busca de melhores tratamentos e análises na oncologia. Uma dessas ferramentas é o *Big Data*, que veio para auxiliar na manipulação dos grandes conjuntos de dados existentes hoje [11]. Dados esses que são extremamente relevantes para extrair informações importantes e de interesse, auxiliando em diagnósticos fazendo uma tomada de decisão mais assertiva, além de ajudar a entender ainda mais os padrões de doença. Além disso, a bioinformática tem um papel crucial nas pesquisas oncológicas atrelado à aprendizagem de máquina e inteligência artificial, principalmente no descobrimento de alterações genéticas e no desenvolvimento de novos fármacos e tratamentos mais racionais.

O câncer é caracterizado por função proteica e padrões de transcrição alterados, que são consequência de mutações somáticas e alterações epigenéticas, garantindo vantagem no crescimento de tumores, dessa forma existem diversas ferramentas da bioinformática que podem auxiliar na pesquisa contra o câncer e podemos citar algumas delas:

10.2.1 Análise de expressão gênica:

Um grande aliado na análise de expressão gênica é o Sequenciamento de Nova Geração (NGS), no qual, permite a leitura simultânea de milhões de fragmentos de DNA ou RNA, ocasionando uma abordagem de alto rendimento e mais econômico para análises de sequências genômicas e transcricionais. Ao realizar o sequenciamento de RNA (RNA-Seq), que é uma aplicação específica do (NGS), é possível analisar o transcriptoma de maneira abrangente e detalhada, que no qual, transcriptoma é o conjunto de moléculas de RNA transcritas a partir do material genético de um organismo [1]. O estudo do transcriptoma é fundamental, pois esclarece sobre os componentes e elementos funcionais do genoma [19].

Na análise da expressão gênica é possível identificar quais genes, por exemplo, estão inativos ou ativos em células cancerígenas, analisando dados de expressão gênica é possível extrair essas informações, além de identificar assinaturas genéticas distintas de diferentes tipos de câncer [2]. Assim, o NGS permite que quande quantidade de dados sejam sequenciados, evidenciando a importância do *Big Data* e a mineração de dados na bioinformática.

Algumas etapas envolvidas nesse processamento são:

- Pré-processamento dos dados, no qual, os dados brutos do sequenciamento de RNA são processados.

- Análise diferencial da expressão gênica que compara os níveis de expressão entre diferentes grupos de amostras com o intuito de identificar genes que apresentam alterações significativas na expressão.

- Análise funcional, quando os genes diferencialmente expressos são submetidos a análises funcionais para entender os processos biológicos afetados pelo mesmo.

10.2.2 Análises em proteômica:

A proteômica estuda as interações e funções de um grupo de proteínas ou de uma proteína em uma célula. Os dados proteômicos são úteis na classificação de células e tecidos em diferentes estágios da doença e na compreensão dos diferentes mecanismos biológicos envolvidos. A bioinformática permite o processamento dos dados e identificação das proteínas após serem feitas as técnicas experimentais de proteômica, como, por exemplo, a espectrometria de massa. Esse processamento se dá principalmente através da análise quantitativa da expressão proteica e na anotação funcional das proteínas identificadas. Podendo assim, auxiliar na descoberta e identificação de alvos terapêuticos através da compreensão das funções e interações dessas proteínas no câncer. Assim, a onco proteômica visa estudar a interação das proteínas em uma célula cancerosa por tecnologia proteômica, sendo uma área promissora que se utiliza de biomarcadores tumorais para diagnóstico precoce [16].

10.2.3 Predição e Mineração de Dados:

Com base em dados genéticos e moleculares de pacientes é possível se utilizar de modelos computacionais que vão ajudar na predição a partir dos dados favorecendo uma melhor tomada de decisão no diagnóstico médico. Assim, a mineração de dados que é uma ferramenta importante que envolve modelos estatísticos desempenha papel fundamental ajudando a encontrar padrões

na doença, padrões esses que podem ser fundamentais para ajudar tanto no diagnóstico médico como no prognóstico, se tornando uma ferramenta promissora em virtude do seu alto desempenho [7].

10.3 Por que os bancos de dados são importantes na oncologia?

Os bancos de dados tiveram sua origem por volta de 1960 e são uma forma de organização de informações importantes armazenadas em um sistema de computador, com grande potencial nas pesquisas de diversas áreas [13]. Com o aumento da produção de dados de pesquisas nos últimos anos, os bancos de dados clínicos e genéticos desempenham um papel fundamental na pesquisa, principalmente na oncologia, tendo em vista, que eles têm o potencial de oferecer informações valiosas, auxiliando em novas pesquisas, diagnósticos, tratamentos, predição e prognóstico do câncer. Além das vantagens citadas, podemos acrescentar outras, como, por exemplo:

Organização e armazenamento dos dados clínicos: Os banco de dados clínicos contém informações clínicas extremamente importantes como o estadiamento do tumor, histórico familiar do câncer, tratamentos anteriores, Classificação de Tumores Malignos (TNM) e dados demográficos, oriundos de diversos centros médicos e instituições. Essas e muitas outras informações são extremamente valiosas aos pesquisadores e profissionais da saúde, ao ampliarem o conhecimento científico em pesquisas de evolução do câncer e novos tratamentos.

Epidemiológica: Através dos bancos de dados é possível realizar amplos estudos epidemiológicos sobre o câncer. Com as análises de dados provenientes de bancos de dados confiáveis é possível identificar padrões de prevalência e incidência da doença, correlacionando até mesmo, com a mineração de dados citada anteriormente. Estudar e entender a epidemiologia do câncer é uma questão de saúde pública, pois essas informações podem embasar novas políticas

públicas mais efetivas, direcionar recursos e novas medidas preventivas para a doença.

Inteligência artificial e Aprendizagem de máquinas: Os dados disponíveis nos bancos de dados tem potencial valioso de treinamento de algoritmos de aprendizado de máquina e inteligência artificial que podem ser utilizados para aplicar em modelos preditivos podendo achar padrões e descobertas na medicina de precisão [6].

Existem diversas bases de dados com grande potencial na área de oncologia como os bancos de dados clínicos (Figura 10.2) e os bancos de dados genômicos (Figura 10.3):

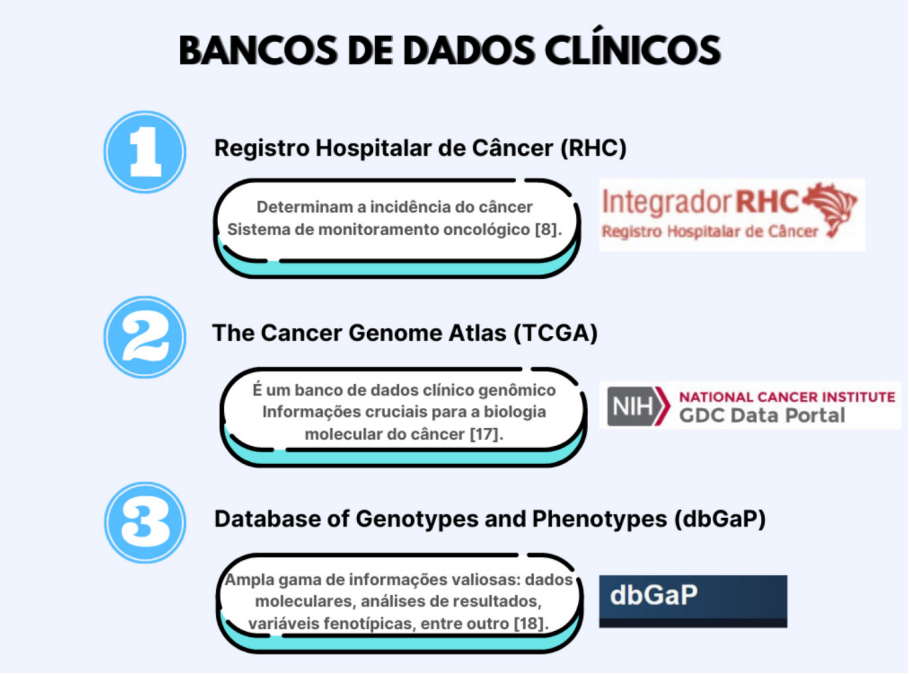

Figura 10.2: Bancos de dados clínicos. Fonte: [8; 17-18].

Registro Hospitalar de Câncer (RHC): É uma fonte de informações que se encontra instalada em diversos hospitais e instituições oncológicas, sejam públicas ou privadas. O RHC do Banco de Dados do Instituto Nacional do Câncer (INCA) do Brasil, implementado em 1983, é o registro mais antigo do Brasil [10]. Os RHC coletam informações de todos os pacientes com

diagnóstico de câncer confirmado e contém dados clínicos com informações importantes como, estadiamento do tumor, historio familiar do câncer, histórico de consumo de bebidas alcoólicas, sexo, localização, tratamentos anteriores e outros. Essas informações desempenham um papel fundamental ao subsidiar estudos prognósticos e de sobrevida de pacientes oncológicos. Além disso, é válido ressaltar que os pacientes têm sua integridade mantida, tendo em vista, que os pacientes não são identificados.

***The Cancer Genome Atlas* (TCGA):** É uma base de dados internacional que disponibiliza dados que visam mapear as alterações genéticas, como mutações, nos vários tipos de câncer, além de disponibilizar dados genômicos e clínicos assim como o GDC. Apenas na última década, o TCGA conseguiu gerar mais de 2,5 petabytes de dados genômicos, epigenômicos, transcriptômicos e proteômicos de cânceres [3]. Além disso, permite realizar análises dos conjuntos de dados de forma integrada.

***Database of Genotypes and Phenotypes* (dbGaP):** É um repositório de dados do *National Institutes of Health* (NIH) dos Estados Unidos que contém dados genômicos e informações clínicas de diversas doenças, incluindo o câncer. E possui informações produzidas por diversos estudos que investigaram a interação de genótipo e fenótipo. Também incluem dados moleculares, imagens médicas e outras informações gerais sobre o estudo e documentos, como protocolos de pesquisa [18]. Essas informações são de extrema importância para diversas novas pesquisas na área genômica.

***Genomic Data Commons* (GDC):** É uma base de dados internacional que compartilha dados genômicos e informações importantes como dados clínicos e genéticos, auxiliando nos avanços das pesquisas de caráter oncológico.

***Cancer Cell Line Encyclopedia* (CCLE):** É um banco de dados que apresenta informações referentes às linhagens das células tumorais, gerando dados de 1000 linhagens celulares de diferentes tecidos [12]. Além disso, o CCLE possui uma ferramenta de visualização de dados intitulada CLiFF.

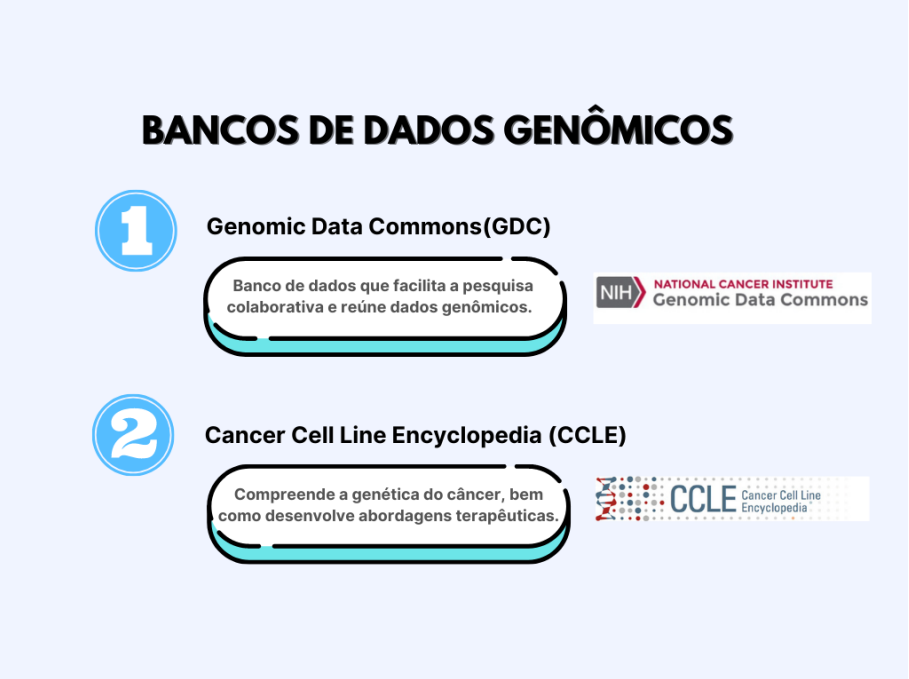

Figura 10.3: Bancos de dados genômicos. Fonte: autoria própria.

Saiba mais 10.1

Este artigo está disponível em https://bioinfo.com.br/bioinformatica-na-luta-contra-o-cancer-os-bancos-de-dados-na-pesquisa-oncologica/

10.4 Referências

[1] CARVALHO, Mayra Costa da Cruz Gallo de; SILVA, Danielle Cristina Gregorio da. Sequenciamento de DNA de nova geração e suas aplicações na genômica de plantas. Ciência Rural, v. 40, p. 735-744, 2010.

[2] CIEŚLIK, Marcin; CHINNAIYAN, Arul M. Cancer transcriptome profiling at the juncture of clinical translation. Nature Reviews Genetics, v. 19, n. 2, p. 93-109, 2018.

[3] DAS, Tonmoy et al. Integration of online omics-data resources for cancer research. Frontiers in Genetics, v. 11, p. 578345, 2020.

[4] GUIDO, Rafael VC; ANDRICOPULO, Adriano D.; OLIVA, Glaucius. Planejamento de fármacos, biotecnologia e química medicinal: aplicações em doenças infecciosas. Estudos avançados, v. 24, p. 81-98, 2010.

[5] HAGEN, Joel B. The origins of bioinformatics. Nature Reviews Genetics, v. 1, n. 3, p. 231-236, 2000.

[6] JOTHI, Neesha et al. Data mining in healthcare–a review. Procedia computer science, v. 72, p. 306-313, 2015.

[7] KAUR, Ishleen; DOJA, M. N.; AHMAD, Tanvir. Data mining and machine learning in cancer survival research: an overview and future recommendations. Journal of Biomedical Informatics, v. 128, p. 104026, 2022.

[8] KLIGERMAN, Jacob. Registro hospitalar de câncer no Brasil. Revista Brasileira de Cancerologia, v. 47, n. 4, p. 357-359, 2001.

[9] LÜHRS, L. et al. Identificação de microssatélites e síntese de primes para erva-mate a partir de programas de bioinformática. 2021.

[10] MINISTÉRIO DA SAÚDE (BR). INSTITUTO NACIONAL DE C NCER JOSÉ ALENCAR GOMES DA SILVA. Informação dos registros hospitalares de câncer como estratégia de transformação: perfil do Instituto Nacional de Câncer José Alencar Gomes da Silva em 25 anos. 2012.

[11] NEURALMED. Big Data na saúde: Por que a importância de olhar para esse universo? Disponível em: https://www.neuralmed.ai/blog/big-data-na-saude. Acesso em: 23 maio 2023.

[12] NUSINOW, David P. et al. Quantitative proteomics of the cancer cell line encyclopedia. Cell, v. 180, n. 2, p. 387-402. e16, 2020

[13] ORACLE. O que é um Banco de Dados? Disponível em: https://www.oracle.com/br/database/what-is-database/. Acesso em: 23 maio 2023.

[14] SAFADY, Nágela G. Bioinformática: união entre ciência e tecnologia. Disponível em: https://blog.varsomics.com/bioinformatica/. Acesso em: 22 maio 2023.

[15] SANTOS, M. de O.; LIMA, F. C. da S. de; MARTINS, L. F. L.; OLIVEIRA, J. F. P.; ALMEIDA , L. M. de; CANCELA, M. de C. Estimativa de Incidência de Câncer no Brasil, 2023-2025. Revista Brasileira de Cancerologia, [S. l.], v. 69, n. 1, p. e–213700, 2023. DOI: 10.32635/2176-9745.RBC.2023v69n1.3700. Disponível em: https://rbc.inca.gov.br/index.php/revista/article/view/3700. Acesso em: 11 jul. 2023.

[16] SHRUTHI, Basavaradhya Sahukar et al. Proteomics: A new perspective for cancer. Advanced biomedical research, v. 5, 2016.

[17] TOMCZAK, Katarzyna; CZERWIŃSKA, Patrycja; WIZNEROWICZ, Maciej. Review The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemporary Oncology/Współczesna Onkologia, v. 2015, n. 1, p. 68-77, 2015.

[18] TRYKA, Kimberly A. et al. NCBI's Database of Genotypes and Phenotypes: dbGaP. Nucleic acids research, v. 42, n. D1, p. D975-D979, 2014.

[19] WANG, Zhong; GERSTEIN, Mark; SNYDER, Michael. RNA-Seq: a revolutionary tool for transcriptomics. Nature reviews genetics, v. 10, n. 1, p. 57-63, 2009.