

biotechnology
DNA
research
bioinformatics

CIÊNCIAS BIOLÓGICAS

BIOINFORMÁTICA UMA CIÊNCIA SEM FRONTEIRAS

ORGANIZADORES

IGOR LUIZ VIEIRA DE LIMA SANTOS
CARLIANE REBECA COELHO DA SILVA
RAFAEL TRINDADE MAIA

1ª

Edição

Acesso livre ao E-Book em
WWW.EDITORASCIENCE.COM.BR

 EDITORA
SCIENCE
ANO 2026



biotechnology
DNA
research
bioinformatics

CIÊNCIAS BIOLÓGICAS

BIOINFORMÁTICA UMA CIÊNCIA SEM FRONTEIRAS

ORGANIZADORES

IGOR LUIZ VIEIRA DE LIMA SANTOS
CARLIANE REBECA COELHO DA SILVA
RAFAEL TRINDADE MAIA

1ª

Edição

Acesso livre ao E-Book em

WWW.EDITORASCIENCE.COM.BR

CAMPINA GRANDE-PB
 EDITORA
SCIENCE
ANO 2026

Todos os Direitos Desta Edição Reservados à
© 2026 EDITORA SCIENCE
Av. Marechal Floriano Peixoto. 5000.
Campina Grande, PB, 58434-500.
CNPJ: 42.754.503/0001-00

REGISTRO CBL (Câmara Brasileira do Livro)

Dados Internacionais de Catalogação na Publicação (CIP)
(Câmara Brasileira do Livro, SP, Brasil)

Bioinformática uma ciência sem fronteiras [livro eletrônico] / organizadores Igor Luiz Vieira de Lima Santos, Carliane Rebeca Coelho da Silva, Rafael Trindade Maia. -- 1. ed. -- Campina Grande,
PB : Carliane Rebeca Coelho da Silva, 2026.
PDF

Bibliografia
ISBN 978-65-01-90653-9

1. Biologia 2. Ciências da vida 3. Códigos
4. Computação - Estudo e ensino 5. Inteligência artificial 6. Programação (Computadores) I. Santos, Igor Luiz Vieira de Lima. II. Silva, Carliane Rebeca Coelho da. III. Maia, Rafael Trindade.

26-331493.0

CDD-570

Índices para catálogo sistemático:

1. Ciências biológicas 570

Maria Alice Ferreira - Bibliotecária - CRB-8/7964



<https://doi.org/10.56001/26.9786501906539>

Para consulta na CBL acesse: <https://www.cbldados.org.br/isbn/pesquisa/>



Editora–Chefe

Pós-Dra. Carliane Rebeca Coelho da Silva

Editores Organizadores

Prof. Dr. Igor Luiz Vieira de Lima Santos

Pós-Dra. Carliane Rebeca Coelho da Silva

Prof. Dr. Rafael Trindade Maia

Editoração e Diagramação

Corpo Técnico da Editora Science

Revisão Principal/Por Pares

Os Autores / Revisores *Ad Hoc* / Corpo Editorial / Organizadores

Revisão Final

Pós-Dra. Carliane Rebeca Coelho da Silva

Programas Registrados de Design

©Canva Pro Registered Design



Copyright © 2026 Editora Science

Copyright Textual © 2026 Os autores

Copyright da Edição © 2026 Editora Science

Todos os Direitos e os Termos de Cessão de Direitos Autorais para esta edição foram cedidos à Editora Science pelos próprios autores.

Declaração de Direitos

Todos os direitos reservados.

Qualquer parte deste livro pode ser reproduzida, transmitida de qualquer forma ou por qualquer meio, eletrônico, mecânico, fotocópia, microfilmagem, gravação ou de outra forma, desde que citada a fonte. Creative Commons Attribution 4.0 International (CC BY-NC-ND 4.0). <https://creativecommons.org/licenses/by-nc-nd/4.0/>

Todos os artigos de autoria inédita, revisão, comentários, opiniões, resultados, conclusões ou recomendações são de inteira responsabilidade do(s) autor(es), e não refletem necessariamente as opiniões dos editores e/ou da empresa.

Para cópias impressas, para compras em massa e/ou informações sobre este e outros títulos da © Editora Science, entre em contato com a editora pelo telefone: Tel.: +55-83-991647953; E-mail: contato@editorascience.com ou editorascience@gmail.com

Siga nossas redes sociais fique por dentro das novidades e amplie o alcance dos nossos livros:

Facebook: <http://www.facebook.com/editorascience>

Instagram: <https://www.instagram.com/editorascience>

© 2026 EDITORA SCIENCE

Editora-Chefe:

PÓS-DRA. CARLIANE REBECA COELHO DA SILVA (EDITORA-CHEFE)

Conselho Editorial:

PÓS-DRA. CARLIANE REBECA COELHO DA SILVA (EDITORA-CHEFE)

PROF. DR. IGOR LUIZ VIEIRA DE LIMA SANTOS (UFCEG)

DRA. LUCIANA AMARAL DE MASCENA COSTA (UFRPE)

PÓS-DRA. AYRLES FERNANDA BRANDÃO DA SILVA (UFCE)

Corpo Editorial:

DR. MARCUS VINICIUS PERALVA SANTOS (IFTO)

DR. RÔMULO ALVES DE OLIVEIRA (IFSE)

DRA. ROSEANNE SANTOS DE CARVALHO (IFSE)

PÓS-DRA. CARLIANE REBECA COELHO DA SILVA (EDITORA-CHEFE)

DRA. FERNANDA MIGUEL DE ANDRADE (FMS)

DR. MILTON GONÇALVES DA SILVA JUNIOR (UNIARAGUAIA)

DRA. WELMA EMIDIO DA SILVA (FIS)

DRA. AYRLES FERNANDA BRANDÃO (UFCE)

DR. GABRIEL PARISOTTO (UNISUAM)

DR. IGOR LUIZ VIEIRA DE LIMA SANTOS (UFCEG)

ME. LÚCIA MAGNÓLIA ALBUQUERQUE SOARES DE CAMARGO (UNIFACISA)

DRA. LUCIANA AMARAL DE MASCENA COSTA (UFRPE)

ME. MARCELO SALVADOR CELESTINO (UNESP)

PÓS-DRA. ELIANA NAPOLEÃO COZENDEY DA SILVA (FIOCRUZ-ENSP)

DR. EDIGAR HENRIQUE VAZ DIAS (UFCAT)

DR. HENRIQUE MACIEL VIEIRA DE MORAES (UFRJ)

DR. CRISTIANO CUNHA COSTA (UFS)

MSC. DANIEL DA SILVA GOMES (UFPB)

DRA. FRANCIELI DE FATIMA MISSIO (UFMS)

DR. JOSÉ OLÍVIO LOPES VIEIRA JÚNIOR (UENF)

DRA. NARA HELENA TAVARES DA PONTE (UEAP)

DR. LUIZ ALEXANDRE VALADÃO DE SOUZA (SME-RJ)

PÓS-DRA. MICHELE APARECIDA CERQUEIRA RODRIGUES (UFLO)

PÓS-DR. MARCOS PEREIRA DOS SANTOS (FACUR)

LICENSE PUBLICATION DETAILS

Copyright © 2026 Editora Science

Copyright Notice

All content in this work, except where otherwise noted, is licensed under a Creative Commons Attribution 4.0 International (CC BY-NC-ND 4.0) license which permits copying, distribution, and adaptation of the work, provided the original work is properly cited and any changes from the original work are properly indicated. Any altered, transformed, or adapted form of the work may only be distributed under the same or similar license to this one.

© 2026 by Carliane Rebeca Coelho da Silva is licensed under Attribution-NonCommercial-NoDerivatives 4.0 International 



**Attribution-NonCommercial-
NoDerivatives 4.0 International
(CC BY-NC-ND 4.0)**

HOW CITE THIS BOOK:

NLM Citation

Santos ILVL, Silva CRC, Maia RT editor. *Bioinformática uma Ciência sem Fronteiras*. 1st ed. Campina Grande (PB): Editora Science; 2026.

APA Citation

Santos, I. L. V. L.; Silva, C. R. C. & Maia, R. T. (Eds.). (2026). *Bioinformática uma Ciência sem Fronteiras*. (1st ed.). Editora Science.

ABNT Brazilian Citation NBR 6023:2018

SANTOS, I. L. V. L.; SILVA, C. R. C.; MAIA, R. T. **Bioinformática uma Ciência sem Fronteiras** 1. ed. Campina Grande: Editora Science, 2026.

WHERE ACCESS THIS BOOK:

www.editorascience.com.br/

Sumário

CAPÍTULO 1 **1**

POTENCIAL GENÔMICO DE *TREMELLA MESENERICA* NA BIOSÍNTESE DE METABÓLITOS SECUNDÁRIOS VIA FERRAMENTAS *IN SILICO* **1**

GENOMIC POTENTIAL OF *TREMELLA MESENERICA* IN THE BIOSYNTHESIS OF SECONDARY METABOLITES VIA *IN SILICO* TOOLS 1

DOI: <https://doi.org/10.56001/26.9786501906539.01> **1**

Dielson Márcio Fonseca Maués 1

Hellen Kempfer Philippsen 1

Leonardo Castelo Branco Carvalho 1

CAPÍTULO 2 **17**

PROSPECÇÃO SISTEMÁTICA E APLICADA DA HERANÇA GENÉTICA DA MIOPIA UTILIZANDO FERRAMENTAS DE BIOINFORMÁTICA **17**

SYSTEMATIC AND APPLIED PROSPECTING OF THE GENETIC INHERITANCE OF MYOPIA USING BIOINFORMATICS TOOLS 17

DOI: <https://doi.org/10.56001/26.9786501906539.02> **17**

Amanda Geovana Pereira de Araújo 17

Carliane Rebeca Coelho da Silva 17

Igor Luiz Vieira de Lima Santos 17

CAPÍTULO 3 **36**

PROSPECÇÃO SISTEMÁTICA E APLICADA DA HERANÇA GENÉTICA DA DEPRESSÃO UTILIZANDO FERRAMENTAS DE BIOINFORMÁTICA **36**

SYSTEMATIC AND APPLIED PROSPECTING OF THE GENETIC INHERITANCE OF DEPRESSION USING BIOINFORMATICS TOOLS 36

DOI: <https://doi.org/10.56001/26.9786501906539.03> **36**

Tainá Oliveira de Araújo 36

Carliane Rebeca Coelho da Silva 36

Igor Luiz Vieira de Lima Santos 36

CAPÍTULO 4 **54**

DESENVOLVIMENTO DE UMA VACINA BASEADA EM LINHAGEM *STAPHYLOCOCCUS AUREUS* RF122 CONTRA MASTITE BOVINA **54**

DEVELOPMENT OF A VACCINE BASED ON THE *STAPHYLOCOCCUS AUREUS* RF122 STRAIN AGAINST BOVINE MASTITIS 54

DOI: <https://doi.org/10.56001/26.9786501906539.04> **54**

Tatiane Batista dos Santos 54

Marianne Celestino Andrade 54

Kelisson Alves Sousa	54
Enzo Henrike Bitencourt de Moraes	54
Pedro Henrique da Silva Rodrigues	54
Victória Garcia Peres	54
Lorena Emília Sena Lopes	55

CAPÍTULO 5 **75**

IN SILICO IDENTIFICATION AND CHARACTERIZATION OF CALEOSIN GENES IN CASHEW (<i>ANACARDIUM OCCIDENTALE</i> L.)	75
IDENTIFICAÇÃO E CARACTERIZAÇÃO IN SILICO DOS GENES DA CALEOSINA EM CAJU (<i>ANACARDIUM OCCIDENTALE</i> L.)	75
DOI: https://doi.org/10.56001/26.9786501906539.05	75
Maria Cidinaria Silva Alves	75
Rousilândia de Araujo Silva	75
Gleyce Hellen de Almeida de Souza	75
José Eduardo Souza Echeverria	75
Fabíola Lucini	75
Ruana Carolina Cabral da Silva	76

CAPÍTULO 6 **93**

BIOINFORMÁTICA: A BÚSSOLA PARA VACINAS PERSONALIZADAS CONTRA O MELANOMA	93
BIOINFORMATICS: THE COMPASS FOR PERSONALIZED VACCINES AGAINST MELANOMA	93
DOI: https://doi.org/10.56001/26.9786501906539.06	93
Maria Cidinaria Silva Alves	93
Rousilândia de Araujo Silva	93
Jakeline Moreira da Silva	93
José Lucas de Araújo	93
José Cláudio Gomes de Araújo	93
José Eduardo Souza Echeverria	94
Gleyce Hellen de Almeida de Souza	94
Fabíola Lucini	94
Ruana Carolina Cabral da Silva	94

CAPÍTULO 7 **103**

CARACTERIZAÇÃO GENÔMICA DE DETERMINANTES DE VIRULÊNCIA E RESISTÊNCIA EM <i>ESCHERICHIA COLI</i> ASSOCIADA À PIELONEFRITE: ESTUDO DA CEPA CFT073	103
GENOMIC CHARACTERIZATION OF VIRULENCE AND RESISTANCE DETERMINANTS IN <i>ESCHERICHIA COLI</i> ASSOCIATED WITH PYELONEPHRITIS: A STUDY OF STRAIN CFT073	103

DOI: https://doi.org/10.56001/26.9786501906539.07	103
Ruana Carolina Cabral da Silva	103
Gleyce Hellen de Almeida de Souza	103
Fabíola Lucini	103
Jakeline Moreira da Silva	103
José Lucas de Araújo	104
José Cláudio Gomes de Araújo	104
José Eduardo Souza Echeverria	104
Lucas Dionizio Leite de Abreu	104
Rousilândia de Araujo Silva	104
Maria Cidinaria Silva Alves	104

CAPÍTULO 8 **116**

DA SEQUÊNCIA À VARIANTE: FUNDAMENTOS E ETAPAS DA ANÁLISE BIOINFORMÁTICA DE VARIANTES GENÉTICAS	116
FROM SEQUENCE TO VARIANT: FOUNDATIONS AND STAGES OF BIOINFORMATIC ANALYSIS OF GENETIC VARIANTS	116
DOI: https://doi.org/10.56001/26.9786501906539.08	116
Ruana Carolina Cabral da Silva	116
Gleyce Hellen de Almeida de Souza	116
Fabíola Lucini	116
Lais Albuquerque de Oliveira	116
Jakeline Moreira da Silva	116
José Lucas de Araújo	117
José Cláudio Gomes de Araújo	117
José Eduardo Souza Echeverria	117
Rousilândia de Araujo Silva	117
Maria Cidinaria Silva Alves	117

CAPÍTULO 9 **128**

IDENTIFICAÇÃO E CARACTERIZAÇÃO BIOINFORMÁTICA DO GENE TP53: ESTUDO DE CASO DE UM SUPRESSOR TUMORAL	128
BIOINFORMATIC IDENTIFICATION AND CHARACTERIZATION OF THE TP53 GENE: A CASE STUDY OF A TUMOR SUPPRESSOR	128
DOI: https://doi.org/10.56001/26.9786501906539.09	128
Ruana Carolina Cabral da Silva	128
Gleyce Hellen de Almeida de Souza	128
Lais Albuquerque de Oliveira	128
Mariana Carvalho Sturaro	128
José Eduardo Souza Echeverria	128
Aline Andrade Martins	129
Fabíola Lucini	129

CAPÍTULO 10 **139**

BASES MOLECULARES DA APOPTOSE MITOCONDRIAL: ORGANIZAÇÃO E DIVERSIDADE ESTRUTURAL DE PROTEÍNAS DA FAMÍLIA BCL-2 E SUAS IMPLICAÇÕES NO CÂNCER 139

MOLECULAR BASIS OF MITOCHONDRIAL APOPTOSIS: ORGANIZATION AND STRUCTURAL DIVERSITY OF BCL-2 FAMILY PROTEINS AND THEIR IMPLICATIONS IN CANCER 139

DOI: <https://doi.org/10.56001/26.9786501906539.10> 139

Lais Albuquerque de Oliveira 139

Gleyce Hellen de Almeida de Souza 139

Ruana Carolina Cabral da Silva 139

CAPÍTULO 11 150

LÚPUS ERITEMATOSO SISTÊMICO: BASES MOLECULARES DA IMUNIDADE INATA E VIAS DE SINALIZAÇÃO NA AUTOIMUNIDADE 150
SYSTEMIC LUPUS ERYTHEMATOSUS: MOLECULAR BASIS OF INNATE IMMUNITY AND SIGNALING PATHWAYS IN AUTOIMMUNITY 150

DOI: <https://doi.org/10.56001/26.9786501906539.11> 150

Ruana Carolina Cabral da Silva 150

Gleyce Hellen de Almeida de Souza 150

Jonatas Thiago dos Santos Tenan 150

Lais Albuquerque de Oliveira 150

Mariana Carvalho Sturaro 151

Rousilândia de Araujo Silva 151

Maria Cidinaria Silva Alves 151

José Eduardo Souza Echeverria 151

Fabíola Lucini 151

CAPÍTULO 12 165

IDENTIFICAÇÃO BIOINFORMÁTICA DE ALVOS MOLECULARES PARA TERAPIAS COMBINADAS EM DOENÇAS METABÓLICAS: UMA ABORDAGEM INTEGRATIVA 165

BIOINFORMATIC IDENTIFICATION OF MOLECULAR TARGETS FOR COMBINATION THERAPIES IN METABOLIC DISEASES: AN INTEGRATIVE APPROACH 165

DOI: <https://doi.org/10.56001/26.9786501906539.12> 165

Ruana Carolina Cabral da Silva 165

Gleyce Hellen de Almeida de Souza 165

Fabíola Lucini 165

Luana Rossato 165

CAPÍTULO 13 179

SEQUENCIAMENTO DE NOVA GERAÇÃO E BIOINFORMÁTICA NO DIAGNÓSTICO DE DOENÇAS HUMANAS: APLICAÇÕES, AVANÇOS E PERSPECTIVAS	179
NEXT-GENERATION SEQUENCING AND BIOINFORMATICS IN THE DIAGNOSIS OF HUMAN DISEASES: APPLICATIONS, ADVANCES, AND PERSPECTIVES	179
DOI: https://doi.org/10.56001/26.9786501906539.13	179
Ruana Carolina Cabral da Silva	179
Gleyce Hellen de Almeida de Souza	179
Fabíola Lucini	179
Luana Rossato	179
 <u>CAPÍTULO 14</u>	 <u>195</u>
PUBLIQUE COM A SCIENCE EM FLUXO CONTÍNUO	195
<i>PUBLISH WITH SCIENCE IN CONTINUOUS FLOW</i>	195
DOI: https://doi.org/10.56001/26.9786501906539.14	195
AUTORES	195
AUTORES	195
AUTORES	195
 <u>CAPÍTULO 15</u>	 <u>196</u>
PUBLIQUE COM A SCIENCE EM FLUXO CONTÍNUO	196
<i>PUBLISH WITH SCIENCE IN CONTINUOUS FLOW</i>	196
DOI: https://doi.org/10.56001/26.9786501906539.15	196
AUTORES	196
AUTORES	196
AUTORES	196
 <u>SOBRE OS ORGANIZADORES DO LIVRO DADOS CNPQ:</u>	 <u>197</u>

PREFÁCIO À 1ª EDIÇÃO

A bioinformática consolidou-se, nas últimas décadas, como uma das áreas mais estratégicas e transformadoras da ciência contemporânea. Ao integrar biologia, computação, estatística e matemática, essa ciência rompe fronteiras tradicionais do conhecimento e se estabelece como um pilar essencial para a compreensão dos sistemas biológicos em sua complexidade, impulsionando avanços significativos na genética, na biotecnologia, na medicina, na farmacologia e em diversas outras áreas do saber.

O livro “Bioinformática: Uma Ciência sem Fronteiras” nasce da necessidade de reunir, em uma única obra, fundamentos teóricos sólidos e aplicações práticas atualizadas, capazes de atender tanto à formação acadêmica quanto às demandas crescentes da pesquisa científica moderna. Ao longo de seus capítulos, o leitor é convidado a explorar conceitos essenciais, ferramentas computacionais, metodologias analíticas e estudos aplicados que evidenciam o papel central da bioinformática na era dos grandes volumes de dados biológicos.

Esta obra destaca-se por sua abordagem interdisciplinar e por valorizar a produção científica de pesquisadores qualificados — mestres, doutores e pós-doutores — comprometidos com a qualidade, a inovação e a ética na ciência. O conteúdo apresentado reflete não apenas o estado da arte da bioinformática, mas também os desafios e perspectivas futuras de um campo em constante evolução, que exige atualização contínua, pensamento crítico e integração entre diferentes áreas do conhecimento.

Ao transcender limites geográficos e disciplinares, “Bioinformática: Uma Ciência sem Fronteiras” reafirma o papel da ciência como instrumento de transformação social, inovação tecnológica e avanço do conhecimento humano.

Que esta obra sirva como fonte de aprendizado, inspiração e estímulo à pesquisa, contribuindo para a formação de profissionais e pesquisadores capazes de enfrentar os desafios científicos do presente e construir soluções para o futuro.

Boa Leitura
Os Organizadores

CAPÍTULO 1

POTENCIAL GENÔMICO DE *Tremella mesenterica* NA BIOSÍNTESE DE METABÓLITOS SECUNDÁRIOS VIA FERRAMENTAS *IN SILICO*

GENOMIC POTENTIAL OF *Tremella mesenterica* IN THE BIOSYNTHESIS OF SECONDARY METABOLITES VIA *IN SILICO* TOOLS

DOI: <https://doi.org/10.56001/26.9786501906539.01>

Submetido em: 15/12/2025

Revisado em: 28/12/2025

Publicado em: 29/01/2026

Dielson Márcio Fonseca Maués

Universidade Federal Rural da Amazônia, Instituto Socioambiental e dos Recursos
Hídricos, Belém-PA

<http://lattes.cnpq.br/8341762657179200>

Hellen Kempfer Philippsen

Universidade Federal Rural da Amazônia, Instituto Socioambiental e dos Recursos
Hídricos, Belém-PA

<http://lattes.cnpq.br/9963999800396051>

Leonardo Castelo Branco Carvalho

Universidade Federal Rural da Amazônia, Instituto Socioambiental e dos Recursos
Hídricos, Belém-PA

<http://lattes.cnpq.br/8955609343848651>

Resumo

A ascensão da tecnologia e avanço do poder computacional permitiu o desenvolvimento da bioinformática, principalmente, no que diz respeito às ciências ômicas. Com tais recursos aplicados possibilitam que institutos de ciência e pesquisa otimizem resultados, facilitando a gestão de dados biológicos com mais eficiência. A biotecnologia, ao desfrutar de artifícios da bioinformática, promove alcançar meios viáveis e inovadores para determinar estratégias de exploração de recursos genéticos e bioativos de modo sustentável,

diminuindo os custos laboratoriais em experimentação primária. Nesta perspectiva, este trabalho objetivou explorar o perfil para a produção de metabólitos secundários a partir da análise genômica de *Tremella mesenterica*. As ferramentas de bioinformática utilizada foi o genbank para a obtenção do genoma do organismo; o Antismash para predição e similaridade genética; o programa KEGG para descrição das vias biossintéticas dos metabólitos secundários; banco de dados auxiliares como MIBiG e Pfam para determinação das funções biológicas dos bioativos. Para sustentar o embasamento das substâncias previstas foram realizadas buscas no ambiente Pubchem e revisão literárias para a averiguar suas propriedades químicas e efeitos biológicos em testes laboratoriais. O resultado constatou que este fungo apresenta potencial para síntese de esqualestatina e ácido clavárico, com capacidades antimaláricos, inibição da síntese de colesterol e, efeitos antitumorais e anti-inflamatório, respectivamente. Após a resultante, conclui-se que as ferramentas in silico são fundamentais na análise genômica associado a investigação preditiva de recursos genéticos e moleculares de organismos e que pesquisas futuras in vitro com o fungo *T. mesenterica* para produção de esqualestatina e ácido clavárico.

Palavras-Chave: bioinformática, genômica, metabólitos secundários, *Tremella mesenterica*

Abstract

The rise of technology and advancements in computing power have enabled the development of bioinformatics, particularly in omics sciences. These applied resources enable science and research institutes to optimize results, facilitating more efficient biological data management. By leveraging bioinformatics capabilities, biotechnology fosters viable and innovative means of determining strategies for exploiting genetic and bioactive resources sustainably, reducing laboratory costs in primary experiments. With this in mind, this study aimed to explore the profile for secondary metabolite production based on the genomic analysis of *Tremella mesenterica*. The bioinformatics tools used were GenBank to obtain the organism's genome; Antismash for prediction and genetic similarity; the KEGG program to describe the biosynthetic pathways of secondary metabolites; and auxiliary databases such as MIBiG and Pfam to determine the biological functions of bioactives. To support the predicted substances, searches in the Pubchem environment and literature reviews were conducted to determine their chemical properties and biological effects in laboratory tests. The results showed that this fungus has the potential to synthesize squalenstatin and clavarinic acid, with antimalarial properties, inhibition of cholesterol synthesis, and antitumor and anti-inflammatory effects, respectively. The results suggest that in silico tools are essential for genomic analysis associated with predictive investigation of genetic and molecular resources of organisms, and that future in vitro research with the fungus *T. mesenterica* is needed for the production of squalenstatin and clavarinic acid.

Keywords: Bioinformatics, genomics, secondary metabolites, *Tremella mesenterica*.

Introdução

Com o advento da tecnologia nos últimos séculos, novas ferramentas e recursos, sobretudo, computacionais tornaram mais precisas as análises e pesquisas, seja no campo da saúde, do meio ambiente, da engenharia, da química, até a biotecnologia. Esta ascensão, cada vez mais desenvolvida e aperfeiçoada, faz com que empresas e institutos científicos lancem alto investimento tecnológico para extrair grandes volumes de dados e tornar seus exames mais finos, estratégicos e inovadores (Pereira, 2024). Nesse contexto, o setor biotecnológico, vem sendo bastante favorecido pelas técnicas biocomputacionais, pois tendo em vista a competitividade do mercado, somado às exigências ambientais de desenvolvimento sustentável, tais artifícios permitem a exploração dos recursos naturais e conhecimento da biodiversidade de modo mais

eficiente, com potencial de redução de custos, bem como na diminuição de impactos sobre o meio ambiente (De carvalho, 2024). Essa visão é interessante pois, a partir de dados genômicos sequenciados, é possível obter noções de recursos genéticos a investigar, como também estudar elementos genômicos entre organismos que podem oferecer produtos similares para uso humano. O papel da bioinformática é fundamental pois possibilita análises e ensaios *in silico* nas ciências ômicas (genômica, metagenômica e proteômica) como mecanismo de usufruir do poder tecnológico e computacional no desenvolvimento científico e biotecnológico (Matos *et al.*, 2023).

Tais tecnologias são determinantes para análises de clusters de genes biossintéticos (BGC), os quais estão associados a expressão de metabólitos secundários de muitos organismos. Metabólitos secundários são compostos naturais produzidos, principalmente, por plantas e microrganismos a fim de obter, proteção contra estresse abiótico e biótico. A produção destas substâncias pode ser influenciada por fatores genéticos, ambientais (temperatura, pH, luminosidade, efeito sazonal) e edáficos (solo). Podem ser caracterizados conforme com a composição química em ácidos orgânicos, seu efeito no meio e/ou em outros organismos e função dentro do agente produtor.

Dentro da infinita biodiversidade mundial, fungos são conhecidos por apresentarem grande volume de excreções metabólicas e variados compostos de diferentes naturezas (Sheldrake, 2021). Alguns grupos como os Basidiomicetos e Ascomicetos, são destacados por possuírem grande arsenal metabólico e boas possibilidades de bioativos de interesse biotecnológico.

Certas espécies do gênero *Tremella* (basidiomicetos) correspondem a organismos de importância ecológica por serem decompositores de matéria orgânica, alimentícia por possuir propriedades nutricionais que podem ser consumidas, e biotecnológica pela produção de enzimas auxiliaadoras em processos industriais. Entretanto, a espécie *Tremella mesenterica*, no momento, apresenta poucos trabalhos abordando suas características genéticas e moleculares, revelando a carência de estudos sobre a possibilidade de contribuição deste organismo na biotecnologia. Assim, o alvo do presente estudo é explorar o genoma da espécie *Tremella mesenterica* com propósito de investigar sequências de genes associadas a expressão de metabólitos secundários utilizando ferramentas de bioinformática, a fim de verificar possíveis aplicações biotecnológicas desta espécie.

Metodologia

- **Busca do genoma**

O genoma da espécie *T. mesenterica* foi consultada no banco de dados do Genbank, no ambiente virtual NCBI (National Center for Biotechnology Information), utilizando o filtro Genome e a busca pela sequência FASTA (Rangwala *et al.*, 2021). As informações relativas da montagem do genoma estão apresentadas na tabela 1.

Tabela 1: Dados da montagem do genoma

Acesso	Cobertura do Genoma	Genes (CDS)	Conteúdo GC	Nível da Montagem	Tamanho do Genoma
GCA_004117975.1	185,0x	8123	46,5	Andaime	27,1MB

Fonte: Os autores, 2025.

- **Predição dos metabólitos secundários**

Os clusters gênicos de metabólitos secundários foram identificados através da ferramenta de predição antiSMASH fungal (Antibiotics and Secondary Metabolite Analysis Shell), versão 8.0.1, utilizando os critérios para fungos (Blin *et al.*, 2025). Tendo como input o arquivo formato FASTA do genoma, as operações empregadas para a precisão da descoberta dos cluster foi o relaxed para maior expressividade de BGC, com todos os algoritmos extras como Known Cluster Blast, Active Site Finder, Cluster Blast, Cluster Pfam analysis, Sub Cluster Blast e Pfam-based GO term annotation ofertados pelo antiSMASH ativados (Blin *et al.*, 2021). A máquina utilizada para as análises apresenta as seguintes especificações: 11th Gen Intel(R) Core(TM) i5-11320H de processador e 16.0 GB de RAM.

Esta análise, foram realizadas filtragem e classificação dos agrupamentos a partir do predito pelo antiSMASH com a utilização de banco de dados auxiliares como o MIBiG (Minimum Information about a Biosynthetic Gene cluster), versão 2.0, compreendido como um repositório de BGC (Kautsar *et al.*, 2020); e o banco de dados Pfam (Protein Families Database) para exame dos genes associados a síntese de produtos naturais e suposições de funções biológicas (El-gebali *et al.*, 2019).

- **Predição das rotas bioquímicas**

Para a caracterização da rota metabólica foi utilizado o programa KEGG (Kyoto Encyclopedia of Genes and Genomes), versão 111.0, no qual foi usado os arquivos de referência das sequências genomas e bioativos identificados, conjuntamente. O filtro adotado baseou no módulo pathway kegg relativo à via principal da biossíntese de maior expressão e similaridade predita pelo antiSMASH (Kanehisa e Sato, 2020).

Potencial biotecnológico com base nas literaturas

Os efeitos dos bioativos preditos foram sustentados a partir das revisões no banco de dado PubChem (Kim *et al.*, 2023). Para as buscas funcionais dos compostos, foi usado a nomenclatura dos bioativos como referência (“squalestatin” e “clavaric acid”).

Além disso, para as buscas do potencial dos bioativos a partir das revisões literárias foram acessadas pelas Plataformas do Periódicos Capes, Scopus, Pubmed e Google Acadêmico, utilizando os nomes dos supostos compostos, com filtros para obtenção de artigos, dissertações e teses nas Língua Portuguesa e Inglesa, considerando todo o intervalo de ano desde a primeira publicação. Os estudos selecionados correspondem os de maior sucesso registrado, relativo somente aos elementos preditos pelo presente trabalho.

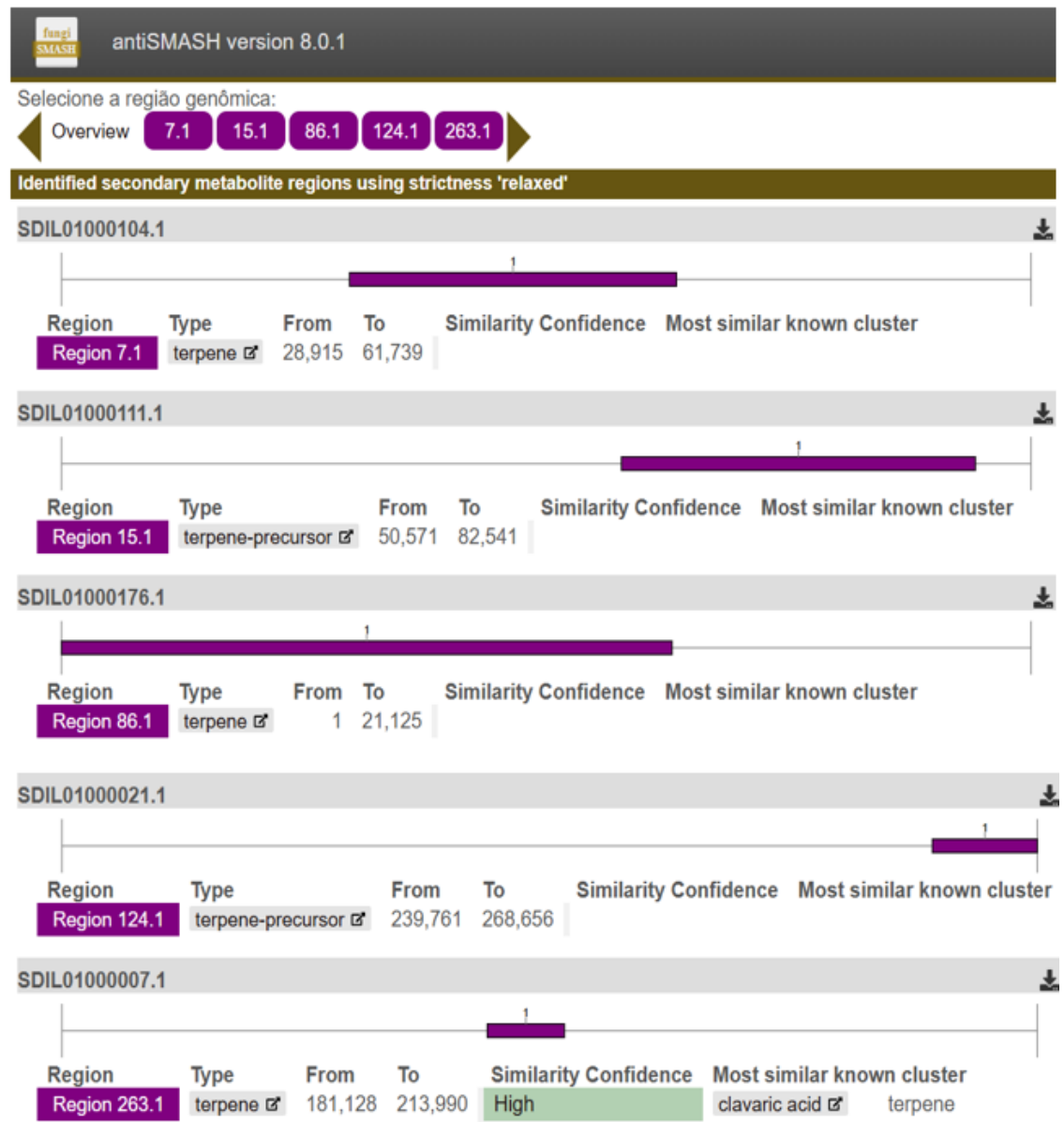
Resultados e discussões

- **Os BGCs de *Tremella mesenterica***

A partir da análise do genoma de *T. mesenterica* utilizando o programa antiSMASH, foi possível constatar 5 regiões de maior expressão associados a clusters biossintéticos, identificados como Região 7.1 (R1), Região 15.1 (R2), Região 86.1 (R3), Região 124.1 (R4) e Região 263.1 (R5), todos contendo agrupamentos para produção de Terpenos.

Dentre os fragmentos genômicos, a região R5 apresentou 100% de similaridade com os BGC e funções já determinadas no banco de dados do programa, sendo o agrupamento mais semelhante ao composto ácido clavárico (AC). Já das regiões R1 à R4 não apresentou grande similaridade com nenhum agrupamento semelhante do banco de dados, levando a supor que os BGC destas regiões podem estar associadas a vias de biossíntese de novas moléculas ou que pertencem a vias não caracterizadas de moléculas já conhecidas, entretanto, os perfis dos BGC destas regiões direcionam conjuntamente para produção de compostos como esqualestatina (ácido zaragózico ou AZ).

Figura 1: Mineração genômica: região, tipo de metabolitos, sequência de DNA e composto similar. Identificação de bioativos candidatos com base na similaridade de grupos de genes biossintéticos na base de dados do software.



Fonte: Output AntiSMASH, 2025.

Na R1 e R2 determinou-se, especificamente, dez genes; na R3 determinou-se 7 genes; na R4, onze genes; enquanto na R5, nove genes, atuantes no metabolismo secundário do fungo analisado. Tais genes foram nomeados no Identificador, no qual cada uma das sequências associou a síntese de produtos mais semelhante do banco de dados, participantes da formação dos bioativos preditos (Tabela 2, 3, 4, 5 e 6).

Tabela 2: Dados do BGC de classe tetraterpeno (subclasse carotenoide C40).

Identificador	Tamanho do nucleotídeo	Tamanho do Aminoácido	Categoria	Produto
M231_06493	2802	933	outro	aminopeptidase 2
M231_06494	1335	444	biossintético-adicional	GTP ciclo-hidrolase I Proteína
M231_06495	288	95	outro	ribossomal 60S L37a
M231_06496	2418	805	outro	proteína hipotética
M231_06497	2367	788	outro	proteína hipotética
M231_06498	2313	770	biossintético	proteína hipotética
M231_06499	1521	506	outro	proteína hipotética
M231_06500	2328	775	outro	proteína hipotética
M231_06501	2841	946	outro	proteína hipotética
M231_06502	1386	461	outro	proteína de membrana

Fonte: Os autores, 2025. A denominação “outro” classifica os gens e sua função como não pertencente a nenhuma das categorias definidas no banco de dados do programa. Proteína hipotética: prevista com base em dados genômicos, porém, a função ou existência ainda não foi confirmada em experimentos.

Tabela 3: Dados do BGC de precursor de diterpeno.

Identificador	Tamanho do nucleotídeo	Tamanho do Aminoácido	Categoria	Produto
M231_06696	561	186	outro	proteína hipotética
M231_06697	999	332	outro	proteína hipotética proteína associada
M231_06698	1113	370	outro	ao receptor de serina-treonina quinase

M231_06699	402	133	outro	ribonucleoproteína nuclear pequena D1
M231_06700	1680	559	outro	proteína hipotética
M231_06701	867	288	outro	proteína hipotética geranilgeranil
M231_06702	954	317	biossintético	difosfato sintase, tipo III
M231_06703	1098	365	biossintético-adicional	proteína hipotética
M231_06704	1242	413	outro	proteína hipotética
M231_06705	564	187	outro	proteína hipotética

Fonte: Os autores, 2025. A denominação “outro” classifica os gens e sua função como não pertencente a nenhuma das categorias definidas no banco de dados do programa. Proteína hipotética: prevista com base em dados genômicos, porém, a função ou existência ainda não foi confirmada em experimentos.

Tabela 4: Dados do BGC de triterpeno (esqualeno).

Identificador	Tamanho do nucleotídeo	Tamanho do Aminoácido	Categoria	Produto
M231_07848	186	61	outro	Subunidade J da ATPase transportadora de H ⁺ tipo F
M231_07849	978	325	outro	proteína hipotética
M231_07850	1476	491	biossintético	farnesil-difosfato farnesiltransferase
M231_07851	1332	443	outro	proteína hipotética
M231_07852	1524	507	outro	proteína hipotética
M231_07853	609	202	outro	proteína hipotética
M231_07854	5991	1996	outro	proteína hipotética

Fonte: Os autores, 2025. A denominação “outro” classifica os gens e sua função como não pertencente a nenhuma das categorias definidas no banco de dados do programa. Proteína hipotética: prevista com base em dados genômicos, porém, a função ou existência ainda não foi confirmada em experimentos.

Tabela 5: Dados do BGC do precursor de sesquiterpeno.

Identificador	Tamanho do nucleotídeo	Tamanho do Aminoácido	Categoria	Produto
M231_02527	1071	356	outro	proteína 1 de tRNA 2-tiolação citoplasmática
M231_02528	1284	427	outro	mRNA metiltransferase
M231_02529	486	161	outro	proteína de montagem de clatrina
M231_02530	393	130	outro	enzima conjugadora de ubiquitina E2 W
M231_02531	1545	514	outro	proteína hipotética
M231_02532	1314	437	outro	proteína hipotética
M231_02533	1059	352	biossintético	farnesil difosfato sintase
M231_02534	399	132	outro	proteína hipotética
M231_02535	966	321	outro	subunidade complexa arp2/3 de 34 kDa
M231_02536	1029	342	outro	proteína hipotética
M231_02537	1032	343	outro	proteína hipotética

Fonte: Os autores, 2025. A denominação “outro” classifica os gens e sua função como não pertencente a nenhuma das categorias definidas no banco de dados do programa. Proteína hipotética: prevista com base em dados genômicos, porém, a função ou existência ainda não foi confirmada em experimentos.

Tabela 6: Dados do BGC do triterpeno (subclasse esteroide C30).

Identificador	Tamanho do nucleotídeo	Tamanho do Aminoácido	Categoria	Produto
M231_01091	423	140	outro	proteína hipotética
M231_01092	795	264	outro	proteína hipotética
M231_01093	975	324	outro	proteína hipotética
M231_01094	1941	646	outro	proteína hipotética
M231_01095	759	252	biossintético-adicional	hidroxiacilglutaciona hidrolase
M231_01096	2166	721	biossintético	lanosterol sintase
M231_01097	4272	1423	outro	proteína hipotética

M231_01098	909	302	outro	proteína hipotética citocromo c
M231_01099	1122	373	outro	peroxidase, mitocondrial

Fonte: Os autores, 2025. A denominação “outro” classifica os gens e sua função como não pertencente a nenhuma das categorias definidas no banco de dados do programa. Proteína hipotética: prevista com base em dados genômicos, porém, a função ou existência ainda não foi confirmada em experimentos.

• **Caracterização da rota metabólica dos compostos**

Como ressaltado anteriormente, as regiões do genoma de *T. mesenterica* determinou a predição de substâncias terpenoides. Os compostos dessa natureza antes de sua produção apresentam uma rota inicial ou rota principal que possuem papel fundamental na síntese da estrutura principal dos terpenos.

Nesse sentido, os dados resultantes do KEGG descreveram a rota inicial do metabólito secundário a partir de uma atividade derivada do processo de glicólise, no qual partindo da Acetil-CoA, uma série de cinco reações são ocasionados até a formação do Difosfato de isopentenil. Nesta etapa, as enzimas atuantes definidas foram C-acetiltransferase, hidroximetilglutaril-CoA sintase, hidroximetilglutaril-CoA redutase, fosomevalonato quinase e difosomevalonato descarboxilase.

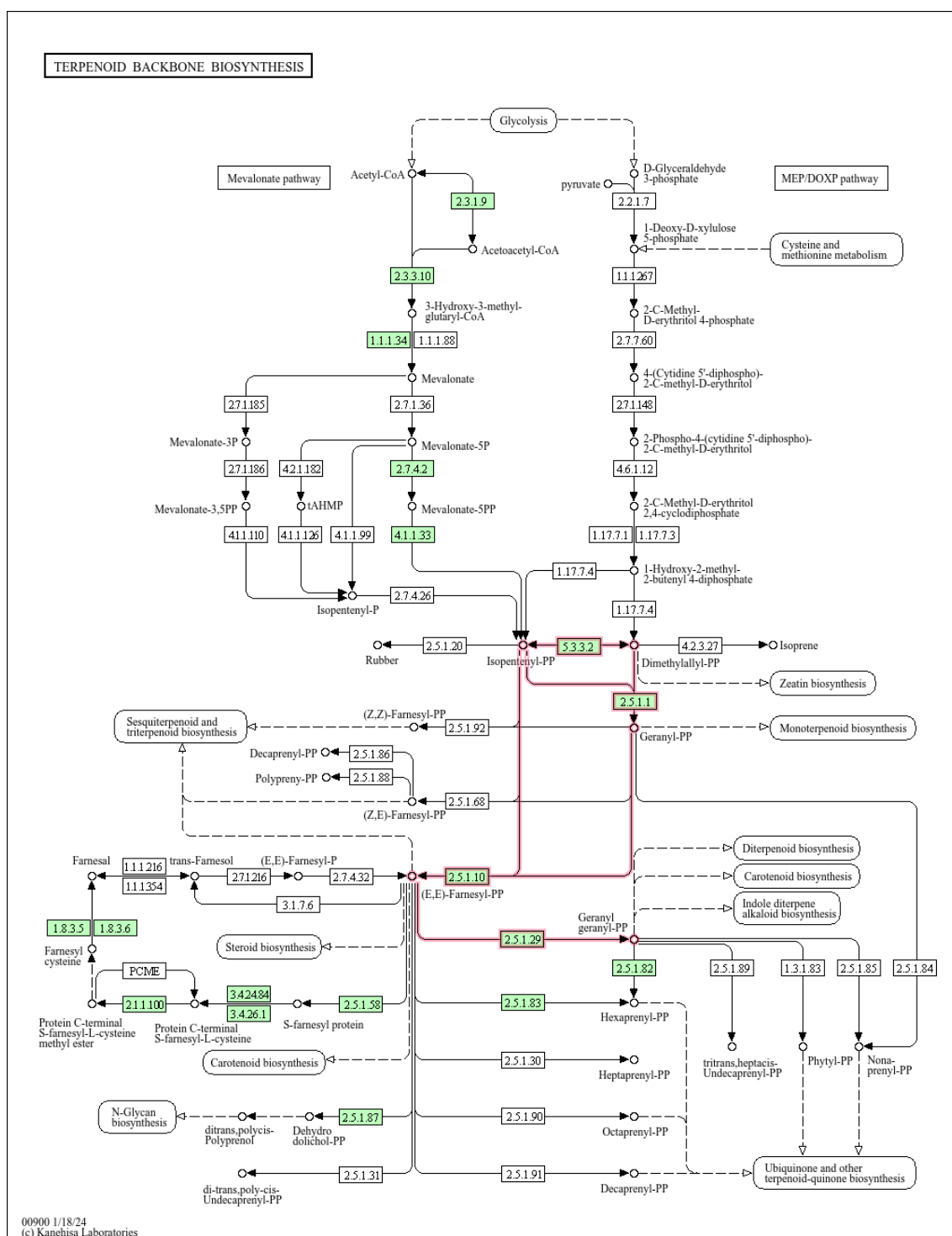
Com o Isopentenil-PP foi ilustrado duas vias alternativas: 1) a formação do Geranil-PP até o Nonaprenil-PP (neste caminho não foi caracterizado as enzimas participantes neste fragmento de via) e; 2) formação do Farnesil-PP até o Hexaprenil-PP, com participação das enzimas geranilgeranil difosfato sintase e hexaprenil difosfato sintase. Sendo ambas as cadeias chave que direcionam a corrente biossintética para síntese final de compostos terpenóides e quinonas.

A via metabólica específica do bioativo predito pelo antiSMASH, ácido clavárico, até então, não está integralmente descrito na literatura, no entanto, como trata de um bioativo triterpeno, enquadra-se na rota desta classificação. Ou seja, a cadeia principal resultante atua, conjuntamente, com trans-Farnesil difosfato para formar o presqualeno-PP que, por sua vez, formará o esqualeno ou esqualeno-2,3-epoxido, por meio das atividades das farnesil-difosfato farnesiltransferase e esqualeno monooxigenase compondo as estruturas bases complementares deste bioativo.

Quanto a via metabólica do bioativo, esqualestatina, não houve resultantes especificadas de descrição pelo KEGG, porém, pelo fato do composto ser de natureza

policetídeo, condiz com rota desta classe estimada pelo programa. Isto é, a partir de moléculas como metilmalonil-CoA, propanoil-CoA e cetil-CoA advinda da via primária do metabolismo, são rearranjadas por meio blocos enzimáticos multimodulares denominados sintases de policetídeos (PKS) que promovem condensação das unidades de acetila e malonila (similar a biossíntese de ácidos graxos). Algumas das enzimas modificadoras atuantes na síntese de policetídeos são a cetorredutase (KR), responsável pela redução da carbonila ao álcool, a desidratase (DH), que elimina água levando a formação da olefina e, a enoila redutase (ER), que converte a olefina na cadeia saturada.

Figura 2: Via principal da biossíntese dos terpenos. Descrição: 2.3.1.9 (C-acetiltransferase); 2.3.3.10 (hidroximetilglutaril-CoA sintase); 1.1.1.34 (hidroximetilglutaril-CoA redutase); 2.7.4.2 (fosfomevalonato quinase); 4.1.1.33 (difosfomevalonato descarboxilase); 2.5.1.10, 2.5.1.29 (geranilgeranil difosfato sintase); e 2.5.1.82 (hexaprenil difosfato sintase). A linha destacada em vermelho retrata o caminho chave de reações e moléculas importantes para síntese de metabólitos secundários (destaque para Policetídeos e Terpenóides).



Fonte: Output KEGG, 2025.

A esqualostatina, também conhecida como ácido zaragózico, família a qual pertence, são compostos produzidos por alguns fungos filamentosos. Os primeiros trabalhos sobre esta substância foram em 1993, com isolados de *Sporormiella intermedia*, *Leptodontium elatius* e *Amauroascus niger*. Quimicamente, o ácido Zaragózico é um

ácido carboxílico, no qual suas cadeias laterais podem variar em 1-álquil e 6-ácil (Bergstrom *et al.* 1995; Dufresne *et al.*, 1993).

Alguns estudos e revisões relatam que este composto possui atividade contra a enzima fitoeno sintase em vegetais, portanto, inibem a síntese de carotenoide (Kamdern e Boyom, 2023; Neudert *et al.*, 1998; Baxter *et al.*, 1992). Em outros organismos, esta substância atua como inibidor da enzima esqualeno sintase em mamíferos, assim, impede a síntese de colesterol (Kamdern e Boyom, 2023; Bergstrom *et al.*, 1995) ou podem apresentar efeitos anti Tripanossomatídeos (Kamdern e Boyom, 2023; Silva, 2012).

Nessa perspectiva, a potencial produção de AZ por *T. mesenterica* pode ser direcionada para aplicações farmacológicas. Um estudo de Silva (2012) abordou a possibilidade de desenvolvimento de drogas antimaláricos a partir da esqualostatina. A autora não deixa claro a fonte da extração do composto, porém, seu trabalho revelou o efeito antiparasitário moderado contra *Plasmodium falciparum* geneticamente modificado, a partir de experimentos *in vitro* do microrganismo, visto que foi observado a inibição do desenvolvimento e crescimento do protozoário, bem como alcançou resultados expressivos associando o bioativo com outras substâncias como nerolidol e com risedronato, nos qual ambas as concentrações observou cerca de 50% de inibição, demonstrando interação sinérgicas entre os compostos. Ademais, neste mesmo cenário, Gabriel *et al.* (2015) observaram a redução da biossíntese de fitoeno (PTY) e morte dos *P. falciparum* durante o ciclo intraeritrocítico usando o bioativo esqualostatina e apontando, portanto, que a PTY é uma enzima chave para a biossíntese de carotenoides do parasita que é fundamental para o crescimento intraeritrocítico.

Outra possível utilidade deste bioativo, é em uso medicamentoso para tratamento da hipercolesterolemia em seres humanos. Bergstrom *et al.* (1995), isolou as esqualenos sintases de fígado de camundongos e, a partir das aplicações *in vitro* de AZ foi notado redução da atividade enzimática. O mecanismo avaliado consistiu na inibição da esqualeno sintase pelo AZ que é competitiva em relação ao F-PP (pirofosfato), sendo uma das justificativas associada a similaridade estrutural entre o AZ e o F-PP (ambos têm um núcleo central altamente ácido com duas longas caudas hidrofóbicas anexadas) o qual direcionou à tese de que os ZAs inibem a esqualeno sintase imitando efetivamente a ligação do F-PP à enzima.

O AC é um composto triterpeno produzido por alguns fungos, no qual foi isolada de *C. truncatus* pela primeira vez em 1998, porém podem apresentar variações conforme o organismo produtor (Jayasuriya *et al.*, 1998).

Das poucas investigações acerca do ácido clavárico, os estudos o associam como uma substância antitumoral e anti-inflamatório. O experimento de Lingham *et al.* (1998), por meio de triagem aleatória de extratos de produtos naturais e usando-as em ensaio bioquímico de alto rendimento em células cancerígenas de camundongos, descobriu um composto combatente de enzimas mutantes em células tumorais, o AC. Este bioativo foi o único inibidor específico das proteínas RAS ou farnesil-proteína transferase (um catalisador essencial para oncogenes) que está ativo na maior parte dos cânceres humanos. O AC demonstrou ser competitivo nas atividades das proteínas oncogênicas, no qual interferiu, portanto, no desenvolvimento das células mutantes dos mamíferos.

Já Yazdani *et al.* (2022), através de análises bioquímicas e comparação molecular refinadas, demonstrou em teste *in vitro* de células cancerígenas de cólon humano ações moderadas de citotoxicidade de triterpenos fúngicos, trazendo a abordagem que alguns triterpenos fúngicos são potenciais agentes citotóxicos e quimiossensibilizadores para superar a resistência a múltiplas drogas por parte linhagens de células cancerígenas. Posteriormente, o AC foi apresentado com efeito anti-inflamatório em ensaios do composto em placas contendo enzimas ciclooxigenase (COX- 1 e COX-2) e outras contendo lipoxigenase (5-LOX e 15-LOX), cujas enzimas desempenham papel fundamental na inflamação, na percepção da dor, como mediadores pró-inflamatórios, sendo importante na resposta inflamatória a lesões e propagação da inflamação. Os resultados evidenciaram ação anti-inflamatória moderada do AC sob as enzimas das células de adenocarcinoma de cólon humano, principalmente, em COX-2.

Conclusão

Com base na sequência genômica é possível realizar bioprospecção de recursos genéticos, bioativos de interesse em organismos por meio de ferramentas de bioinformática. Tais artifícios são fundamentais em análises iniciais, sobretudo, em genomas de espécies recém-descobertas.

É evidente que a mineração genômica de espécie por meio *in silico* de metabólitos secundários é vantajosa no gerenciamento do tempo de análise, visto que é possível alcançar resultados favoráveis com base em ferramentas de predição quando requer respostas ou soluções em curto período de tempo ou em tempo imediato exigido.

Neste trabalho foi possível a detecção por ferramentas de bioinformática de grupos de genes envolvidos na síntese dos seguintes metabólitos secundários do fungo *T. mesenterica*: ácido clavárico e ácido zaragózico. O presente estudo aponta que

provavelmente a espécie *T. mesenterica* pode ser produtora desses importantes metabólitos para a biotecnologia e sugere que futuras pesquisas *in vitro* podem ser realizadas com o fungo para induzir a produção dos compostos.

Referências

BAXTER, A. *et al.* Squalastatin 1, a potent inhibitor of squalene synthase, which lowers serum cholesterol *in vivo*. **Journal of Biological Chemistry**, v. 267, n. 17, p.11705-11708, 1992.

BERGSTROM, J. D. *et al.* Discovery, biosynthesis, and mechanism of action of the zaragozic acids: potent inhibitors of squalene synthase. **Annual review of microbiology**, v. 49, n. 1, p. 607-639, 1995.

BLIN, K. *et al.* antiSMASH 8.0: extended gene cluster detection capabilities and analyses of chemistry, enzymology, and regulation. **Nucleic acids research**, p. gkaf334, 2025.

BLIN, K. *et al.* antiSMASH 6.0: improving cluster detection and comparison capabilities **Nucleic Acids Research** (2021) doi: 10.1093/nar/gkab335.

DE CARVALHO, A. M. **O impacto da tecnologia no mercado de trabalho e as mudanças no ambiente de produção**. Revista Evidência, v. 6, 2024.

DUFRESNE C. *et al.* Zaragozic acids D and D2: potente inhibitors of squalene synthase and of Ras farnesyl-protein transferase. **J. Nat. Prod.** 1993, 56 (11): 1923–9. doi:10.1021/np50101a009.

EL-GEBALI, S. *et al.* The Pfam protein families database in 2019. **Nucleic Acids Research**, v.47, n. D1, p. D427 D432, 2019.

JAYASURIYA, H. *et al.* Clavaric Acid: A Triterpenoid Inhibitor of Farnesyl- Protein Transferase from *Clavariadelphus truncatus*. **Journal of natural products**, v. 61, n. 12, p. 1568-1570, 1998.

KAMDEM, B. P.; BOYOM, F. F. Inhibitors of Farnesyl Diphosphate Synthase and Squalene Synthase: Potential Source for Anti-Trypanosomatidae Drug Discovery. **Drugs and Drug Candidates**, v. 2, n. 3, p. 624-652, 2023.

KANEHISA, M.; SATO, Y. KEGG Mapper for inferring cellular functions from protein sequences. **Protein Science**. 2020; 29: 28–35. <https://doi.org/10.1002/pro.3711>.

KAUTSAR, S. A. *et al.* MIBiG 2.0: A repository for biosynthetic gene clusters of known function. **Nucleic Acids Research**, v. 48, n. D1, p. D454 D458, 2020.

KIM, S. *et al.* PubChem 2023 update. **Nucleic acids research**, v. 51, n. D1, p. D1373-D1380, 2023.

LINGHAM, R. B. *et al.* Clavaric acid and steroidal analogues as Ras-and FPP- directed inhibitors of human farnesyl-protein transferase. **Journal of medicinal chemistry**, v. 41, n. 23, p. 4492-4501, 1998.

MATOS, D. F. *et al.* BIOINFORMÁTICA APLICADA ÀS CIÊNCIAS ÔMICAS: UMA BREVE REVISÃO BIBLIOGRÁFICA. TECNOLOGIA E INOVAÇÃO NA AGRICULTURA: **APLICAÇÃO, PRODUTIVIDADE E SUSTENTABILIDADE EM PESQUISA**, v. 1, p. 23-43, 2023.

NEUDERT, U.; MARTÍNEZ-FÉREZ, I. M.; FRASER, P. D; SANDMANN, G. Expression of an active phytoene synthase from *Erwinia uredovora* and biochemical properties of the enzyme. **Biochimica et Biophysica Acta (BBA)-Lipids and Lipid Metabolism**, v. 1392, n. 1, p. 51-58, 1998.

PEREIRA, T. T. S. **Bioinformática**. Editora Senac São Paulo, 2024.

RANGWALA, S. H. *et al.* Accessing NCBI data using the NCBI sequence viewer and genome data viewer (GDV). **Genome research**, v. 31, n. 1, p. 159-169, 2021.

SHELDRAKE, M. **Trama da vida: como os fungos constroem o mundo**. Fósforo, 2021.

SILVA, M. F. **Estudos in vitro de potenciais antimaláricos nos estágios intraeritrocítico de Plasmodium falciparum**. 2012. Dissertação (Mestrado em Biologia da Relação Patógeno-Hospedeiro) - Instituto de Ciências Biomédicas, University of São Paulo, São Paulo, 2012. doi:10.11606/D.42.2012.tde-23042013-122652. Acesso em: 2024-09-16.

YAZDANI, M. *et al.* Triterpenes from *Pholiota populnea* as cytotoxic agents and chemosensitizers to overcome multidrug resistance of cancer cells. **Journal of natural products**, v. 85, n. 4, p. 910-916, 2022.

CAPÍTULO 2

PROSPECÇÃO SISTEMÁTICA E APLICADA DA HERANÇA GENÉTICA DA MIOPIA UTILIZANDO FERRAMENTAS DE BIOINFORMÁTICA

SYSTEMATIC AND APPLIED PROSPECTING OF THE GENETIC INHERITANCE OF MYOPIA USING BIOINFORMATICS TOOLS

DOI: <https://doi.org/10.56001/26.9786501906539.02>

Submetido em: 05/01/2026

Revisado em: 20/01/2026

Publicado em: 03/02/2026

Amanda Geovana Pereira de Araújo

Universidade Federal de Campina Grande, Centro de Educação e Saúde, Cuité-PB

<http://lattes.cnpq.br/3946322725458190>

Carliane Rebeca Coelho da Silva

Pós-doutorado em Biotecnologia, Universidade Federal Rural de Pernambuco,

Recife-PE

<https://orcid.org/0000-0002-9462-5823>

Igor Luiz Vieira de Lima Santos

Universidade Federal de Campina Grande, Centro de Educação e Saúde, Cuité-PB

<https://orcid.org/0000-0003-3438-0446>

Resumo

A miopia é o distúrbio visual mais comum em todo o mundo. Essa acuidade ocorre devido à focalização da imagem antes desta chegar à retina. Este trabalho objetiva analisar genes envolvidos na progressão da miopia com ferramentas de bioinformática, realizar estudos comparativos, e construir primers para estudos futuros. Foi realizada uma metodologia com características qualitativas e quantitativas, utilizando técnicas de padronização, selecionando, segundo os bancos de dados, os artigos mais representativos acerca do tema, após isso, os primers foram confrontados com programas de análises, na tentativa de confirmar a especificidade do gene RASGRF1 com a progressão da doença. Conforme a pesquisa do NCBI, 22 genes foram descritos influenciadores da miopia hereditária, já o consórcio internacional de erro refrativo identificou 119 fatores genéticos para a miopia. Em contrapartida, o gene utilizado para construção de primers, o RASGRF1, não foi descrito em nenhum desses bancos. Porém, foi descrito na literatura como um estudo novo influenciador da miopia. Nesse contexto, cinco séries de primers apresentaram ser insatisfatórias, mas, outras cinco séries apresentaram fatores satisfatórios, o que indicou uma alta similaridade entre o gene e os indivíduos comparados. Além disso, foi realizado um apanhado

BIOINFORMÁTICA UMA CIÊNCIA SEM FRONTEIRAS, 1ª Edição. EDITORA SCIENCE – 2026.

geral de genes potencialmente influenciadores da miopia, como o RASGRF1, APLP2, LEPREL1, LAMA2, CPSF1, GJD2, ZIC2 entre outros. Conclui-se, então que foram sistematizados vários genes potencialmente envolvidos nesse acometimento e além disso, cinco séries de primers foram satisfatórias com os resultados para o gene RASGRF1, os quais, servirão de base para a estruturação real dos possíveis testes in vitro futuros.

Palavras-chave: Distúrbio; primers; gene

Abstract

Myopia is the most common visual disorder in the world. This concern is due to the focus of the image before access to the retina. This work aims to analyze genes in the progress of myopia with bioinformatics tools, perform comparative studies, and build primers for future studies. Are presented a methodology with qualitative and quantitative characteristics, using standardization techniques, after databases, the most representative articles on the subject, after which the primers were confronted with the analysis programs in an attempt to confirm a specificity of the RASGRF1 gene with progression of the disease. According to a NCBI survey, 22 genes were inherited from hereditary myopia. But the international consortium of refractive error identified 119 genetic factors for myopia. In contrast, the gene used to construct primers, RASGRF1, was not described in any of these banks. However, it was in the literature as a new study that influenced myopia. This context, five series of primers, is unsatisfactory, but, other five, the module of satisfactors, which indicated a similarity between the gene and the compared. In addition, a general set of myopia influencing genes, such as RASGRF1, APLP2, LEPREL1, LAMA2, CPSF1, GJD2, ZIC2, and others were performed. It is concluded that these were systematized several more effective genes in this process and in addition, the five series of primers were satisfactory with the results of the RASGRF1 gene, which will serve as a basis for a real structuring of the possible future tests in vitro.

Keywords: Distúrbio; primers; gene.

Introdução

A miopia é o distúrbio visual mais comum em todo o mundo. Essa problemática visual ocorre devido à focalização da imagem antes desta chegar à retina, a qual vai converter o estímulo luminoso em estímulo nervoso e endereçá-lo ao cérebro, para que as imagens sejam interpretadas. Por esta razão, uma pessoa míope visualiza os objetos mais distantes como se estivessem desfocados, e os objetos mais próximos, com nitidez.

Dessa forma, a miopia pode ser classificada como axial, que é causada pelo alongamento do globo ocular; não síndrômica, se essa ocorrer sem fatores externos, e miopia síndrômica, caso seja associada ao sinal de outra doença ocular ou sistêmica. Além disso, um pequeno número de pessoas desenvolve a miopia degenerativa, sendo uma importante causa de cegueira nos países desenvolvidos.

A incidência de miopia está aumentando significativamente. “Quase 80% dos indivíduos em algumas populações sofrem com esse problema comprometedor da visão” (BENAVENTE-PEREZ et al., 2016). Segundo (ZHANG, 2015), embora existam numerosos métodos de melhorar a visão turva associada com miopia, inclusive usando lentes corretivas ou refrativa e cirurgia, possíveis intervenções para a patogênese da miopia foram intensamente estudadas.

Tendo isso em vista, pesquisas afirmam que a miopia mais frequente ocorre e se desenvolve durante a idade escolar em muitas áreas do mundo, como nos Estados Unidos, Singapura, Austrália e Taiwan. Ademais, é importante salientar que outras anomalias oculares, assim como a miopia, têm o caráter genômico envolvido como a coloração dos olhos, a heterocromia, o retinoblastoma, o glaucoma, a síndrome de Marfan, entre tantos outros.

Só na década de 60 começou-se a pensar sobre os determinantes hereditários da miopia, elucidando a interação entre fatores genéticos e ambientais. Os indícios das bases genéticas hereditárias estão correlacionados com a patologia, acarretando problemas no aprendizado do indivíduo mais jovem e, conseqüentemente, afetando a qualidade de vida destes. Por outro lado, pessoas com idade mais avançada também são acometidas pela miopia e afetadas por fatores ambientais, comprometendo suas tarefas e impondo restrições para o desenvolvimento de atividades físicas. Isto gera os problemas psicossociais potenciais dessa patologia, como exemplo disso, a depressão.

A genética tem grande influência sobre a miopia, e embora básica, é uma ciência de entendimento complexo e aplicado. Na prática, atualmente, em todas as áreas da vida contemporânea, existe a engenharia genética enredada, e não é diferente com a miopia, que tem caráter hereditário e associação genômica ampla. Essa associação estuda os genes que estão envolvidos nos diversos tipos de miopia e seus caminhos específicos, dentre eles, estão os genes que se submetem ao transporte de íon, ao regulamento do potencial da membrana e a atividade de canal, os que são responsáveis pelo regulamento do oxigênio nos olhos entre tantos outros.

Assim sendo, estudos apontam que crianças com os pais míopes são mais prováveis de herdar a doença em comparação a aqueles com pais não míopes. Porém, (GOLDSCHMIDT, 2014) explica que ter dois pais míopes, geralmente representa um maior risco do que ter apenas um, porém, embora essas correlações sejam consistentes, com base na genética da miopia, elas não a estabelecem, pois, além dos erros refrativos entre pais e filhos e os valores da hereditariedade, também existem aspectos ambientais envolvidos, e que também contribuem para o desenvolvimento da miopia, como sua influência na educação.

Por conseguinte, este trabalho pauta-se na responsabilidade de aprofundar e organizar os conhecimentos sobre a miopia, uma vez que o tema é bastante complexo e muitos podem ser os genes envolvidos e acima de tudo, como eles podem atuar no acometimento e evolução da doença que atinge muitas pessoas no Brasil e no mundo. O objetivo geral do trabalho foi analisar sistematicamente os genes envolvidos nas vias de progressão da miopia utilizando ferramentas de bioinformática. Dessa forma, a pesquisa é justificada pelo interesse geral sobre esta acuidade visual e expectativa pela possibilidade da melhoria da qualidade de vida dos

BIOINFORMÁTICA UMA CIÊNCIA SEM FRONTEIRAS, 1ª Edição. EDITORA SCIENCE – 2026.

indivíduos, utilizando as ferramentas e informações necessárias acerca do problema. O estudo é relevante para a ciência por possibilitar a compilação de dados a respeito das influências hereditárias sobre a miopia, uma vez que é generalista e disperso o conhecimento sobre os genes envolvidos e sua forma de atuar no desenvolvimento da doença. Desse modo, é possível ter uma visão mais global a respeito do tema possibilitando um maior conhecimento para a possível realização de tratamentos mais efetivos atenuando a progressão e grau do problema.

Metodologia

- **Local e Equipe de Estudo**

O trabalho foi realizado nas dependências do Centro de Educação e Saúde da Universidade Federal de Campina Grande com o grupo de pesquisa BASE (Biotecnologia Aplicada a Saúde e Educação). O grupo de pesquisa e a Universidade ofereceram toda a infraestrutura computacional e de rede necessária para a realização do trabalho.

- **Metodologia Geral**

Trata-se de um estudo exploratório com potencial tecnológico com características qualitativas e quantitativas como ferramenta subsidiária de informações essenciais para atingir os objetivos propostos. Foram utilizadas técnicas padronizadas de coleta de dados utilizando como base os bancos de dados de informações genéticas.

Nesses bancos foram identificados artigos tratando a respeito do tema, bem como os conteúdos sobre aconselhamento genético sempre visando o tema proposto. Na base de dados mais sólida consultada até o momento foram catalogados 2.071 artigos versando sobre o referido tema utilizando o termo “hereditary myopia”, destes foram selecionados os artigos mais recentes e representativos da influência genética sobre o tema.

- **Análises das Sequências Envolvidas com a Miopia**

Foi construída uma base de dados composta por centenas de seqüências dos genes identificados para miopia provenientes dos Bancos de Dados GenBank (NCBI – National Center for Biotechnology Information) (<https://www.ncbi.nlm.nih.gov/>) e das vias bioquímicas (KEGG – Kyoto Encyclopedia of Genes and Genomes) (<http://www.genome.jp/kegg/>).

Após alinhadas pelo programa Clustal-Omega (CHOJNACKI et al., 2017) (<https://www.ebi.ac.uk/Tools/msa/clustalo/>) elas foram visualizadas pelo programa BioEdit v7.0.4 (Tom Hall, Ibis Therapeutics, Carlsbad, CA, USA) (HALL, 1999) (<http://www.mbio.ncsu.edu/BioEdit/bioedit.html>) ou pelo programa MUSCLE

(Multiple Sequence Comparison by Log- Expectation), MUSCLE é utilizado para alcançar melhor precisão média e melhor velocidade do que o ClustalW2 ou T-Coffee, dependendo das opções escolhidas (<https://www.ebi.ac.uk/Tools/msa/muscle/>).

As seqüências foram submetidas aos programas para a construção dos primers do NCBI (Primer 3 - Pick Primers) (<http://primer3.ut.ee/>) (YE et al., 2012). Os pares de oligonucleotídeos gerados foram confrontados novamente com os bancos de dados e com programas de análises como o Oligo Analysis Tool para confirmação da especificidade dos mesmos para os genes de interesse na progressão da miopia com a finalidade de estudos futuros, tudo isto para escolher os que melhor se adequariam aos pretensos testes *in vitro*.

- **Viabilidade**

Este trabalho está em acordo com a Lei Nº 13.123, de 20 de maio de 2015 que define “patrimônio genético” como “informação de origem genética de espécies vegetais, animais, microbianas ou espécies de outra natureza, incluindo substâncias oriundas do metabolismo destes seres vivos” e “conhecimento tradicional associado” como “informação ou prática de população indígena, comunidade tradicional ou agricultor tradicional sobre as propriedades ou usos diretos ou indiretos associada ao patrimônio genético”. Este trabalho segue as normas do CGEN - Conselho de Gestão do Patrimônio Genético.

Desenvolvimento

- **Resultados E Discussões Finais**

Com as consultas aos bancos de dados de seqüências genômicas e de conhecimento associado ao componente genético da miopia, foram constatados um total de 22 genes descritos como influenciadores da miopia hereditária potencialmente relevantes para este acometimento. A Tabela 1 mostra todos os genes descritos como envolvidos no acometimento de algum grau de miopia em humanos e seus respectivos, símbolos, substantivos, descrição de função, localização genômica, cromossomo, orientação e éxons.

É possível notar, apenas analisando a tabela, que existe uma multilocalidade bastante ampla na distribuição gênica potencialmente influenciadora da miopia onde os genes estão distribuídos nos cromossomos 1, 3, 4, 6, 7, 8, 9, 11, 12, 13, 17, 18, 19, 20, 21 e X. Estes genes demonstram o grau de dificuldade em entender e identificar certos tipos de miopia, bem como seu componente genético associado, visto que existe uma gama bastante grande de possibilidades e isso só poderá ser elucidado com o aumento dos estudos a respeito do tema. Singh e Tyagi (2018) realizaram uma análise multivariada dos genes e da genética envolvida

BIOINFORMÁTICA UMA CIÊNCIA SEM FRONTEIRAS, 1ª Edição. EDITORA SCIENCE – 2026.

com as doenças oculares utilizando uma metodologia genômica aplicada. Eles identificaram componentes genéticos associados aos mais diversos tipos de desordens oculares ligadas ou não a idade, os seus resultados podem ser vistos na Tabela 2, onde um total de 91 genes foram descritos como influenciadores dos acometimentos oculares em diversos graus.

O consórcio internacional de erro refrativo e miopia (CREAM) publicou recentemente o maior estudo genético mundial de miopia, que identificou 119 fatores genéticos para a miopia. Segundo a pesquisa, isso quadruplicou o número de fatores de riscos genéticos conhecidos que desempenham um papel em todos os tipos de células da retina, contudo a quantidade de genes e de cromossomos envolvidos descritos vem crescendo diariamente. A maioria deles está envolvida no processamento de luz, pigmentação da retina e desenvolvimento ocular. Isso tudo pode apoiar a suposição de que a luz insuficiente é um importante fator no desenvolvimento da miopia (FLITCROFT et al., 2018).

Com base nos argumentos citados, foi visto que a miopia é a patologia da visão mais comum e quando está em uma dioptria (grau) baixa, dificulta ainda mais a percepção do problema. Sendo assim, a miopia classifica-se de três formas: a baixa, a moderada e a alta. Todas podem sofrer influências genéticas e ambientais que devem ser avaliadas para analisar a condição gene-específica relacionada aos três estágios da doença. Com base nos dados acima e em tantos outros disponíveis nos bancos de dados, foi constatado que o gene *RasGRF1* está fortemente ligado ao desenvolvimento molecular da alta miopia, apesar deste gene não estar descrito nas tabelas de Singh e do NCBI. O MedGen, banco de dados médicos do NCBI, ainda reporta 6 variantes clínicas responsáveis pelo desenvolvimento ocular anormal, além de outros acometimentos ao indivíduo: A Síndrome de Stickler Tipo I e mais duas variantes dela a Síndrome de Stickler Tipo I Ocular não Sindrômica e a Síndrome de Stickler Pura; Síndrome de Marshall; Glaucoma de Ângulo Aberto Primário Juvenil Série I; e por fim A Síndrome Monossômica Parcial do Cromossomo 5 (Cri-du-Chat). O banco de dados OMIM (Online Mendelian Inheritance in Human) ainda revela 70 acometimentos humanos influenciadores da miopia conforme Tabela 3.

Não existe uma única causa para a miopia, podendo a etiologia ser genética ou ambiental. Três fatores são importantes para o seu desenvolvimento, o primeiro é a relação entre o esforço visual para perto e uma fraca acomodação, o segundo está associado a predisposição hereditária e o terceiro fator é a relação entre a pressão intra-ocular e debilidade escleral. Dessa forma, existe maior influência dos dois primeiros fatores, em graus diferentes, no período inicial da miopia, enquanto o terceiro fator é mais frequente nos graus mais avançados, causando sua progressão.

Em relação ao primeiro fator, o esforço visual resultante do trabalho para perto, somado à influência de distúrbios refrativos acarreta uma desfocalização óptica, durante o período de crescimento do globo, fazendo com que o eixo ântero-posterior se tornasse mais longo. Esse tipo de miopia não costuma ultrapassar 3 graus. Quanto ao segundo fator, de predisposição hereditária, estudos mostram que a herança pode ser autossômica dominante, recessiva e poligênica. Na miopia autossômica dominante este acometimento se desenvolve tardiamente na infância e usualmente não atinge altos graus. Já na miopia autossômica recessiva é característica de comunidades com alta frequência de consanguinidade, mas também está relacionada a alguns casos esporádicos. Analisando estudos, estabeleceram que o desenvolvimento da miopia segue um modelo poligênico e multifatorial.

Na miopia patológica existe sempre uma história familiar, havendo evidências de ser decorrente de herança autossômica recessiva. No terceiro fator, a debilidade da esclera pode estar relacionada a fibrilogênese defeituosa, a qual pode ser congênita ou resultante de doenças sistêmicas. Assim, existe uma resposta inadequada ao crescimento do globo e subsequente alongamento, mesmo com a pressão intra-ocular normal. O alongamento se localiza principalmente na porção posterior do globo, produzindo alterações adversas da coróide e retina.

Tabela 1. Genes potencialmente envolvidos em algum grau do acometimento hereditário da miopia e suas características.

Organismo	Símbolo	Substantivos	Descrição do Gene	Mapa de Localização	Cromossomo	Orientação da Fita de DNA	Contagem de Éxons
Homo sapiens	BMP2	BDA2A, SSFSC, BMP2	bone morphogenetic protein 2	20p12.3	20	plus	3
Homo sapiens	CFH	AHUS1, AMBP1, ARMD4, ARMS1L3, FH, FHL1, HF, HF1, HF2, HUS, CFH	complement factor H	1q31.3	1	plus	25
Homo sapiens	COL11A2	DFNA13, DFNB53, FBCG2, HKE5, OSMEDA, OSMEDB, PARP, STL3	collagen type XI alpha 2 chain	6p21.32	6	minus	70
Homo sapiens	COL1A1	EDSARTH1, EDSC, OI1, OI2, OI3, OI4	collagen type I alpha 1 chain	17q21.33	17	minus	51
Homo sapiens	CRYAA	CRYA1, CTRCT9, HSPB4	crystallin alpha A	21q22.3	21	plus	4
Homo sapiens	FHIT	AP3Aase, FRA3B	fragile histidine triad	3p14.2	3	minus	20
Homo sapiens	GATA4	ASD2, TACHD, TOF, VSD1	GATA binding protein 4	8p23.1	8	plus	11
Homo sapiens	HGF	DFNB39, F-TCFB, HPTA, SF, HGF	hepatocyte growth factor	7q21.11	7	minus	20
Homo sapiens	IGF1	IGF, IGF-I, IGFI, MGF	insulin like growth factor 1	12q23.2	12	minus	7
Homo sapiens	MET	AUTS9, DFNB97, HGFR, RCCP2, c-Met	MET proto-oncogene, receptor tyrosine kinase	7q31.2	7	plus	24
Homo sapiens	MMP1	CLG, CLGN	matrix metalloproteinase 1	11q22.2	11	minus	10
Homo sapiens	MMP9	CLG4B, GELB, MANDP2, MMP-9	matrix metalloproteinase 9	20q13.12	20	plus	13
Homo sapiens	MYOC	GLC1A, GPOA, JOAG, JOAG1, TIGR	myocilin	1q24.3	1	minus	3
Homo sapiens	OPA1	BERHS, MGM1, MTDPS14, NPG, NTG, largeG	OPA1 mitochondrial dynamin like GTPase	3q29	3	plus	32
Homo sapiens	PAX6	AN, AN2, ASGD5, D11S812E, FVH1, MGDA, WAGR	paired box 6	11p13	11	minus	17
Homo sapiens	PROM1	AC133, CD133, CORD12, MCDR2, MSTP061, PROML1, RP41, STGD4	prominin 1	4p15.32	4	minus	35
Homo sapiens	RPGR	COD1, CORDX1, CRD, PCDX, RP15, RP3, XLRP3, orf15	retinitis pigmentosa GTPase regulator	Xp11.4	X	minus	19
Homo sapiens	TGFB1	CED, DPD1, IBDIMDE, LAP, TGFB, TGFbeta	transforming growth factor beta 1	19q13.2	19	minus	7
Homo sapiens	TIMP2	CSC-21K, DDC8	TIMP metalloproteinase inhibitor 2	17q25.3	17	minus	5

Homo sapiens	TJP2	C9DUPq21.11, DFNA51, DUP9q21.11, PFIC4, X104, ZO2	tight junction protein 2	9q21.11	9	plus	30
Homo sapiens	TTR	ATTR, CTS, CTS1, HEL111, HsT2651, PALB, TBPA	transthyretin	18q12.1	18	plus	4
Homo sapiens	VDR	NR1H1, PPP1R163	vitamin D receptor	12q13.11	12	minus	12

Fonte: NCBI (<https://www.ncbi.nlm.nih.gov/gene>)

Tabela 2. Genes / variantes genéticas ligadas a doenças / perturbações oculares humanas comuns.

No	Disease	Gene/variant	Age of disease or disorder onset (a)
1	AMD	NOS2A, CFH, CF, C2, C3, CFB, HTRA1/LOC, MMP-9, TIMP-3, SLC16A8, etc.	Old
2	Cataract	GEMIN4, CYP51A1, RIC1, TAPT1, TAF1A, WDR87, APE1, MIP, Cx50/GJA3 & 8, CRYAA, CRYBB2, PRX, POLR3B, XRCC1, ZNF350, EPHA2, etc.	Old
3	Glaucoma	CALM2, MPP-7, Optineurin, LOX1, CYP1B1, CAV1/2, MYOC, PITX2, FOXC1, PAX6, CYP1B1, LTBP2, etc.	Over 40 except congenital form that can affect an infant
4	Inherited optic neuropathies	Complex I or ND genes, OPA1, RPE65, etc.	Young males
5	Marfan syndrome	FBN1, TGFB2, MTHFR, MTR, MTRR, etc.	Born with disorder but may not be diagnosed until later in life
6	Myopia	HGF, C-MET, UMODL1, MMP-1/2, PAX6, CBS, MTHFR, IGF-1, UHRF1BP1L, PTPRR, PPFIA2, P4HA2, etc.	Typically progresses until about age 20
7	Polypoidal choroidal vasculopathies	C2, C3, CFH, SERPING1, PEDF, ARMS2-HTRA1, FGD6, ABCG1, LOC387715, CETP, etc.	Between ages 50 and 65
8	Retinitis pigmentosa	RPGR, PRPF3, HK1, AGBL5, etc.	Between 10 and 30
9	Stargardt's disease	ABC1, ABCA4, CRB1, etc.	Signs may appear in early childhood to middle age
10	Uveal melanoma	PTEN, BAP1, GNAQ, GNA11, DDEF1, SF3B1, EIF1AX, CDKN2A, p14ARF, HERC2/OCA2, etc.	50 to 80

Fonte: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5767668/table/ijo-11-01-117-t01/?report=objectonly>. (SINGH e TYAGI, 2018.)

Tabela 3. Heranças Mendelianas Humanas responsáveis pelo surgimento da miopia.

ACOMETIMENTOS HUMANOS	
1. SINUS NODE DISEASE AND MYOPIA	35. ARTHROGRYPOSIS, DISTAL, TYPE 9; DA9
2. MYOPIA 18, AUTOSOMAL RECESSIVE; MYP18	36. MICROCEPHALY WITH OR WITHOUT CHORIORETINOPATHY, LYMPHEDEMA, OR MENTAL RETARDATION; MCLMR
3. MYOPIA 1, X-LINKED; MYP1	37. EHLERS-DANLOS SYNDROME, CLASSIC TYPE, 1; EDSCL1
4. ACHROMATOPSIA 3; ACHM3	38. ALAGILLE SYNDROME 1; ALGS1
5. NIGHT BLINDNESS, CONGENITAL STATIONARY, TYPE 1A; CSNB1A	39. FERRITIN LIGHT CHAIN; FTL
6. DEAFNESS AND MYOPIA; DFNMYP	40. COLLAGEN, TYPE XI, ALPHA-1; COL11A1
7. OPHTHALMOPLÉGIA, FAMILIAL STATIC	41. SPASTIC ATAXIA 7, AUTOSOMAL DOMINANT; SPAX7
8. STICKLER SYNDROME, TYPE I; STL1	42. ALBINISM, OCULAR, TYPE I; OA1
9. HYPERFERRITINEMIA WITH OR WITHOUT CATARACT; HRFTC	43. TRANSFORMING GROWTH FACTOR, BETA-1; TGFB1
10. BEHR SYNDROME; BEHR	44. ACHROMATOPSIA 2; ACHM2
11. COLLAGEN, TYPE II, ALPHA-1; COL2A1	45. KERATOCONUS 1; KTCN1
12. ALPORT SYNDROME 1, X-LINKED; ATS1	46. FORKHEAD BOX C2; FOXC2
13. MARSHALL SYNDROME; MRSHS	47. RETINITIS PIGMENTOSA; RP
14. GLAUCOMA, PRIMARY OPEN ANGLE; POAG	48. NEPHROTIC SYNDROME, TYPE 5, WITH OR WITHOUT OCULAR ABNORMALITIES; NPHS5
15. WAGNER VITREORETINOPATHY; WGVPR	49. CATARACT 21, MULTIPLE TYPES; CTRCT21
16. ECTOPIA LENTIS ET PUPILLAE	50. OTOSPONDYLOMEGAEPIPHYSEAL DYSPLASIA, AUTOSOMAL DOMINANT; OSMEDA
17. RETINITIS PIGMENTOSA 2; RP2	51. CORNEAL DYSTROPHY, AVELLINO TYPE; CDA
18. BRAIN SMALL VESSEL DISEASE WITH OR WITHOUT OCULAR ANOMALIES; BSVD	52. COLLAGEN, TYPE IX, ALPHA-3; COL9A3
19. MICROCORIA, CONGENITAL	53. ANTERIOR SEGMENT DYSGENESIS 1; ASGD1
20. STICKLER SYNDROME, TYPE II; STL2	54. COLLAGEN, TYPE XI, ALPHA-2; COL11A2
21. CUTIS LAXA, AUTOSOMAL RECESSIVE, TYPE IIA; ARCL2A	55. PAPILLORENAL SYNDROME; PAPRS
22. CATARACT 6, MULTIPLE TYPES; CTRCT6	56. POTOCKI-LUPSKI SYNDROME; PTL
23. OCULODENTODIGITAL DYSPLASIA, AUTOSOMAL RECESSIVE	57. CALCIUM CHANNEL, VOLTAGE-DEPENDENT, ALPHA-1F SUBUNIT; CACNA1F
24. ZIMMERMANN-LABAND SYNDROME 1; ZLS1	58. SMITH-MAGENIS SYNDROME; SMS
25. HYPOMAGNESEMIA 3, RENAL; HOMG3	59. V-KI-RAS2 KIRSTEN RAT SARCOMA VIRAL ONCOGENE HOMOLOG; KRAS
26. GLAUCOMA 1, OPEN ANGLE, A; GLC1A	60. PAIRED BOX GENE 6; PAX6
27. HOMOCYSTINURIA DUE TO CYSTATHIONINE BETA-SYNTHASE DEFICIENCY	61. GAP JUNCTION PROTEIN, ALPHA-8; GJA8
28. ALPORT SYNDROME 2, AUTOSOMAL RECESSIVE; ATS2	62. CATARACT 1, MULTIPLE TYPES; CTRCT1
29. HYPEREKPLEXIA 2; HKPX2	63. SPASTIC PARAPLEGIA 9A, AUTOSOMAL DOMINANT; SPG9A
	64. EDICT SYNDROME; EDICT

30. COWDEN SYNDROME 1; CWS1	65. SCALP-EAR-NIPPLE SYNDROME; SENS
31. SHORT STATURE, OPTIC NERVE ATROPHY, AND PELGER-HUET ANOMALY; SOPH	66. POTOCKI-SHAFFER SYNDROME
32. PSEUDOXANTHOMA ELASTICUM, FORME FRUSTE	67. MEGALOCORNEA; MGC1
33. MENTAL RETARDATION, X-LINKED, SYNDROMIC, CLAES-JENSEN TYPE; MRXSCJ	68. LAMININ, BETA-2; LAMB2
34. EHLERS-DANLOS SYNDROME, DERMATOSPARAXIS TYPE; EDSDERMS h	69. ANTERIOR SEGMENT DYSGENESIS 3; ASGD3
	70. NEUROFIBROMIN 1; NF1

Fonte: <https://www.ncbi.nlm.nih.gov/omim>

O RasGRF1 foi um dos primeiros genes relatados nos primeiros estudos de associação genômica ampla para miopia, sendo altamente expresso em neurônios e retina, ele é crucial para as funções oculares e consolidação da memória visual. Um mecanismo proposto indicou que o gene RasGFR1 tem influência significativa na miopia. Uma sugestão é que esse gene codifica o fator-1 liberador de nucleotídeo guanina específico da proteína Ras, que é altamente expresso na retina e nos neurônios, e então ativa Ras (CHEN, 2015).

Os resultados de estudos associados a genômica ampla (GWAS) segundo (CHEN, 2015), para erros de refração em humanos, apontaram que diversos SNPs podem estar relacionados de modo direto à miopia, cerca de 66,4%. A composição do gene RasGRF1, também conhecido como (GNRP, GRF1, CDC25, GRF55, CDC25L, H-GRF55, PP13187 e ras-GRF1), foi identificada estruturalmente com 30 éxons situados na posição 15q25.1 centimorgans. A sua presença no cromossomo 15 aumenta a lista de cromossomos envolvidos com genes influenciadores da miopia enumerada anteriormente. Ademais, foi visto que sua atuação é maior no cérebro onde promove a troca de GDP/GTP nas GTPases da família Ras, relacionando-se com a transmissão sináptica das respostas dos fotorreceptores.

Além disso, foi visto que o RASGRF1 também é expresso em células β pancreáticas em que regula a massa de células β e em regiões cerebrais específicas, envolvendo o hipocampo e o hipotálamo, dessa forma, ativando RAS a homeostase da glicose e função neuronal. Ainda foi observado que, quando o RASGRF1 possui uma deficiência causa hipo-insulinemia em consequência da redução das células β (BORRÁS, 2011).

De acordo com (FERNÁNDEZ-MEDARDE, 2009), estudos mostraram que pesquisas realizadas tanto *in vitro* quanto *in vivo* em animais como, camundongos deficientes em RasGRF1 demonstraram déficits de aprendizado e memória, isto ocorre

quando este gene é nocauteado ou desligado. Estes estudos evidenciaram pela primeira vez a relevância funcional da sinalização Ras na fotorrecepção de mamíferos e garantiram uma análise mais aprofundada de camundongos nocauteados no gene RasGRF1 como modelos potenciais para analisar mecanismos moleculares subjacentes a doenças humanas influenciadas por fotorrecepção defeituosa.

Adicionalmente aos modelos animais estudados, estudos de intervenção em humanos têm mostrado que os inibidores muscarínicos previnem o desenvolvimento de miopia, e a diminuição do ácido retinóico coroidal, isto foi detectado em modelos animais de miopia. Sendo assim, entende-se que o RasGRF1 seja um forte gene associado com alta miopia (CHEN, 2015).

O RasGRF1 é um gene com 3350 nucleotídeos altamente conservado entre espécies de primatas, roedores, baleias, golfinhos, ungulados e vários mamíferos placentários, isto possibilita seu uso eficaz como gene indicador em organismos modelo para estudos comparativos e por homologia em humanos, onde estas sequências ultrapassam os 90% de similaridade com taxas de E-Value em 0.0, o que indica uma alta similaridade entre os indivíduos comparados conforme Figuras 1 e 2.

Figura 1: BLAST (Basic Local Alignment Searching Tool) do gene RasGRF1, onde é possível observar os valores dos Scores das sequências desse gene em comparação com humanos.

Sequences producing significant alignments:

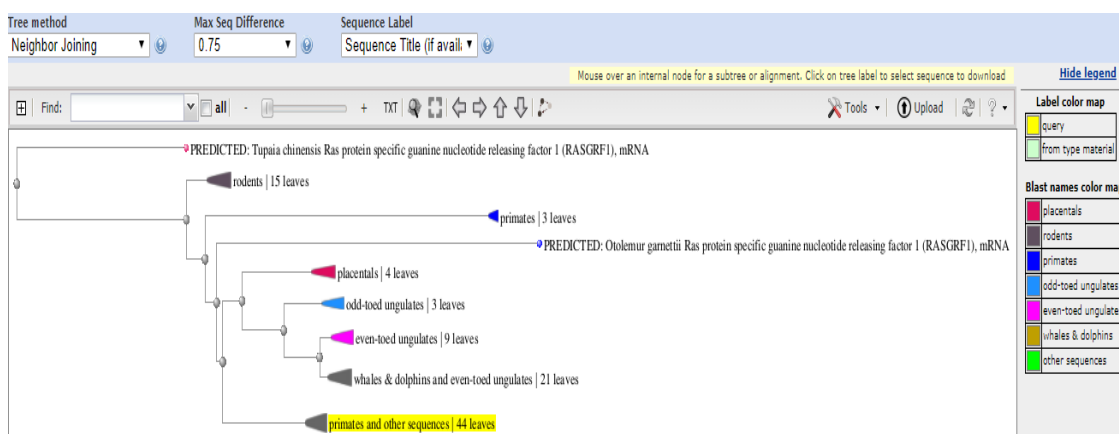
Select: [All](#) [None](#) [Selected: 0](#)

Alignments [Download](#) [GenBank](#) [Graphics](#) [Distance tree of results](#)

	Description	Max score	Total score	Query cover	E value	Ident	Accession
☐	Homo sapiens RASGRF1 gene for RASGRF1	6556	6556	100%	0.0	100.00%	LS482415.1
☐	Human ORFeome Gateway entry vector pENTR223-RASGRF1 complete sequence	6545	6545	99%	0.0	100.00%	LT741038.1
☐	Synthetic construct Homo sapiens clone ccsbBroadEn_11089 RASGRF1 gene, encodes complete protein	6545	6545	99%	0.0	100.00%	KJ901695.1
☐	Homo sapiens Ras protein-specific guanine nucleotide-releasing factor 1 (RASGRF1), complete cds	6545	6545	99%	0.0	100.00%	BC040275.1
☐	PREDICTED: Homo sapiens Ras protein specific guanine nucleotide releasing factor 1 (RASGRF1), transcript variant X5, mRNA	6300	6300	96%	0.0	99.97%	XM_017022458.2
☐	Homo sapiens Ras protein specific guanine nucleotide releasing factor 1 (RASGRF1), transcript variant 3, mRNA	6300	6300	96%	0.0	99.97%	NM_001145648.1
☐	Synthetic construct DNA clone, pF1KB9099, Homo sapiens RASGRF1 gene for Ras protein-specific, guanine nucleotide-releasing factor 1, complete cds, without stop	6300	6300	96%	0.0	99.97%	AB385192.1
☐	PREDICTED: Homo sapiens Ras protein specific guanine nucleotide releasing factor 1 (RASGRF1), transcript variant X7, mRNA	6257	6257	96%	0.0	99.71%	XM_017022459.2
☐	PREDICTED: Homo sapiens Ras protein specific guanine nucleotide releasing factor 1 (RASGRF1), transcript variant X4, mRNA	6257	6257	96%	0.0	99.71%	XM_017022457.2
☐	PREDICTED: Homo sapiens Ras protein specific guanine nucleotide releasing factor 1 (RASGRF1), transcript variant X3, mRNA	6257	6257	96%	0.0	99.71%	XM_017022456.2
☐	PREDICTED: Pan paniscus Ras protein specific guanine nucleotide releasing factor 1 (RASGRF1), transcript variant X4, mRNA	6255	6255	96%	0.0	99.74%	XM_003810718.3
☐	PREDICTED: Pan troglodytes Ras protein specific guanine nucleotide releasing factor 1 (RASGRF1), transcript variant X4, mRNA	6239	6239	96%	0.0	99.65%	XM_001153395.4
☐	PREDICTED: Pan paniscus Ras protein specific guanine nucleotide releasing factor 1 (RASGRF1), transcript variant X3, mRNA	6213	6213	96%	0.0	99.47%	XM_008963072.1
☐	PREDICTED: Pan troglodytes Ras protein specific, guanine nucleotide releasing factor 1 (RASGRF1), transcript variant X3, mRNA	6196	6196	96%	0.0	99.39%	XM_009429555.2
☐	PREDICTED: Gorilla gorilla gorilla Ras protein specific guanine nucleotide releasing factor 1 (RASGRF1), transcript variant X4, mRNA	6183	6183	96%	0.0	99.36%	XM_019011188.1
☐	PREDICTED: Gorilla gorilla gorilla Ras protein specific guanine nucleotide releasing factor 1 (RASGRF1), transcript variant X3, mRNA	6141	6141	96%	0.0	99.09%	XM_019011185.1
☐	PREDICTED: Pongo abelli Ras protein specific guanine nucleotide releasing factor 1 (RASGRF1), transcript variant X2, mRNA	6023	6023	96%	0.0	98.99%	XM_009429555.2

Fonte: NCBI-BLASTN

Figura 2: Utilizando o algoritmo Neighbor Joining foi possível construir uma árvore filogenética representando a similaridade entre as sequências gênicas das espécies para o gene RasGRF1

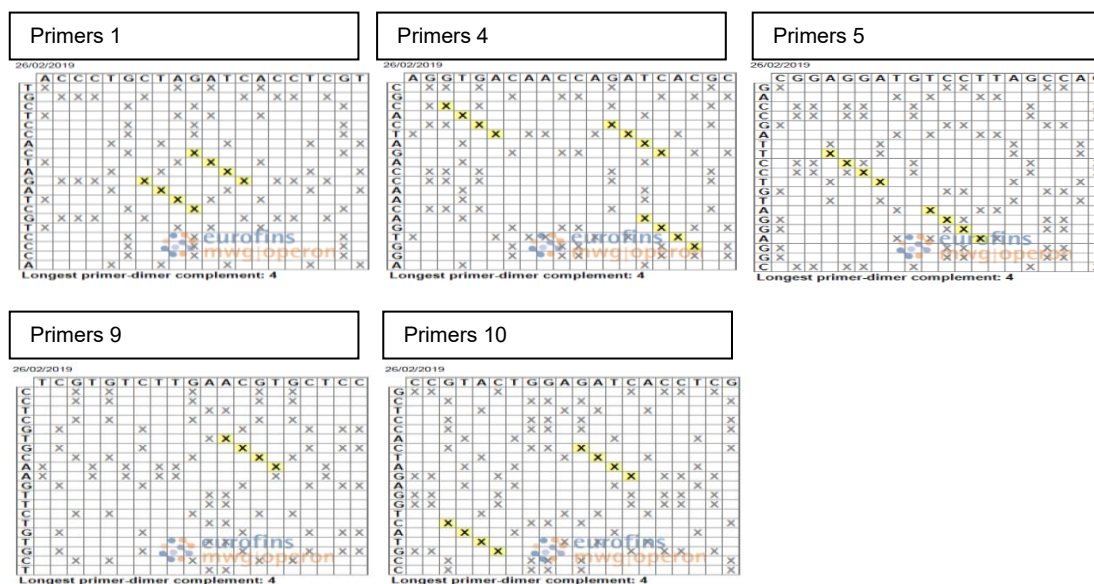


Fonte: NCBI-Phylogenetic Tree

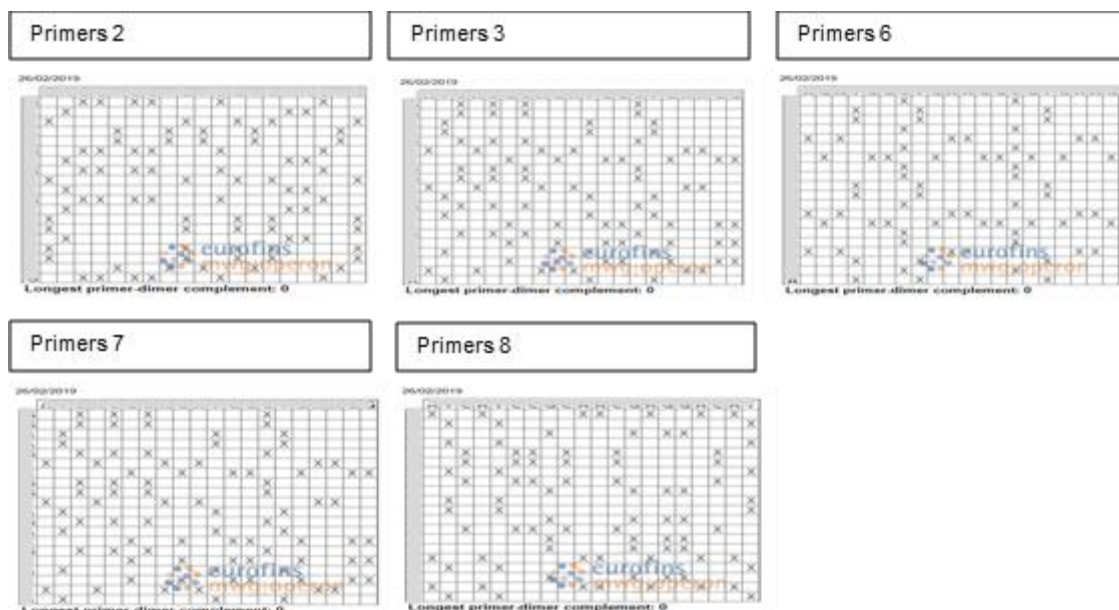
Tendo como base as Figuras 1 e 2, é possível a construção de primers universais que podem ser inicialmente utilizados em organismos modelo com maior facilidade, para depois extrapolar esses testes em humanos com o intuito de desvendar as diferenças genéticas existentes entre os genes de indivíduos acometidos ou não pela miopia, utilizando o gene RasGRF1 como modelo inicial. Os primers para o gene RasGRF1 foram construídos utilizando a ferramenta PickPrimers do NCBI com a ID de sequência LS482415.1, após esse passo, todos os primers foram confrontados com o programa Oligo Analysis Tool para verificação da validade das construções e sua eficácia *in silico* para possível posterior aplicação *in vivo*. Foram construídas 10 séries de primers com base nas sequências obtidas dos bancos de dados para o gene RasGRF1, desses primers, as séries 1, 4, 9, 10 e 5 não passaram nos testes *in silico* para sua construção, favorecendo o aparecimento de dímeros nas possíveis reações *in vitro*, conforme áreas amarelas sombreadas no Quadro 1.

Por outro lado, as séries de primers 2, 3, 6, 7 e 8 apresentaram estruturas satisfatórias na análise *in silico*. Principalmente, a série 2 por amplificar um fragmento de 905pb do gene. Isto indica que esses primers podem ser utilizados para os mais diversos fins em biologia molecular, amplificando fragmentos do gene RasGRF1 de humanos e de tantos outros organismos para análises de homologia e procura por mutações relacionadas ao acometimento pela miopia conforme Quadro 2.

Quadro 1: Diagramas de rejeição por formação de dímeros complementares das séries de primers analisadas.



Quadro 2: Diagramas de aceitação por ausência de dímeros complementares das séries de primers analisadas. Sequências sombreadas devido a possibilidade de deposição patentária.



Estudos de associação genômica ampla (GWAS) para erro de refração demonstraram que polimorfismos de nucleotídeo único (SNPs) em RasGRF1 e GJD2 estavam associados a erro de refração e miopia. O gene GJD2 em 15q14 codifica uma proteína específica do neurônio conexina 36, que é uma proteína específica do neurônio de uma família de proteínas de membrana integrais. O CX36 forma canais entre membranas adjacentes de células neuronais e está presente em fotorreceptores, células

bipolares e anácrinas que ao permitir o transporte intercelular de pequenas moléculas e íons, desempenha um papel essencial no processo de transmissão dos circuitos elétricos da retina. (KUNCEVICIENE, 2018). Isto é um forte indicativo de que o desenvolvimento cerebral das células neuronais está diretamente associado ao desenvolvimento das funções oculares. Isto pode ser devido a própria relação objetiva e necessariamente perfeita para seu funcionamento mútuo recebendo os estímulos do ambiente e convertendo em imagem para o cérebro. Ademais, pesquisas revelaram que esses genes são importantes para a transmissão e processamento de sinais visuais, assim como a miopia comum e o erro de refração. Nesse contexto, a expressão do RasGRF1 é regulada por receptores muscarínicos e o ácido retinóico onde os inibidores muscarínicos impedem o desenvolvimento da miopia (KUNCEVICIENE, 2018).

Conforme Kiefer, 2013 o gene ZIC2 controla dois fragmentos da progressão de células ganglionares retinianas ipsilaterais, ligadas a repulsão axonal na linha média do quiasma óptico, como também regula a composição das projeções axonais em seus alvos finais no cérebro. Assim, ZIC2 e RasGRF1 são genes de vulnerabilidade, tanto para miopia comum como também para alta miopia (OISHI, 2013). Além disso, o ZIC2 controla o refinamento de inputs específicos nos olhos com base nas informações visuais, regulando diretamente a expressão do transportador de serotonina (Sert) que está envolvido na modulação de mecanismos dependentes de atividade durante a construção dos circuitos sensoriais. Dessa forma, de acordo com (GARCÍA-FRIGOLA, 2010), a manifestação ectópica de ZIC2 induz a um fenótipo de falta de refinamento de seus axônios, um defeito, que é revertido após a aplicação de fluoxetina, mostrando, pelo menos em parte, que esse fenótipo é causado pela indução de Sert.

Outro exemplo enredado com o desenvolvimento da miopia é o gene APLP2 que codifica a proteína 2 semelhante ao precursor amiloide (APLP2), que é um membro da família APP (proteína precursora amiloide) integrando APP, APLP1 e APLP2. A coesão desta proteína e da APP é fundamental para mediar transmissão neuromuscular, aprendizagem espacial e plasticidade sináptica. A atuação do gene APLP2 está vigorosamente ligada à miopia em um padrão de primata (TKATCHENKO, 2015).

LEPREL1 é um gene que realiza um papel importante no desenvolvimento ocular e quando ocorre alguma mutação homozigótica de perda de função ele pode causar miopia severamente alta e início de catarata. (MORDECHAI, 2011). Além disso, outras modificações no gene, que codifica a prolil 3-hidroxilase-2 (P3H2), foram recentemente identificadas em indivíduos com miopia grave não síndrômica hereditária recessiva

(HUDSON, 2015). Genes multivariados são fortes candidatos com a alta miopia, como o LEPREL1 e GRM6 (WAN et al., 2018).

O gene CPSF1 é um estudo recente realizado na Lituânia e está associado a causa molecular da miopia. Embora as doenças oculares em humanos, incluindo a miopia, permaneçam desconhecidas neste gene, estudos em peixes-zebra foram realizados e os resultados sugeriram que CPSF1 estava relacionado no desenvolvimento ocular e na projeção axônio-cérebro de células ganglionares da retina. Nessa conjuntura, esses resultados sugerem que as mutações no CPSF1 podem ser uma nova causa e destacam um mecanismo potencialmente novo para a doença (OUYANG, 2019). Assim, este gene foi altamente expresso na retina, sugerindo um importante papel na função retiniana. O estudo em peixes-zebra de CPSF1 resultou em uma implicação na regulação do tamanho dos olhos. Dessa forma, a pesquisa afirma que as mutações do gene CPSF1 estão envolvidas na projeção do axônio e esses resultados podem fornecer novos caminhos para explorar mais a patogênese molecular da miopia, bem como as funções do CPSF1 no desenvolvimento do olho e da via visual (OUYANG, 2019).

Para os fatores genéticos relatados associados à miopia listados neste trabalho, alguns podem ser muito preliminares, enquanto outros podem já ter evidências confirmatórias ou contraditórias. Portanto, muitos dos genes ou loci relacionados à miopia precisam ser ainda mais validados. Além disso, para a maioria dos loci ou SNPs associados, os defeitos genéticos exatos e suas consequências funcionais relacionadas que contribuem para a miopia aguardam ainda a descoberta. Nenhum caminho específico subjacente à miopia foi esclarecido ainda, o que impede a priorização de genes candidatos. Apesar disso, o RasGRF1 pode ser uma possibilidade concreta de influência nesse acometimento. Portanto, em conjunto com estudos funcionais e uma combinação de diferentes tecnologias seria essencial para melhorar nossa compreensão dos fatores genéticos na miopia, levando às melhorias na previsão do início, prevenção e tratamento no futuro.

Levando em consideração as análises desenvolvidas e a condição multifatorial da expressão gênica que pode gerar a miopia fica claro que não existe apenas um gene envolvido no acometimento ou apenas uma via específica causadora deste problema. Isso configura que ainda estamos por descobrir vários outros fatores associados ao acometimento pela miopia. Pesquisas demonstraram que os genes subjacentes a todos os picos de miopia de alto grau, exceto possivelmente para o cromossomo Xq28, permanecem não identificados, sugerindo heterogeneidade genética substancial

(YOUNG, 2009). Apesar disso, esses dados são de 2009, desde então vários estudos mais recentes têm sido conduzidos nas mais diversas partes do mundo, indicando possíveis associações individuais para certos acometimentos oculares, dentre eles, a miopia. Também deve-se levar em consideração o papel dos fatores ambientais influenciando os genes e resultando no acometimento de forma mais efetiva. Estudos na busca por genes candidatos ressaltam que a miopia é muito complexa que é improvável que os estudos desses genes únicos demonstrem uma influência tão elevada para explicar a maioria dos casos de susceptibilidade. Existe uma necessidade de uma abordagem genômica ampla para explorar as contribuições e interações relativas entre genes candidatos conhecidos e, possivelmente novos genes que predisõem a maior susceptibilidade à miopia.

Conclusões

Este trabalho ofereceu uma visão bem abrangente e complexa dos estudos sobre a genética molecular da miopia, utilizando as ferramentas computacionais disponíveis para análises genéticas por meio da bioinformática. Foram identificados e sistematizados em tabelas vários genes potencialmente envolvidos nesse acometimento. Um estudo comparativo em organismos modelo possibilitou verificar a alta homologia do gene RasGRF1 entre várias espécies, principalmente de mamíferos. Foram construídas 5 séries de primers específicos para esse gene levando em consideração sua eficiência *in silico*, são elas as séries 2, 3, 6, 7 e 8 que servirão de base para a estruturação real dos possíveis testes *in vitro*.

Referências Bibliográficas

- BENAVENTE-PEREZ, Alexandra et al. Abstracts from the 15th International Myopia Conference. In: **Eye and Vision**. BioMed Central, 2016. p. 28.
- BSC, Afsaneh Khetrapal. **Genética da Visão**. Disponível em: <https://www.newsmedical.net/health/Genetics-of-Eyesight.aspx> > Acesso em 27 de fevereiro de 2019.
- BORRÁS, Consuelo et al. RasGRF1 deficiency delays aging in mice. **Aging (Albany NY)**, v. 3, n. 3, p. 262, 2011.
- CHEN, Ting et al. Polymorphism in the RasGRF1 gene with high myopia: A meta-analysis. **Molecular vision**, v. 21, p. 1272, 2015.
- CHOJNACKI, Szymon et al. Programmatic access to bioinformatics tools from EMBL-EBI update: 2017. **Nucleic acids research**, v. 45, n. W1, p. W550-W553, 2017.

FERNÁNDEZ-MEDARDE, Alberto et al. RasGRF1 disruption causes retinal photoreception defects and associated transcriptomic alteration. **Journal of neurochemistry**, v. 110, n. 2, p. 641-652, 2009.

FLITCROFT, D. Ian et al. Novel myopia genes and pathways identified from syndromic forms of myopia. **Investigative ophthalmology & visual science**, v. 59, n. 1, p. 338-348, 2018.

GARCÍA-FRIGOLA, Cristina; HERRERA, Eloísa. Zic2 regulates the expression of Sert to modulate eye-specific refinement at the visual targets. **The EMBO journal**, v. 29, n. 18, p. 3170-3183, 2010.

GOLDSCHMIDT, E.; JACOBSEN, N. Genetic and environmental effects on myopia development and progression. **Eye**, v. 28, n. 2, p. 126, 2014.

HALL, Thomas. A. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. **Nucleic Acids Symp. Ser.**, v.41, p.95-98, 1999.

HUDSON, David M. et al. Post-translationally abnormal collagens of prolyl 3-hydroxylase-2 null mice offer a pathobiological mechanism for the high myopia linked to human LEPREL1 mutations. **Journal of Biological Chemistry**, v. 290, n. 13, p. 8613-8622, 2015.

KIEFER, Amy K. et al. Genome-wide analysis points to roles for extracellular matrix remodeling, the visual cycle, and neuronal development in myopia. **PLoS genetics**, v. 9, n. 2, p. e1003299, 2013.

KUNCEVICIENE, Edita et al. Heritability of myopia and its relation with GDJ2 and RASGRF1 genes in Lithuania. **BMC ophthalmology**, v. 18, n. 1, p. 124, 2018.

LI, Jiali; ZHANG, Qingjiong. Insight into the molecular genetics of myopia. **Molecular vision**, v. 23, p. 1048, 2017

LIN, Fang-yu et al. Controversial opinion: evaluation of EGR1 and LAMA2 loci for high myopia in Chinese populations. **Journal of Zhejiang University-SCIENCE B**, v. 17, n. 3, p. 225-235, 2016.

M, Kanehisa, M. **Post-genome Informatics**, Oxford University Press (2000). Disponível em: <https://www.genome.jp/kegg/kegg1.html> > Acesso em 27 cde fevereiro de 2019

MORDECHAI, Shikma et al. High myopia caused by a mutation in LEPREL1, encoding prolyl 3-hydroxylase 2. **The American Journal of Human Genetics**, v. 89, n. 3, p. 438-445, 2011.

OISHI, Maho et al. Association between ZIC2, RASGRF1, and SHISA6 genes and high myopia in Japanese subjects. **Investigative ophthalmology & visual science**, v. 54, n. 12, p. 7492-7497, 2013.

OUYANG, Jiamin et al. CPSF1 mutations are associated with early-onset high myopia and involved in retinal ganglion cell axon projection. **Human molecular genetics**, 2019.

PEREIRA, Manuel Monteiro. **Miopia**. Disponível em: <https://www.saudebemestar.pt/pt/clinica/ofthalmologia/miopia/> > Acesso em 27 de fevereiro de 2019

SINGH, Mahavir; TYAGI, Suresh C. Genes and genetics in eye diseases: a genomic medicine approach for investigating hereditary and inflammatory ocular disorders. **International journal of ophthalmology**, v. 11, n. 1, p. 117, 2018.

TERRI L, Young. **The Molecular Genetics of Human Myopia: An Update**. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3718050/> > Acesso em 27 de fevereiro de 2019

TERRI L. Young. **A tool for finding specific primers**. Disponível em: https://www.ncbi.nlm.nih.gov/tools/primer-blast/primertool.cgi?ctg_time=1551214940&job_key=eXOm5vjX9X_SRWVAaCBBchI7UEA_KEtdPg > Acesso em 27 de fevereiro de 2019

THOMPSON, J. D. et al. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. **Nucleic Acids Res**, v.22, n.22, p.4673-4680, 1994.

TKATCHENKO, Andrei V. et al. APLP2 regulates refractive error and myopia development in mice and humans. **PLoS genetics**, v. 11, n. 8, p. e1005432, 2015.

WAN, Ling et al. Exome sequencing study of 20 patients with high myopia. **PeerJ**, v. 6, p. e5552, 2018. YOUNG, Terri L. The **molecular** genetics of human myopia: an update. **Optometry and vision science: official publication of the American Academy of Optometry**, v. 86, n. 1, p. E8, 2009.

ZHANG, Xiaoyu; QU, Xinhua; ZHOU, Xingtao. Association between parental myopia and the risk of myopia in a child. **Experimental and therapeutic medicine**, v. 9, n. 6, p. 2420-2428, 2015.

CAPÍTULO 3

PROSPECÇÃO SISTEMÁTICA E APLICADA DA HERANÇA GENÉTICA DA DEPRESSÃO UTILIZANDO FERRAMENTAS DE BIOINFORMÁTICA

SYSTEMATIC AND APPLIED PROSPECTING OF THE GENETIC INHERITANCE OF DEPRESSION USING BIOINFORMATICS TOOLS

DOI: <https://doi.org/10.56001/26.9786501906539.03>

Submetido em: 15/12/2025

Revisado em: 28/12/2025

Publicado em: 29/01/2026

Tainá Oliveira de Araújo

Universidade Federal de Campina Grande, Centro de Educação e Saúde, Cuité-PB

<http://lattes.cnpq.br/3946322725458190>

Carlíane Rebeca Coelho da Silva

Pós-doutorado em Biotecnologia, Universidade Federal Rural de Pernambuco,

Recife-PE

<https://orcid.org/0000-0002-9462-5823>

Igor Luiz Vieira de Lima Santos

Universidade Federal de Campina Grande, Centro de Educação e Saúde, Cuité-PB

<https://orcid.org/0000-0003-3438-0446>

Resumo

A depressão é a alteração patológica do humor mais comum em todo o mundo. Caracteriza-se como um transtorno afetivo de origem multifatorial que compromete a homeostasia da saúde mental do indivíduo, acarretando sintomas neurodegenerativos, cognitivos, psicotrópicos e no estado do humor geral. Este trabalho objetiva analisar a herança genética da depressão utilizando ferramentas de bioinformática. Foi realizada uma metodologia com características qualitativas e quantitativas, utilizando técnicas padronizadas, selecionando artigos nos bancos de dados mais representativos acerca do tema. Além disso, estudos in silico foram realizados para a identificação dos genes associados com a depressão e seus efeitos principalmente os genes MAOA e MAOB. Os resultados demonstraram que atualmente estão descritos 684 genes que podem estar relacionados com a progressão da doença catalogados como

influenciadores da depressão hereditária. Isto demonstra claramente a complexidade do componente genético associado aos transtornos de humor que, de modo amplo, pode afetar várias características psíquicas do organismo. Conclui-se com isso que a depressão está diretamente ligada aos transtornos de humor e que sofrem uma influência genética potencial capaz de alterar os comportamentos dos indivíduos possuidores de certos genes com funcionamento alterado.

Palavras-chave: Transtornos Mentais; Marcadores Moleculares; gene.

Abstract

Depression is the most common pathological mood disorder worldwide. It is characterized as an affective disorder of multifactorial origin that compromises the homeostasis of the individual's mental health, resulting in neurodegenerative, cognitive, psychotropic symptoms and in the general mood. This work aims to analyze the genetic inheritance of depression using bioinformatics tools. A methodology with qualitative and quantitative characteristics was carried out, using standardized techniques, selecting articles in the most representative databases on the topic. In addition, *in silico* studies were performed to identify the genes associated with depression and their effects, mainly the MAOA and MAOB genes. The results showed that 684 genes that may be related to the progression of the disease are currently described as influencing hereditary depression. This clearly demonstrates the complexity of the genetic component associated with mood disorders which, broadly, can affect various psychic characteristics of the organism. We conclude with this that depression is directly linked to mood disorders and that they suffer a potential genetic influence capable of altering the behaviors of individuals possessing certain genes with altered functioning.

Keywords: Mental Disorders, Molecular Genetics, Genes

Introdução

O Transtorno Depressivo Maior (TDM) compreende uma das formas mais comuns de patologia psiquiátrica. É a alteração afetiva mais estudada e comentada na atualidade, essa problemática mental é classificada como um transtorno de humor afetivo, regendo as atitudes dos sujeitos modificando a percepção de si mesmos, onde os indivíduos passam a enxergar seus problemas como grandes catástrofes. Por esta razão, uma pessoa depressiva apresenta principalmente um quadro de tristeza profunda e persistente, pensamentos de morte, desesperança, inutilidade, falta de energia e humor apático (ESTEVES *et al.*, 2006).

Dessa forma, segundo a classificação do Código Internacional de Doenças (CID-10) identifica uma série de variantes clínicas da depressão, que são categorizadas de acordo com a presença de sintomas psiquiátricos, a gravidade e a sua recorrência. F32.1 Episódios de depressão moderada (quatro ou mais sintomas exibidos), F32.2 Episódios de depressão grave sem sintomas psicóticos (geralmente apresenta diversos sintomas, com autoestima reduzida, pensamentos e tentativas de suicídio), F33.3 Transtorno depressivo recorrente, episódio atual grave com sintomas psicóticos (caracterizada pela presença de traços psicóticos, essa condição pode envolver alucinações e delírio) (SHADRINA *et al.*, 2018).

A depressão apresenta-se como uma das principais causas de incapacidade em todo o mundo, sendo responsável pela limitação do funcionamento físico, pessoal e social, em que contribui de forma significativa para a mortalidade prematura por suicídio (GONÇALVES *et al.*, 2018).

A incidência desta patologia está aumentando significativamente no mundo inteiro. Segundo a Organização Mundial da Saúde estima-se que 300 milhões de pessoas, de todas as idades sofram com esse transtorno.

Os transtornos depressivos constituem um grave problema de saúde pública, devido aos altos índices de morbimortalidade e sua alta prevalência, repercussões na saúde geral e impacto psicossocial, acarretando uma série de consequências negativas e afetando significativamente a qualidade de vida dos indivíduos (GONÇALVES *et al.*, 2018). Tal enfermidade predispõe o surgimento de outras patologias médicas como as doenças cardíacas, cerebrovasculares e arteriais periféricas (DASKALOPOULOU *et al.*, 2016). Assim, estima-se que os indivíduos deprimidos têm 45% (IC95%: 1,29- 1,63) maior risco de Acidente Vascular Encefálico (AVE) do que os não deprimidos e 25% (IC95%: 1,11-1,40) maior risco de mortalidade relacionado ao AVE (PAN *et al.*, 2011).

Tendo isso em vista, a depressão acarreta uma série de consequências desfavoráveis, pois este distúrbio vem acompanhado na maioria das vezes de uma série de sintomas e pode agravar doenças pré-existentes ou até mesmo provocar o aparecimento de novas a depender do grau que se estabelece. Por esta razão, tal transtorno do humor depressivo passará a ser a principal causa relacionada a sobrecarga de doenças em todo o mundo até o ano de 2030 (WHO, 2018).

A depressão é diferente das flutuações regulares de humor e das respostas emocionais aos problemas da vida cotidiana. Assim, a sua etiologia é bastante complexa e variada, podendo estar relacionada aos fatores ambientais como drogas, bebidas alcoólicas e estresse e os fatores genéticos e epigenéticos (HERNÁNDEZ *et al.*, 2016).

Os fatores de riscos genéticos desempenham um papel essencial no desenvolvimento do TDM, e mesmo que básico, é complexo e pouco compreendido. Assim sendo, estudos realizados com gêmeos apontam que cerca de 37-50% da origem dos transtornos do humor pode ser explicada pela genética; acredita-se que fatores ambientais expliquem o restante (SADOCK *et al.*, 2007; SULLIVAN *et al.*, 2000).

Na prática, atualmente, a engenharia genética está presente em todas as esferas da vida cotidiana, e não é diferente com a depressão, que possui caráter hereditário e associação genômica ampla (GWAS). Á vista disso, com o desenvolvimento da

tecnologia de microchips de DNA tornou-se possível a realização de estudos de GWAS a fim de observar um conjunto de variantes genéticas em todo o genoma de diferentes indivíduos para verificar se existe alguma variante que está associada a esta patologia, bem como os fatores de risco do início da depressão para explicar a sua patogênese. Todavia, para a realização deste tipo de estudo são necessários grandes conjuntos de amostras, de modo que, pacientes com diferentes formas desse distúrbio e milhares de pacientes em metanálises, não identificaram nenhum local específico responsável pela hereditariedade da depressão. Assim, essas falhas para identificar claramente as associações genéticas apontam que a depressão é um distúrbio psiquiátrico multifatorial heterogêneo complicado (SHADRINA *et al.*, 2018).

Por fim, este trabalho pauta-se no princípio de aprofundar e organizar os conhecimentos sobre a depressão, pois é uma patologia de etiologia bastante complexa e variada com muitos genes envolvidos e com alta incidência no Brasil e no mundo. O objetivo geral do trabalho foi analisar sistematicamente os genes envolvidos nas vias de progressão da depressão utilizando ferramentas de bioinformática. A pesquisa é justificada pela relevância social sobre o tema proposto e pela possibilidade da recuperação, promoção e proteção da saúde dos indivíduos e da população, de modo, a possibilitar a melhoria na qualidade de vida pela aquisição e compilação de informações importantes sobre esta problemática. O estudo é de grande relevância para a ciência por possibilitar a compilação de dados a respeito das influências hereditárias sobre a depressão. No entanto, a descoberta de genes da depressão pode representar um grande desafio uma vez que é generalista e disperso o conhecimento sobre os genes envolvidos e sua forma de atuar no desenvolvimento da doença. Desse modo, é possível ter uma visão mais ampla a respeito do tema, possibilitando um maior conhecimento e a possibilidade da melhoria na qualidade de vida dos indivíduos, no diagnóstico e um tratamento mais efetivo, reduzindo a progressão e o grau da doença.

Metodologia

- **Local e Equipe de Estudo**

O trabalho foi realizado nas dependências do Centro de Educação e Saúde da Universidade Federal de Campina Grande com o grupo de pesquisa BASE (Biotecnologia Aplicada à Saúde e Educação). O grupo de pesquisa e a Universidade ofereceram toda a infraestrutura computacional e de rede necessária para a realização do trabalho.

- **Metodologia Geral**

Trata-se de um estudo tecnológico e exploratório, bem como de revisão bibliográfica como ferramenta para a compreensão dos fatores genéticos associados à depressão, além de possuir características qualitativas e quantitativas como ferramenta subsidiária de informações essenciais para atingir os objetivos propostos. Foram utilizadas técnicas padronizadas de coleta de dados utilizando como base os bancos de dados de informações genéticas, com o intuito de expandir expectativas referentes ao tema, proporcionando uma visão conceitual mais ampla.

Nos bancos foram identificados artigos tratando a respeito do tema, bem como os conteúdos sobre aconselhamento genético sempre visando o tema proposto. Na base de dados mais sólida consultada até o momento foram catalogados 553 artigos versando sobre o referido tema utilizando o termo “genetics AND depression hereditary” destes foram selecionados os artigos mais recentes e representativos da influência genética na depressão.

- **Programas Genéticos e Análises dos Genes Envolvidos com a Depressão**

Os genes envolvidos com a depressão foram selecionados utilizando ferramenta de busca implícita ao site do NCBI onde os genes foram identificados para depressão nos Bancos de Dados GenBank (NCBI – National Center for Biotechnology Information) (<https://www.ncbi.nlm.nih.gov/>) e nos bancos de dados das vias bioquímicas (KEGG-Kyoto Encyclopedia of Genes and Genomes) (<http://www.genome.jp/kegg/>) e BRENDA (The Comprehensive Enzyme Information System) (<https://www.brenda-enzymes.org/index.php>). Análises de relações funcionais e interações enzimáticas foram realizadas no STRING (https://string-db.org/cgi/input.pl?sessionId=j1pdzCzBqhTE&input_page_show_search=on).

Após buscados, encontrado e catalogados os genes foram analisados um a um a respeito da sua influência no desenvolvimento da depressão. A partir daí foram selecionados os genes MAOA e MAOB para compor prioritariamente o escopo deste trabalho. Adicionalmente comenta-se brevemente sobre o gene TPH2 também influenciador dos gatilhos que promovem a depressão nos indivíduos.

Bases de dados foram compostas para a futura construção dos *primers* específicos para os genes descritos neste trabalho em pesquisas posteriores.

- **Viabilidade**

Este trabalho está em acordo com a Lei Nº 13.123, de 20 de maio de 2015 que define “patrimônio genético” como “informação de origem genética de espécies vegetais, animais, microbianas ou espécies de outra natureza, incluindo substâncias oriundas do metabolismo destes seres vivos” e “conhecimento tradicional associado” como “informação ou prática de população indígena, comunidade tradicional ou agricultor tradicional sobre as propriedades ou usos diretos ou indiretos associada ao patrimônio genético”. Este trabalho está cadastrado e segue as normas do CGEN - Conselho de Gestão do Patrimônio Genético. Dito isto, a presente pesquisa cumpre todos os requisitos estabelecidos pela referida lei.

Resultados e Discussão

A genética pode estar relacionada com o surgimento de um quadro depressivo. Os genes são importantes no desenvolvimento da depressão, mas não se sabe exatamente como suas ações se manifestam no organismo. Estudos hereditários e genéticos mostram a importância relativa dos genes no desenvolvimento dessa patologia, pois a genética comportamental estabeleceu que este transtorno de humor é parcialmente de origem genética (SULLIVAN *et al.*, 2000).

Apesar disso, os resultados indicam que, com as consultas aos bancos de dados de sequências genômicas e de conhecimento associado ao componente genético da depressão, foram constatados um total de 684 genes descritos a seguir como influenciadores da depressão hereditária potencialmente relevantes para este acometimento. Assim, nota-se que existe uma multilocalidade bastante ampla na distribuição gênica potencialmente influenciadora da depressão. Dessa forma, a identificação dos genes que têm influência no risco de desenvolver o TDM é a primeira etapa para a compreensão da biologia da depressão.

Vários genes candidatos têm sido investigados em estudos de associação genômica ampla (GWAS) que podem contribuir para o desenvolvimento do TDM, a análise sistemática nos bancos de dados biológicos realizada por este trabalho identificou como resultado 684 genes associados a esse processo, são eles: AANAT, ABCA13, ABCB1, ABCB7, Abcc1, Abcc2, ABCC3, ABCC9, ABL1, ACAMPAMENTO, acrB, ADAR, ADCY8, Adcyap1, ADH1B, ADIPOQ, Adipoq, ADM, Adora1, ADORA1, ADORA2A, Adora2a, ADRA2A, ADRB1, ADRB2, AGER, Ager, AGT, AGTR1,

Agtr1a, Ahr, Ahr, AKR1C4, Akt1, AKT1, Akt1, ALDH2, ANKK1, ANXA1, ANXA2, APAF1, APC, APLICATIVO, Aplicativo, Aplicativo, Apln, APOA4, APOE, Aqp4, AQP4, AR, Arco, Arco, ARCO, ARHGEF3, Arntl, ARRB1, ÁS, ASMT, ATF2, Atf3, ATP1A3, ATP7B, ATXN3, ATXN7, AVP, Avp, AVPR1B, BATE-PAPO, Bax, Bax, Bcl2, Bcl2l1, Bcl2l1, Bdkrb1, BDNF, Bdnf, Bdnf, Bhlhe40, BRCA2, BSN, C19orf12, C5, CACNA1A, CACNA1C, Cacna1c, Cacna1g, Calca, Camk2a, CAMK2A, Camk2b, Cartpt, Casp3, Casp3, Cav1, Cbln1, Cbln1, CCL2, Ccl2, CCL3, CCL5, Ccn2, CCND1, CCR2, CCR5, Cd274, CD38, CDC42BPB, CDH1, CDH13, CDK5, Cdk5, Cdkn1b, Chrm1, CHRM2, Chrm2, CHRNA3, CHRNA4, CHRNA5, Chrna7, CHRNA4, Chrn4, CLSTN2, CLU, Cmk1r1, CNP, CNR1, CNR2, Cnr2, CNTNAP2, COL1A1, COMT, CPE, CPT1A, CR1, CREB1, Creb1, Creb1, CREB5, CRH, CRHBP, CRHR1, Crhr2, CRP, Cry2, CRY2, CSF2RB, CSTB, CTNNB1, Ctnnb1, CX3CL1, CXCL12, CXCL8, CXCR4, Cybb, CYP27A1, CYP2B6, CYP2C19, CYP2D6, DAO, DAOA, DCC, DCTN1, Ddit4, DGKH, DICER1, DIO1, Dio2, DISC1, Disco1, DKK1, DLG4, DMD, Dmd, DNASE1, DPP4, DRD1, DRD2, Drd2, Drd2, DRD4, DRD5, DSCAM, DTNBP1, EGF, Egr1, Egr1, Ehmt2, Eif2ak3, Eif4e, Emx1, ENO2, Ephb2, ERBB4, ESM1, ESR1, Esr1, ESR2, ETERNO, FAAH, FABP2, FADD, FAS, Fgfr1, FKBP5, Fkbp5, FMR1, Fmr1, FNDC5, Fos, FOSB, FOXL1, FOXP2, FTO, Fumin, Gabbr1, Gabbr1, Gabra2, GABRA6, GAD1, GAD2, Gad2, GALR1, GALR2, GALR3, Gap43, GAROTA, Garota, GATO, Gato, GBA, GC, Gch1, GCK, GCLC, GCLM, GDNF, GFAP, GH1, Ghrh, GHRL, GHSR, GJA1, GLI1, GLI2, GLS2, GLUL, GNB3, GPC3, GPX1, Gria1, Gria1, GRIA1, Gria2, Gria2, GRIA2, GRIA3, GRIA4, GRIK3, GRIN1, Grin2a, GRIN2A, GRIN2B, GRIN2C, Grin2c, GRK2, GRM1, GRM2, GRM3, GRM4, GRM5, GRM7, GRM8, Grpr, GSK3B, Gsk3b, Gsk3b, H19, HAT1, Hcn1, Hcrt, HCRTR2, Hdac5, Hdac6, HES6, Hgf, HIF1A, Hif1a, Hip1, HMCN1, Hmgb1, HMOX1, Hmox1, Hnf4a, Homens1, HOMER1, Hras, HSD11B1, HSD11B2, HSPA4, Hspa4, HTR1A, HTR1B, HTR2A, Htr2a, HTR2C, Htr2c, HTR3A, Htr3a, HTR3B, HTR3D, Htr4, Htt, HTT, ICAM1, Id3, ID3, IDO1, Ido1, IFNA1, IFNG, IGF1, Igf1, IGF1R, IGF2, IGF2R, Il1, IL10, Il10, Il17a, IL18, IL1A, IL1B, Il1b, IL1R2, IL1RN, IL2, IL20RB, IL33, IL4, IL5, IL6, Il6, IL6R, IL7, IRF1, ITGA2, ITGAL, ITGB1, JAK1, KCNB1, Kcnd2, KCNJ10, Kcnj11, KCNJ5, KCNK2, Kdm4a, Keap1, Kitlg, KMT2B, KRAS, L1CAM, L1cam, LCT, LEP, Lep, LEPR, Lif, LRRK2, Lrrk2, LTA, LZTS1, MAOA, MAOB, MAP1B, MAP2K1, Mapk1, Mapk14, Mapk3, Mapk8, MAPT, Mapt, MAX, Mbd1, MCPH1, Mecp2, Mef2c, MEN1, Mer, Mir124a-1, MIR132, Mir134,

MIR145, MIR146A, MIR155, Mir16, MIR182, MIR21, Mir21a, MIR26B, MIR301A, MIR4485, MIR499A, MIR708, MIRLET7B, MMP2, MMP9, Mt1, Mt2, MTHFD1, MTHFR, MTNR1A, MTNR1B, MTOR, Mtor, MTR, MUC4, Myh7, MYO18B, Nampt, NAPAS2, NCAM1, NEGR1, NF1, Nfe2l2, Nfe2l2, Nfkb1, NFKB1, Nfkb1, NGF, Ngf, NLGN1, Nlgn1, Nlgn2, Nlgn3, NLRP12, Nlrp3, NOS1, Nos1, NOS1AP, NOS2, Nos2, Nos2, NOTCH1, NOTCH3, NOVA ENTRADA, NOVA ENTRADA, NOVA ENTRADA, NPAS3, Npas4, Npas4, NPPB, Nppb, NPSR1, Npy, NPY, Nr1h2, NR3C1, Nr3c1, Nr3c1, NR3C2, NR4A2, NRG1, Nrg1, NRXN1, Nsmf, Ntf3, NTRK2, Ntrk2, Ntrk3, NUCB2, °, OPRK1, OPRM1, Oprm1, OR7D4, OTX2, Oxt, OXTR, P2RX2, P2RX7, P2rx7, P2ry1, PAK1, PAK3, Para, PAX6, PCLO, PDE11A, PDE4A, PDE4B, PDSS1, Pdyn, PEA15, PEBP1, Penk, Per1, PER2, Per2, PER3, PF4, Pgr, Pik3cg, PIK3R1, PIWIL1, PLA2G4A, Pla2g4a, PLAT, PLCB1, PLCB4, PLCG1, POLG, POMC, PON1, POSIÇÃO DE LARGADA 2, POU2F1, Ppara, PPARG, PPARGC1A, Ppargc1a, Ppp1r9a, P PP1R9A, PPP3CA, Ppp3r1, Ppp3r2, PRKACA, PRKAR1A, Prkca, PRKCB, PRKCE, PRKCG, PRKN, PRL, PRNP, Prnp, PSEN1, PSMD9, PTEN, Pten, Pten, Ptgs1, PTGS2, Ptgs2, PTK2, PTK2B, PTPRD, PTPRR, PYCARD, PYY, Rac1, RAI1, Rapgef3, RBFOX1, RELN, Reln, RELÓGIO, REN, RGS2, RGS4, RORA, Rps6kb1, RTN4, S100A10, S100a10, S100B, S100b, SAT1, SCD, SEMA3A, SERPINE1, Serpine1, SGCE, SGK1, Sgk1, SHANK2, Shank2, Shank3, SHANK3, SHBG, SIRT1, Sirt1, SIRT2, Sirt2, Slc17a7, Slc18a2, SLC24A1, SLC5A7, SLC6A15, SLC6A2, SLC6A3, Slc6a3, SLC6A4, Slc6a4, Slc7a11, SMCHD1, SMPD1, SNAI1, SNAI2, Snap25, SNCA, Snca, SNTB1, SOD1, SOD2, Sod2, Sorcs2, SORCS2, Sorcs3, Sorrir1, Sorrir1, SP4, Sparc, Srf, SRF, Srf, STAT1, Stim1, STX1A, STXBP1, Syn2, SYNGAP1, Syt1, TAC1, Tac1, TACR1, Tacr1, tat, TBP, Tdo2, TEF, TFAP2B, TGFB1, Tgfb1, Tgfb1, TGFBR3, Thra, Tiam1, TIMP2, Tlr2, Tlr4, TLR4, Tlr4, TNC, TNF, Tnf, Tnf, Tnfrsf12a, Tnfrsf1a, TNFRSF1B, Tnfsf12, TNNT3, TOMM40, TOR1A, TP53, Tp53, TPH1, TPH2, Tph2, TPO, TPPP, TREM1, TRPV1, Trpv1, Tshz1, TTR, TXN, UCN3, Ucp2, VCAN, VDR, VEF, VEGFA, Vegfa, Vegfa, Vip, VPS35, vpu, WFS1, Wnt3, WT1, WWC1, YAP1, ZEB2.

A identificação pontual dos genes envolvidos com a depressão pode permitir estudos futuros tendo como alvo 1 gene específico entre os 684 descritos. Alguns desses genes já estão sendo estudados devido a sua relação com receptores cerebrais que controlam o humor dos indivíduos como é o caso dos genes MAOA e MAOB, que foram selecionados por este trabalho para a construção dos respectivos primers devido a sua

influência já bem catalogada nos principais estudos sobre esta temática. Há que se observar que a gama de genes é bastante diversa e complexa na sua atuação sobre os transtornos depressivos, influenciando vias bioquímicas e receptores e sinalizadores cerebrais diversos capazes de promover alterações de humor.

A monoamina oxidase (MAO) é uma enzima da membrana mitocondrial externa envolvida no processo de degradação de diversas monoaminas liberadas por neurônios e células gliais, no qual catalisam os neurotransmissores de monoamina, incluindo dopamina, serotonina e epinefrina no sistema nervoso central entre outros (MARATHE *et al.*, 2018). Ela é uma Flavoproteína da membrana externa mitocondrial (FAD) que catalisa a desaminação oxidativa de neurotransmissores e aminas biogênicas. Atua sobre aminas primárias e também sobre algumas aminas secundárias e terciárias. Ela difere de amina oxidase primária, pois pode oxidar aminas secundárias e terciárias, mas não a metilamina. Esta enzima é inibida por compostos acetilênicos, como clorgilina, 1-deprenil e pargilina, mas, ao contrário das diaminas oxidases, ela não é inibida por semicarbazida.

Com base na figura 1 percebe-se a influência generalista que esta enzima possui sobre várias funções biológicas, principalmente aquelas ligadas a depressão e estados mentais de ansiedade, pânico entre outros. Além disso, é possível perceber os fármacos que podem ter influência no controle dos transtornos causados pela deficiência desta enzima no sistema biológico humano. Lembrando que o comportamento farmacológico das monoaminas difere sobremaneira das diaminas oxidases como citado anteriormente.

Esta enzima é identificada no BRENDA como interconectada aos seguintes termos como mostrado na figura 1.

Figura 1: Mapa de palavras associadas a enzima monoamina oxidase.



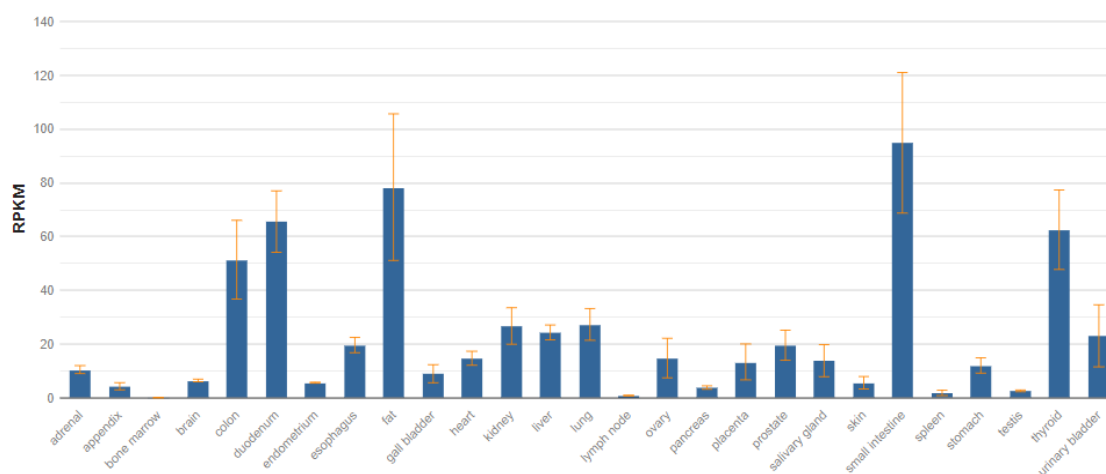
Fonte: <https://www.brenda-enzymes.org/enzyme.php?ecno=1.4.3.4#catecholamine>

As monoaminas são encontradas principalmente no fígado, rins e cérebro. Níveis aumentados ou diminuídos da MAO foram relacionados a uma série de distúrbios neuropsiquiátricos. Nos seres humanos, duas isoenzimas da MAO foram identificadas e codificadas por dois subtipos A e B e podem ser distinguidos de acordo com suas propriedades bioquímicas e farmacológicas: MAO-A que se caracteriza por ter preferência pelo substrato 5-hidroxitriptamina (serotonina) e normalmente é encontrada no sistema nervoso, fígado, trato gastrointestinal e na placenta; e MAO-B que metaboliza preferencialmente o substrato feniletilamina, e é normalmente encontrada no sistema nervoso e nas plaquetas (GOLDBERG *et al.*, 2014).

Existe uma série de informações sobre os achados clínicos envolvendo esses genes na plataforma OMIM (Online Mendelian Inheritance in Human). Esses genes também foram associados a uma variedade de outros transtornos psiquiátricos, incluindo comportamento antissocial e síndrome de Brunner. A prospecção genética desses genes demonstrou que alternativamente variantes transcritas processadas que codificam múltiplas isoformas foram observadas.

Análises de expressão de RNAs utilizando equipamentos sofisticados e técnicas de biologia molecular e genética revolucionárias vêm demonstrando que suas maiores expressões estão amplamente ligadas ao intestino delgado com 94,9 RPKM (Reads per Kilobase per Million) e ao tecido adiposo com 78,4 RPKM, e as menores expressões na medula óssea ≈ 0 RPKM e nos linfonodos 0,92 RPKM como mostra a Figura 2. Este gene foi mapeado no cromossomo sexual (X) com sua localização no braço curto (p) posição (11.3) Centimorgans, contendo 16 éxons (Xp11.3).

Figura 2: Principais locais de expressão das proteínas MAOA (Monoamino Oxidase A), expressa em quantidades de RNAs lidos RPKM/amostra.



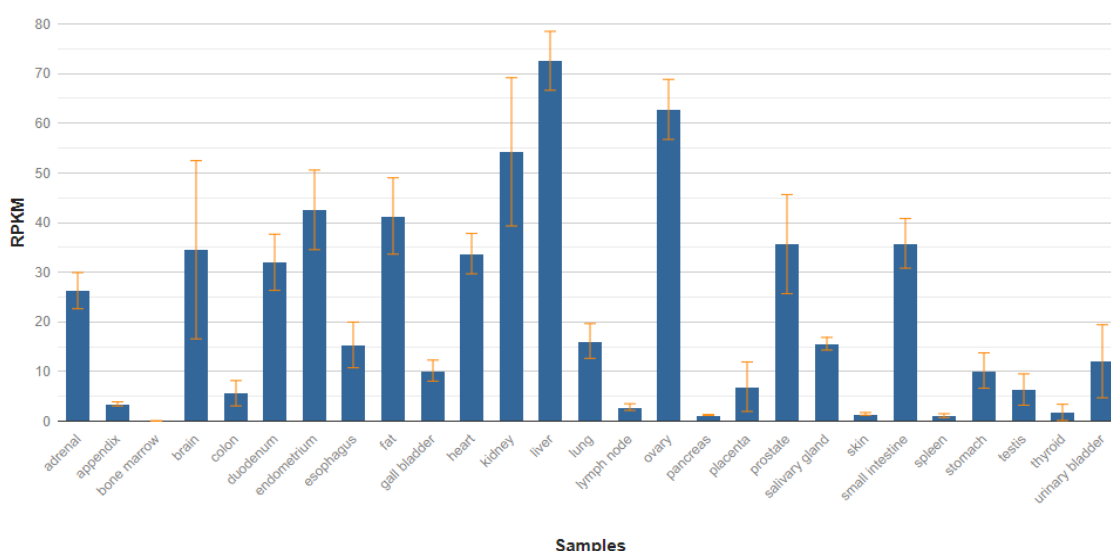
Fonte: <https://www.ncbi.nlm.nih.gov/gene/4128>. Samples

De acordo com (CASOS *et al.*, 2005), estudos mostraram que pesquisas realizadas em animais como camundongos transgênicos isolados onde a integração do transgene causou a deficiência da MAOA. Essa deficiência na MAOA, no cérebro dos filhotes, aumentou a concentração de serotonina até 9 vezes e a imunorreatividade do tipo serotonina estava presente nos neurônios catecolaminérgicos, causando tremores, dificuldade de correção motora e medo. Já nos cérebros de filhotes e adultos, as concentrações de norepinefrina foram aumentadas em até duas vezes e alterações citoarquiteturais foram observadas no córtex somatossensorial. Os adultos manifestaram uma síndrome comportamental distinta, incluindo agressão aumentada nos machos.

Além do gene MAOA foi observada e verificada a função do seu vizinho o gene MAOB. Os resultados mostram que esse gene codifica uma proteína que pertence à família da flavina monoamina oxidase.

Verificou-se que a MAOB degrada preferencialmente a benzilamina e a feniletilamina. Ela também catalisa a desaminação oxidativa de aminas biogênicas e xenobióticas e tem importantes funções no metabolismo de aminas neuroativas e vasoativas no sistema nervoso central e nos tecidos periféricos. Sua maior expressão está associada no fígado com 72,6 RPKM e no ovário com 62,8 RPKM e as menores expressões também na medula óssea ≈ 0 RPKM e no baço com 1,113 RPKM (Figura 3). Este gene foi mapeado no cromossomo sexual (X) com sua localização no braço curto (p) posição (11.3) Centimorgans (Xp11.3), contendo 18 éxons.

Figura 3: Principais locais de expressão das proteínas MAOB (Monoamino Oxidase B), expressa em quantidades de RNAs lidos RPKM/amostra.

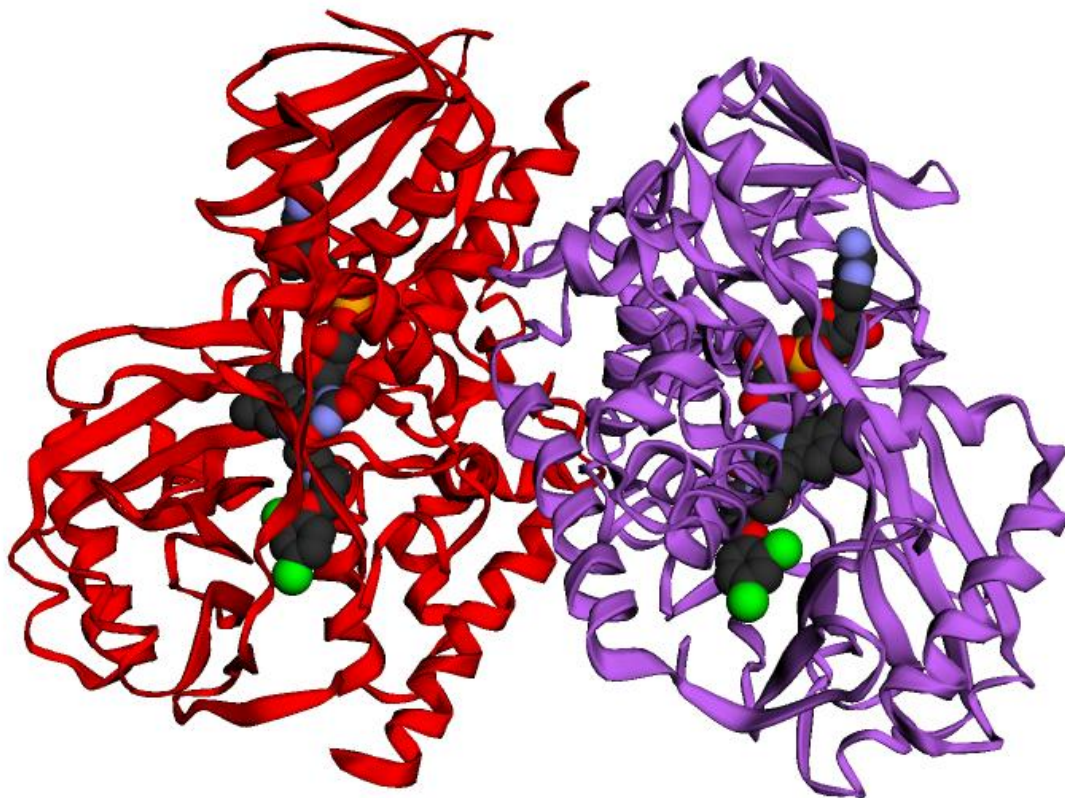


Fonte: <https://www.ncbi.nlm.nih.gov/gene/4129>

De acordo com (GRIMSBY *et al.*, 1997), estudos mostraram que pesquisas realizadas em animais como, camundongos demonstraram que a inativação direcionada de MAOB aumentou os níveis de beta-feniletilanina (PEA), mas não os de serotonina, noradrenalina ou dopamina, demonstrando um papel primário da MAOB no metabolismo da PEA. A PEA é responsável pela modulação do humor e do efeito. Assim, camundongos deficientes de MAOB mostraram uma reatividade aumentada ao estresse.

A estrutura tridimensional da monoamina oxidase A humana recombinante (hMAO A) como seu aduto inibido pela clorgilina está descrita na figura 4.

Figura 4: Estrutura tridimensional resolvida de uma proteína MAO hipotética complexada com a clorgilina.



Fonte:<https://www.ebi.ac.uk/thornton-srv/databases/cgi-bin/pdbsum/GetPage.pl?pdbcode=2bxx&template=3Dmoljsbox.html>

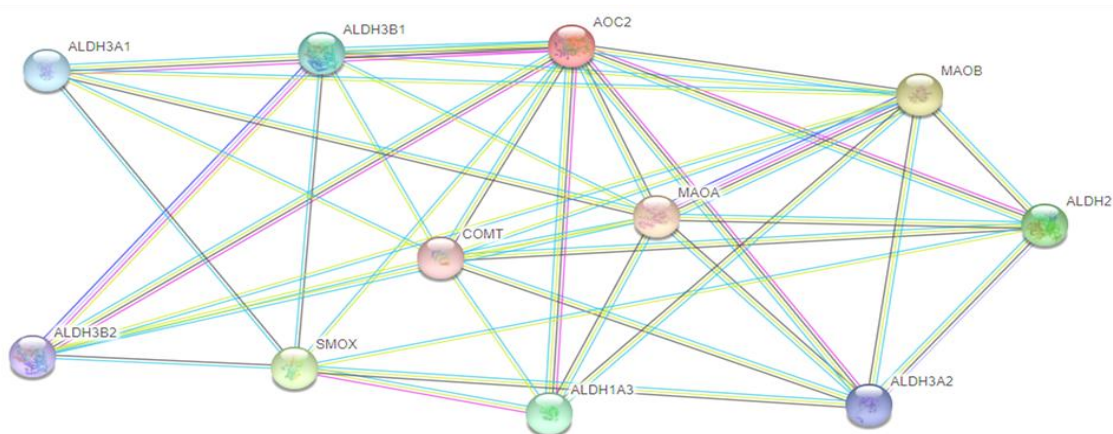
Embora a dobra da cadeia de MAO A humana (hMAO A) seja semelhante à de MAO A de rato e MAO B humano (hMAO B), a hMAO A é única por se cristalizar como um monômero e exibir o comportamento hidrodinâmico de solução de uma forma monomérica em vez de dimérica como a forma de hMAO B e MAO A de rato. O sítio ativo de hMAO A consiste em uma única cavidade hidrofóbica de aproximadamente 550 Å³, que é menor do que a determinada a partir da estrutura de hMAO B inibida por deprenil (aproximadamente 700 Å³), mas é maior do que a MAO A de rato (aproximadamente 450 Å³). Um componente importante da estrutura do sítio ativo de hMAO A é a conformação em alça dos resíduos 210-216, que difere daquela de hMAO B e MAO A de rato. A origem desta alteração estrutural é sugerida como resultado de interações de longo alcance na forma monomérica da enzima. Além de servir de base para o desenvolvimento de inibidores específicos de hMAO A, esses dados apoiam a proposta de que hMAO A envolve uma mudança da forma dimérica para a monomérica por meio de uma mutação Glu-151 -> Lys que é específica de hMAO A (ANDRÉS *et al.*, 2004).

Essas considerações colocam em controvérsia o uso de MAO A de fontes não humanas no desenvolvimento de drogas para uso em humanos. Visto que a farmacodinâmica dos compostos pode ser diferente entre os organismos modelos testados e os resultados esperados em humanos devido as conformações diferenciadas das proteínas e suas interações com os fármacos.

Polimorfismos nesses dois genes já foram identificados e investigados quanto a susceptibilidade para o transtorno de humor e no desenvolvimento de processos bioquímicos que influenciam os receptores neuronais e causadores de mal-estar biopsicossocial e resultando em comportamentos depressivos. Na depressão, ocorre a diminuição de neurotransmissores, como as aminas biogênicas, a serotonina, dopamina e noradrenalina, liberados na fenda sináptica que são responsáveis pela regulação do humor e respostas emocionais, mas a bomba de recaptção e a enzima (MAO) continuam trabalhando normalmente. Entretanto, um neurônio receptor captura menos neurotransmissores e o sistema nervoso funciona com menos neurotransmissores que normalmente seria preciso (ANDRADE *et al.*, 2003).

As principais relações bioquímicas conhecidas das monoaminas oxidases estão descritas na figura 5.

Figura 5. Genes diretamente relacionados com a MAO A.



Fonte: <https://string-db.org/cgi/network.pl?taskId=EjMm8SxyLduN>

Outro exemplo enredado com o desenvolvimento da depressão é o gene da enzima Triptofano Hidroxilase, que se encontra em dois genes diferentes: TPH1 e TPH2. TPH1 localiza-se no braço curto do cromossomo autossômico 11, na posição 15.1 (11p15.5), possui 13 exons é expresso principalmente em tecidos periféricos e a TPH2 localiza-se no braço longo do cromossomo 12, na posição 21.1 (12q21.1) encontra-se predominantemente nos neurônios (WALTHER *et al.*, 2003; BICALHO, 2007; ZHANG

et al., 2011). Assim, essa enzima realiza a modificação do triptofano até que seja convertido em serotonina (CHICHON *et al.*, 2008).

A enzima triptofano hidroxilase 2 é sintetizada nos corpos celulares dos neurônios serotoninérgicos, quando expressa e codificada catalisa o primeiro e limitante passo na biossíntese da serotonina, mutações nesse gene podem estar associadas a doenças psiquiátricas, como a depressão (BICALHO, 2007).

Polimorfismos no gene TPH2 foram associados com a redução dos níveis de serotonina na fenda sináptica do sistema nervoso central, sugerindo que certos polimorfismos possuem a capacidade de alterar a atividade da via serotoninérgica. Sendo assim, os fatores genéticos têm um papel essencial no surgimento do humor depressivo, que é causado por uma vulnerabilidade genética hereditária.

De acordo com Alenina *et al.*, (2009), estudos mostraram que pesquisas realizadas em animais como, camundongos machos nocauteados de TPH2 apresentaram níveis significativamente baixos de serotonina em todas as regiões cerebrais analisadas. Eles demonstraram um sono diurno mais prolongado, respiração reprimida, controle alterado da temperatura corporal e diminuição da frequência cardíaca e pressão arterial durante a noite. Já as fêmeas com ausência na enzima TPH2 apresentaram assistência materna prejudicada e aumento da agressão.

Tendo isso em vista, é notória a provável herança multifatorial e ambiental da depressão onde inúmeros genes podem ter influência no seu desenvolvimento fazendo com que seja difícil um tratamento pontual. Entretanto, existem estudos do tipo Genome-Wide Association Studies (GWAS) que podem contribuir para a elucidação de painéis genéticos influenciadores desta patologia e consequentemente contribuindo para um diagnóstico adequado e um tratamento farmacológico individualizado para cada caso. Com o mapeamento genético proveniente desta investigação torna-se possível a realização de procedimentos mais específicos, a fim de promover, prevenir, proteger, recuperar e reabilitar a saúde dos indivíduos. Além da utilização de fármacos mais direcionados a regular cada via enzimática comprometida e desencadeadora desse transtorno. Contudo, é possível perceber a necessidade de um trabalho monumental para identificação da função de cada gene e seu papel no TDM ou em tantos outros transtornos comportamentais que podem ter influência direta de cada um dos 684 genes identificados primariamente neste trabalho.

Conclusões

Com o exposto e através da identificação dos estudos conclui-se que o Distúrbio Depressivo Maior é uma doença psiquiátrica extenuante que afeta o indivíduo em sua plenitude, prejudicando a sua homeostasia e causando efeitos negativos a suas funções vitais. Observou-se no presente estudo que os fatores genéticos têm um papel importante no surgimento e desenvolvimento desta patologia, sendo causada por uma vulnerabilidade biológica hereditária.

Diante disso, este trabalho ofereceu uma visão bem abrangente e complexa dos estudos sobre a genética molecular da depressão, utilizando as ferramentas computacionais disponíveis para análises genéticas por meio da bioinformática.

Foram identificados e sistematizados 684 genes potencialmente envolvidos nesse acometimento.

Em um estudo futuro como desdobramento deste trabalho serão construídas séries de primers específicos para os genes aqui direcionados como o MAOA e MAOB levando em consideração sua eficiência *in silico* que servirão de base para a estruturação real dos possíveis testes *in vitro*.

Conclui-se que, apesar de possível, a utilização de organismos modelo e o desenvolvimento de tratamentos utilizando esses organismos deve ser bem elaborado para evitar o desenvolvimento de fármacos que não surtam efeito no objetivo pretendido devido as peculiaridades estruturais de cada molécula em seu respectivo organismo.

Referências

ALENINA, N; KIKIC, D; TODIRAS, M, et al. Retardo de crescimento e controle autonômico alterado em camundongos sem serotonina cerebral. **Proc. Nat. Acad. Sci.** 106: 10332-10337, 2009.

ANDRADE R; PEDRÃO L. Algumas Considerações sobre a Utilização de Modalidades Terapêuticas não Tradicionais pelo Enfermeiro na Assistência de Enfermagem Psiquiátrica. **Revista Latino Americana de Enfermagem**, v.13, n. 5, p.737 – 42 set/ out. 2005. Disponível em: <<http://www.scielo.br/pdf/rlae/v13n5/v13n5a19.pdf>>. Acesso em: 15 fev. 2020.

ANDRÉS, A.M.; SOLDEVILA, M., NAVARRO, A. *et al.* Positive selection in MAOA gene is human exclusive: determination of the putative amino acid change selected in the human lineage. **Hum Genet** 115, 377–386 (2004). <https://doi.org/10.1007/s00439-004-1179-6>

BICALHO MA. Estudo de polimorfismo do gene da triptofano hidroxilase 2 (TPH2) em idosos com depressão de início tardio. Dissertação (Mestrado em Farmacologia Bioquímica e Molecular). Universidade Federal de Minas Gerais, Minas gerais. 2007.

BRASIL. Ministério da Saúde. Secretaria de Atenção à Saúde. Departamento de Atenção Básica. Saúde mental / Ministério da Saúde, Secretaria de Atenção à Saúde, Departamento de Atenção Básica, Departamento de Ações Programáticas Estratégicas. – Brasília: Ministério da Saúde, 2013.

CASOS O; SEIF I; GRIMSBY J; GASPAR P, et al. Comportamento agressivo e quantidades alteradas de serotonina cerebral e noradrenalina em camundongos sem MAOA. **Science** 268: 1763-1766, 1995.

CHICHON WI; MATTHEISEN M; GEORGI A, et al. Brain-specific tryptophan hydroxylase 2 (TPH2): a functional Pro206Ser substitution and variation in the 5-region are associated with bipolar affective disorder. **Human Molecular genetics**, v. 17, p.87-97, 2008. doi: 10.1093/hmg/ddm286.

DASKALOPOULOU M. et al. Depressão como fator de risco para a apresentação inicial de doze doenças arteriais cardíacas, cerebrovasculares e periféricas: estudo de ligação de dados de 1,9 milhão de mulheres e homens.2016. **PLoS ONE** 11, e0153838.

ESTEVES FC; GALVAN AL. Depressão numa contextualização contemporânea. **Aletheia** [online]. 2006, n.24, pp. 127-135. ISSN 1413-0394.

GRIMSBY J; TOTH, M; CHEN, K, et al. Resposta ao estresse aumentada e beta-feniletilamina em camundongos deficientes em MAOB. **Nature Genet.** 17: 206-210, 1997.

GOLDBERG JS. et al. Revisiting the monoamine hypothesis of depression: a new perspective. **Perspectives in medicinal chemistry**, v.6, p.1–8, 2014. Acesso em: 18 fev. 2020.

GONCALVES AMC; et al. Prevalência de depressão e fatores associados em mulheres atendidas pela Estratégia de Saúde da Família. **J. bras. psiquiatr.** Rio de Janeiro, v. 67, n. 2, p. 101-109, June 2018. Disponível em:<<http://dx.doi.org/10.1590/0047-2085000000192>>. Acesso em: 17 jan.2020.

HERNÁNDEZ CH; CORONEL LP; AGUILAR CJ; RODRÍGUEZ CE. Neurobiology of major depression and its pharmacologic treatment. **Rev Salud Mental.** 2016. Disponível em:<<http://dx.doi.org/10.17711/SM.0185-3325.2015.067>>. Acesso em: 17 jan. 2020.

MARATHE, SV et al. Effects of Monoamines and Antidepressants on Astrocyte Physiology: Implications for Monoamine Hypothesis of Depression. **Journal of experimental neuroscience**, v.12, 2018, Acesso em: 18 fev. 2020.

PAN A; SUN, Q; OKEREKE OI; REXRODE KM; HU FB. Depressão e risco de morbimortalidade e acidente vascular cerebral: uma metanálise e revisão sistemática. 2011. **JAMA** 306, 1241–1249.

RESOLUÇÃO COFEN 468/2014. Atuação do enfermeiro no aconselhamento genético.

SADOCK BJ; KAPLAN HI; SADOCK VA. Sinopse da psiquiatria de Kaplan & Sadock: ciências comportamentais / psiquiatria clínica. 2007. Filadélfia: Lippincott Williams & Wilkins.

SHADRINA M; BONDARENKO EA; SLOMINSKY PA (2018). Fatores genéticos na doença de depressão maior. **Fronteiras em psiquiatria**, 9, 334. doi: 10.3389 / fpsyt.2018.00334. Acesso em: 15 jan. 2020.

SULLIVAN PF; NEALE MC; KENDLER. Epidemiologia genética da depressão maior: revisão e metanálise. 2000. **The American Journal of Psychiatry**, 157 , 1552-1562. : 10.1176 / appi.ajp.157.10.1552.

SZYMON C, ANDREW C, JOON LEE, ANNA F, RODRIGO L. Programmatic access to bioinformatics tools from EMBL-EBI update: 2017, *Nucleic Acids Research*, Volume 45, Issue W1, 3 July 2017, Pages W550–W553, <https://doi.org/10.1093/nar/gkx273>.

THOMPSON J. D. et al. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. **Nucleic Acids Res**, v.22, n.22, p.4673-4680, 1994.

WALTHER DJ; PETER JU; BASHAMMAKH S; HÖRTNAGL H, et al. Synthesis of Serotonin by a Second Tryptophan Hydroxylase Isoform. **Science**, v. 299, n. 5603, p.76, jan. 2003. DOI:10.1126/science.1078197

WORLD HEALTH ORGANIZATION. Folha informativa – Depressão. Geneva: WHO; 2018. Disponível em:<https://www.paho.org/bra/index.php?option=com_content&view=article&id=5635:folha-informativa-depressao&Itemid=1095>. Acesso em: 17 jan.2020.

ZHANG X; NICHOLLS PJ; LAJE G, et al. A functional alternative splicing mutation in human tryptophan hydroxylase-2. Mol. **Psychiatry**, v.16, p.1169-1176, 2011. doi:10.1038/mp.2010.99.

CAPÍTULO 4

DESENVOLVIMENTO DE UMA VACINA BASEADA EM LINHAGEM *Staphylococcus aureus* RF122 CONTRA MASTITE BOVINA

DEVELOPMENT OF A VACCINE BASED ON THE *Staphylococcus aureus* RF122 STRAIN AGAINST BOVINE MASTITIS

DOI: <https://doi.org/10.56001/26.9786501906539.04>

Submetido em: 24/12/2025

Revisado em: 28/01/2026

Publicado em: 21/02/2026

Tatiane Batista dos Santos

Universidade Tiradentes, Instituto de Tecnologia e Pesquisa, Aracaju-SE

<http://lattes.cnpq.br/7997717672688639>

Marianne Celestino Andrade

Universidade Tiradentes, Instituto de Tecnologia e Pesquisa, Aracaju-SE

<http://lattes.cnpq.br/9940366662825033>

Kelisson Alves Sousa

Universidade Tiradentes, Instituto de Tecnologia e Pesquisa, Aracaju-SE

<http://lattes.cnpq.br/3204626138865288>

Enzo Henrike Bitencourt de Moraes

Universidade Tiradentes, Instituto de Tecnologia e Pesquisa, Aracaju-SE

<http://lattes.cnpq.br/4184379006526319>

Pedro Henrique da Silva Rodrigues

Universidade Tiradentes, Instituto de Tecnologia e Pesquisa, Aracaju-SE

<https://lattes.cnpq.br/4918653543203001>

Victória Garcia Peres

Universidade Tiradentes, Instituto de Tecnologia e Pesquisa, Aracaju-SE

<http://lattes.cnpq.br/2209898161403692>

Lorena Emilia Sena Lopes

Universidade Tiradentes, Instituto de Tecnologia e Pesquisa, Aracaju-SE

<http://lattes.cnpq.br/5384001965243468>

Resumo

A Mastite Bovina (MB) refere-se à inflamação da glândula mamária ocasionada pela bactéria *Staphylococcus aureus* RF122. Até o momento, é controlado por meio de antimicrobianos e com ausência de uma profilaxia eficaz. Assim, o objetivo desse estudo foi desenhar uma vacina com o uso de ferramentas de bioinformática para *S. aureus* RF122. Para isso usou-se o software o *Vaxign* para obtenção de proteínas extracelulares de *S. aureus* RF122 logo após foi analisado a antigenicidade pelo *VaxiJen 2.0*, alergenicidade pelo *AlgPred*, predição de epítomos de células B pelo *Immune Epitope Data Base* (IEDB) com repetição dos parâmetros já descritos. Após seleção dos epítomos foi realizada o *docking* molecular com o receptor das células T, Toll-like-2 (TLR-2). Dessa maneira, um total de 85 proteínas foram preditas pelo *Vaxigen* e após caracterizações com o *VaxiJen 2.0*, alergenicidade pelo *AlgPred*, apenas uma seguiu para análises a SA27 (WP_001035596.1) que se trata de uma enzima de choque térmico. Subsequentes análises para a obtenção dos epítomos foi realizada e um total de oito foram identificados pelo IEDB, destes, somente três apresentaram-se antigênicos e não alergênicos e seguiram para etapa do *docking* molecular com servidor *HADDOCK 2.4*. Os epítomos apresentaram interação com o sítio ativo do TLR-2 apresentando interações de hidrogênio e hidrofóbicas indicando que os peptídeos previstos neste estudo são potencialmente capazes de ativar uma resposta imune inata. Contudo, através da bioinformática se pode buscar alvos contra MB, como a proteína SA27, tornando-se um potencial alvo no desenvolvimento vacinal.

Palavras-Chave: Bioinformática, mastite Bovina, vacina.

Abstract

Bovine Mastitis (BM) refers to inflammation of the mammary gland caused by the bacterium *Staphylococcus aureus* RF122. Currently, it is controlled through antimicrobials, but there is no effective prophylaxis. Therefore, the objective of this study was to design a vaccine using bioinformatics tools for *S. aureus* RF122. For this, the *Vaxign* software was used to obtain extracellular proteins of *S. aureus* RF122. Subsequently, antigenicity was analyzed using *VaxiJen 2.0*, allergenicity using *AlgPred*, and B-cell epitope prediction using the *Immune Epitope Data Base* (IEDB), with repetition of the parameters described above. After epitope selection, molecular docking was performed with the T-cell receptor, Toll-like receptor-2 (TLR-2). In this way, a total of 85 proteins were predicted by *Vaxigen*, and after characterization with *VaxiJen 2.0* and allergenicity by *AlgPred*, only one proceeded to analysis: SA27 (WP_001035596.1), which is a heat shock enzyme. Subsequent analyses to obtain epitopes were performed, and a total of eight were identified by IEDB. Of these, only three were antigenic and non-allergenic and proceeded to the molecular *docking* stage with the *HADDOCK 2.4* server. The epitopes showed interaction with the active site of TLR-2, exhibiting hydrogen and hydrophobic interactions, indicating that the peptides predicted in this study are potentially capable of activating an innate immune response. However, through bioinformatics, it is possible to search for targets against MB, such as the SA27 protein, making it a potential target in vaccine development.

Keywords: Bioinformatics, bovine mastitis, vaccine.

Introdução

Staphylococcus aureus RF122 é uma bactéria Gram-positiva com capacidade de secretar exotoxinas responsáveis por causar diversas doenças graves (Abril *et al.*, 2020; Ronco *et al.*, 2018). Essa linhagem afeta bovinos, resultando na redução da qualidade do

leite e ocasionando perdas econômicas significativas para fazendas e indústrias leiteiras. Além disso, pode levar à contaminação microbiana do leite e, em alguns casos, à transmissão para seres humanos (Ronco *et al.*, 2018; Zaatout, Ayachi e Kecha, 2020).

Estima-se que esse agente etiológico seja responsável por cerca de 50% dos casos de mastite. Nos Estados Unidos da América (EUA), a doença causa perdas econômicas de bilhões de dólares para a indústria de laticínios (Cobirka, Tancin e Slama, 2020; El-Sayed e Kamel, 2021). Atualmente, o controle é realizado por meio de medidas preventivas, como a adoção de procedimentos higiênicos rigorosos durante a ordenha, e do tratamento com antimicrobianos convencionais, destacando principalmente uso de antimicrobiano de origem natural. No entanto, esses tratamentos costumam apresentar limitações, como persistência da infecção, resistência bacteriana e alta recorrência (El-Sayed e Kamel, 2021; Ma *et al.*, 2022; Zaatout, Ayachi e Kecha, 2020). Além disso, o uso descontrolado de antibióticos no tratamento da mastite bovina contribui para o aumento da resistência antimicrobiana, um problema crescente que representa um desafio significativo para a saúde pública (Oliver e Murinda, 2012).

Dessa forma, a necessidade de uma buscar por métodos de profilaxia mais eficazes, como a criação de uma vacina contra o patógeno. Para essa patologia, algumas vacinas inativadas convencionais usando bactérias mortas ou vivas atenuadas estão disponíveis. Porém, hoje, é visto que a eficácia dessas abordagens tem sido insatisfatória, pois há um alto custo de número de patógenos que causam a mastite e sua heterogeneidade, altos custos de produção e baixa eficácia (Dzayee *et al.*, 2022; Ewida e Al-Hosary, 2020; Oliver e Murinda, 2012). Foi visto na literatura que vacinas baseadas em epítomos demonstrou ser promissora para várias patologias infecciosas, como infecção por *Neisseria meningitidis* e tuberculose (Bibi *et al.*, 2021; Rostamtabar *et al.*, 2019).

Dessa maneira, a bioinformática, um dos métodos de projeto de vacina multi epítopo contra doenças, fazer uma análise de dados computacionais e modelagem moleculares precisas para solucionar essa problemática, encontrando alvos de maneira mais rápida, fazendo com que haja uma troca de informações sobre os códigos biológicos e computacionais, gerenciando e integrando bases de dados aplicáveis, e construindo sistemas *in silico* para simulação de moléculas (Hashemzadeh *et al.*, 2020; Parvizpour *et al.*, 2020). Por meio da utilização desse método, é possível revelar candidatos e mecanismos relevantes na patogênese da mastite, nunca abordados na literatura científica, esclarecendo e auxiliando em formas de ações de vários princípios (Dzayee *et al.*, 2022; Parvizpour *et al.*, 2020).

Com base nessas observações, é possível perceber a grande necessidade em solucionar essa problemática ocasionada por esse determinado patógeno multirresistente. Assim, este estudo visa avaliar o uso de ferramentas de bioinformática para o desenho de vacinas para *Staphylococcus aureus* RF122, favorecendo uma melhor posição para que consiga prever uma afinidade de ligação entre moléculas, isto é, epítomos predominantes de proteínas imunogênicas em mastite em gado leiteiro e sua interação com receptor específico.

Metodologia

- **Obtenção dos antígenos de *Staphylococcus aureus***

Para a seleção de antígenos secretados de diversas linhagens de *Staphylococcus aureus*, foi utilizado o programa de filtragem *Vaxign – Vaccine Design* (<http://www.violinet.org/vaxign/>), que identifica proteínas com base em critérios pré-definidos. Este programa emprega um sistema de previsão e análise de alvos vacinais fundamentado no princípio da vacinologia reversa.

- **Análise antigênica, alergênica e não tóxica das proteínas**

Foi obtido as proteínas secretadas e em seguida foi usado o *VaxiJen v.2.0* (<http://www.ddgpharmfac.net/vaxijen/VaxiJen/VaxiJen.html>) para verificar o comportamento antigênico de sequências de proteínas usando 0.4 como limiar. Além disso, foi utilizado o *AlgPred* (<http://www.imtech.res.in/raghava/algpred/index.html>) para verificar o potencial alergênico dos alvos com intuito de investigar a propriedade não alergênica e alergênica usando o limiar de -0.4. Essas análises permitiram determinar o potencial antigênico e alergênico das proteínas selecionadas.

- **Obtenção dos epítomos de células B e análise imunológicas**

Para a predição de epítomos imunodominantes de células B, foi utilizado o banco de dados *Immune Epitope Database (IEDB)* (<http://www.iedb.org>). Após a identificação dos epítomos da proteína, foram selecionadas as sequências com comprimento entre 8 e 25 resíduos de aminoácidos utilizando o método *Kolaskar & Tongaonkar Antigenicity* (Ras-Carmona *et al.*, 2022). Em seguida, essas sequências foram analisadas quanto a características importantes associadas à capacidade vacinal, utilizando as ferramentas *VaxiJen* e *AllerTOP* (https://www.ddg-pharmfac.net/allertop_test/). Além

disso, a toxicidade dos epítomos foi prevista pelo servidor ToxinPred (<http://crdd.osdd.net/raghava/toxinpred/>).

- **Concepção da construção final da vacina**

Os epítomos de células B foram incluídos e unidos pelo flexível para a construção da vacina. O ligante KK foi usado para conjugar os epítomos. O ligante utilizado confirma a separação dos epítomos individuais. O ligante KK ou bi-lisina facilita a atividade imunogênica do construto de vacina.

- **Predição de antigenicidade, alergenicidade e análise das propriedades físico-químicas da vacina formulada**

A antigenicidade da vacina candidata final foi analisada pela ferramenta *VaxiJen* v2.0 (<http://www.ddg-pharmfac.net/vaxijen/VaxiJen/VaxiJen.html>), e o limiar de predição foi mantido em 0,4. O comportamento não alérgico foi previsto usando o servidor *AllerTOP* v.2.0. e a toxicidade pelo *ToxinPred*. Em seguida, o servidor *ProtParam* (<https://web.expasy.org/protparam/>) foi utilizado para as várias características físico-químicas da vacina construída, como peso molecular (PM), Ponto isoelétrico (pI), Índice de instabilidade (II), Índice Alifático (IA) e Grande média de hidropatia (*GRAVY*). Além disso, também foi utilizado o servidor *SOLpro* (<http://scratch.proteomics.ics.uci.edu/>) foi usado para prever a propensão da solubilidade da proteína após a superexpressão em *Escherichia coli*.

- **Previsão, refinamento e validação da estrutura terciária da vacina candidata**

O modelo tridimensional (3D) da vacina foi previsto usando o *AlphaFold 3* (<https://alphafoldserver.com/welcome>). O *AlphaFold 3* é um servidor web para previsão de alta precisão de estruturas terciárias contendo quase todos os tipos moleculares presentes no *Protein Data Bank* (PDB).

O refinamento da estrutura terciária é o processo de melhoria da qualidade que é uma etapa vital para o projeto de vacinas. A estrutura 3D recuperada da vacina foi refinada usando o servidor baseado na web *GalaxyRefine* (<http://galaxy.seoklab.org/refine>). Este servidor forneceu uma classificação robusta e geral, realizada por uma técnica de refinamento baseada em *CASP10*.

A estrutura refinada foi visualizada usando o *BIOVIA Discovery Studio* 2019. Além disso, a avaliação do gráfico *Ramachandran* da estrutura da proteína

purificada foi realizada usando o servidor *PROCHECK* (<https://saves.mbi.ucla.edu/>) para analisar sua qualidade geral.

- **Docking molecular - Vacina e TLR-2**

Para verificar a ancoragem entre o receptor Toll-like-2 (TLR-2) e o construto vacinal, utilizou-se o software HADDOCK 2.4 (<https://rascar.science.uu.nl/haddock2.4/submit/1>) verificando a menor pontuação. A estrutura da proteína TLR-2 (5D3I) foi obtida no *Protein Data Bank* (PDB). Antes do procedimento de docking, o ligante foi removido com o auxílio do software Discovery Studio. A visualização do docking molecular foi realizada no PyMOL.

A interação entre o receptor TLR-2 e o construto vacinal ocorreu nos aminoácidos leucina (posição 324) e asparagina (posição 379) (Koymans *et al.*, 2015). Essas interações foram analisadas com base em distâncias intermoleculares, considerando ligações de hidrogênio inferiores a 2,00 Å e interações hidrofóbicas menores que 5,00 Å, conforme identificado pelo Discovery Studio.

Resultados e Discussão

A resposta imune é direcionada a antígenos protetores no hospedeiro, desempenhando um papel crucial na proteção contra doenças infecciosas e não infecciosas (Mukherjee, Huda e Sinha Babu, 2019). Esses antígenos são amplamente reconhecidos como elementos fundamentais na produção de vacinas e são considerados marcadores biológicos relevantes para o diagnóstico de patologias, além de representarem componentes-chave na defesa do organismo contra infecções (Noor *et al.*, 2022).

Apesar dos avanços terapêuticos no combate a doenças infecciosas, os microrganismos patogênicos continuam a representar uma ameaça significativa à saúde pública. Embora as vacinas convencionais tenham sido eficazes contra determinados patógenos, infecções emergentes exigem abordagens inovadoras para o desenvolvimento de imunizantes (Clements e Casani, 2016).

Métodos tradicionais baseados no cultivo de microrganismos e na identificação de seus componentes imunogênicos são muitas vezes demorados e apresentam limitações, como a identificação exclusiva de antígenos produzidos *in vitro*. Além disso, a abundância de proteínas nem sempre se correlaciona com a imunogenicidade, e antígenos expressos *in vivo* durante a patogênese podem não ser detectados em condições laboratoriais. Essa limitação é particularmente desafiadora no caso de microrganismos

não cultiváveis (Gebre *et al.*, 2021). Nesse contexto, os avanços no sequenciamento genômico e o desenvolvimento de tecnologias biotecnológicas têm permitido o surgimento de métodos alternativos, como a vacinologia reversa, que oferece ferramentas para a identificação de antígenos promissores (Sette e Rappuoli, 2010).

A antigenicidade, que determina a capacidade de um antígeno de se ligar aos receptores das células B e T, desempenha um papel crucial na indução da resposta imune e na formação de células de memória (Vaillant, Sabir e Jan, 2024). Nesse estudo, a análise da natureza antigênica do construto vacinal foi realizada para assegurar sua interação eficaz com receptores imunológicos, garantindo sua potencialidade como candidato à vacina.

- **Obtenção dos antígenos de *Staphylococcus aureus* e propriedades da proteína**

As análises realizadas identificaram uma proteína relevante no banco de dados Vaxign – Vaccine Design, um sistema de análise especializado em armazenar antígenos protetores. Foi selecionada a proteína CP27 da linhagem *Staphylococcus aureus* RF122, conhecida como toxina-1 da síndrome do choque tóxico (TSST-1/CP27/YP_415862.1/WP_001035596.1), pertencente ao grupo de superantígenos distintos (SAGs). Este antígeno foi escolhido para análises subsequentes devido às suas características não alergênicas e altamente antigênicas.

Os SAGs são conhecidos por induzir a proliferação específica de células T V β e promover a liberação de citocinas pró-inflamatórias, como interleucina-1 (IL-1), IL-2, IL-6, fator de necrose tumoral alfa (TNF- α), IFN- γ , e quimiocinas CCL2 e CCL3. A liberação descontrolada desses mediadores pode levar a condições graves, como febre, erupções cutâneas, lesão de múltiplos órgãos, coma e até morte por choque severo (Noli Truant *et al.*, 2020).

Além disso, essa resposta inflamatória pode comprometer a eficácia do sistema imunológico, criando um viés para respostas Th1 ou Th17, o que dificulta o recrutamento adequado de células efetoras e prejudica a capacidade da resposta imune adaptativa de combater a infecção por *S. aureus*. Embora os efeitos exatos dos SAGs na mastite ainda não sejam completamente compreendidos, a ativação específica de células T V β em resposta a SAGs do *S. aureus* já foi bem documentada (Noli Truant *et al.*, 2022).

Dentro do contexto da vacinologia reversa, a identificação de proteínas de superfície, como a CP27, começa pelo genoma e não pelo cultivo de microrganismos.

Esse método utiliza técnicas computacionais e reconhecimento de padrões, permitindo identificar antígenos que não são acessíveis por métodos convencionais (Hajighahramani *et al.*, 2017; Parvizpour *et al.*, 2020). A CP27, por exemplo, demonstrou um papel significativo na imunogenicidade, destacando seu potencial como um componente essencial em novas vacinas. Além de acelerar o processo de construção vacinal, essa abordagem reduz o tempo necessário para o desenvolvimento, concentrando-se em análises *in silico* (Hashemzadeh *et al.*, 2020).

O uso dessa técnica não apenas aumenta a confiabilidade na detecção de antígenos, mas também contribui para o avanço da vacinologia reversa (Mukherjee, Huda e Sinha Babu, 2019). O desenvolvimento de ferramentas para a identificação de proteínas extracelulares reforça o potencial dessa metodologia como um exemplo de aplicação bem-sucedida da biotecnologia assistida por computação. Baseado nessa estratégia, a imunoinformática surge como uma área inovadora, integrando bioinformática, imunologia e vacinação. No presente estudo, a aplicação da imunoinformática demonstrou ser um recurso poderoso para validação e apoio à pesquisa, consolidando seu papel como ferramenta fundamental no avanço de projetos de vacinologia reversa (Dzayee *et al.*, 2022; Noor *et al.*, 2022).

A predição de epítomos imunodominantes possibilita o desenvolvimento de estratégias vacinais inovadoras, incluindo vacinas peptídicas, proteicas ou quiméricas. Entre elas, as vacinas quiméricas destacam-se como uma abordagem promissora, devido à sua capacidade de amplificar a resposta imune contra o alvo específico (Hajighahramani *et al.*, 2017; Koymans *et al.*, 2018).

No presente estudo, a utilização de softwares especializados permitiu a seleção de epítomos potenciais na proteína CP27, identificando três regiões promissoras para o desenho vacinal. Essas regiões demonstraram alta capacidade de ligação ao alvo, desencadeando uma cascata imunológica apropriada. Tal característica as torna ideais para a síntese de peptídeos a serem empregados na construção de uma vacina peptídica (Dzayee *et al.*, 2022).

Os resultados da análise realizada pelo servidor *Vaxign*, que abrange os 33 genomas de *Staphylococcus aureus*, estão disponíveis publicamente na página web do *Vaxign*, permitindo acesso gratuito para exploração. Entre as linhagens analisadas, a *S. aureus* RF122 destacou-se, apresentando alvos com peptídeos sinal indicando endereçamento extracelular. Dado o papel crucial dessas proteínas na patogênese,

selecionamos os alvos identificados nessa linhagem para o desenvolvimento de um desenho vacinal.

A partir da linhagem *S. aureus* RF122, foram obtidos 85 antígenos. Destes, realizamos uma análise de acordo com a localização extracelular, identificando a proteína antigênica mais relevante, bem como seu potencial de alergenicidade. Com base nesse levantamento, foi selecionado o antígeno mais promissor para a caracterização e seleção de proteínas apropriadas para o desenvolvimento vacinal.

Utilizando o servidor *VaxiJen v.2.0*, avaliamos a antigenicidade da sequência final prevista, enquanto a alergenicidade foi analisada através do *AlgPred*. Essas análises permitiram a seleção de uma proteína antigênica (com pontuação de 0.80) e não alergênica (com pontuação de -0.53834409). Os resultados da Tabela 1 demonstram a seleção de proteínas e suas isoformas para análise subsequente, com base nesses critérios.

Tabela 1: Proteínas encontradas que passaram na filtragem no *Vaxijen* e *AlgPred*.

ID	Proteína	Localização	Ant	Aler
1	WP_001125613.1	Extracelular	0.65	0.086822409
2	WP_000782469.1	Extracelular	0.60	0.66045685
3	WP_011382241.1	Extracelular	0,53	1.3339441
4	WP_000744080.1	Extracelular	0,76	0.69254625
5	WP_000953751.1	Extracelular	0.54	0.88316978
6	WP_000827752.1	Extracelular	0.51	0.60097516
7	WP_000473600.1	Extracelular	0.67	0.49793782
8	WP_000849075.1	Extracelular	0.51	1.5129988
9	WP_000239604.1	Extracelular	0.54	1.1308036
10	WP_000222258.1	Extracelular	0.55	0.25130738
11	WP_000857485.1	Extracelular	0.61	0.56181524
12	WP_001225031.	Extracelular	0.60	1.9380175
13	WP_001148141.1	Extracelular	0.44	0.57244783
14	WP_000791713.1	Extracelular	0.60	0.08817441
15	WP_000595399.1	Extracelular	0.58	0.31272599
16	WP_001088785.1	Extracelular	0.58	1.1816235
17	WP_000978189.1	Extracelular	0,65	0.9330163
18	WP_000594524.1	Extracelular	0.65	0.78139415
19	WP_000943805.1	Extracelular	0,67	0.13220252
20	WP_000916708.1	Extracelular	0.68	0.3875661
21	WP_011382305.1	Extracelular	0.53	0.30759886
22	WP_000710572.1	Extracelular	0.43	0.80411135
23	WP_000476437.1	Extracelular	0.63	0.68151788
24	WP_011382381.1	Extracelular	0.58	0.62675885
25	YP_416274.1	Extracelular	0.54	0.39807132
26	WP_000676575.1	Extracelular	0.40	1.4764424
27	WP_001035596.1	Extracelular	0.80	-0.53834409

28	WP_001039452.1	Extracelular	0.41	1.8266594
29	WP_000791410.1	Extracelular	0.61	0.19699659
30	WP_000669013.1	Extracelular	0.70	0.61062445
31	WP_011382230.1	Extracelular	0.63	-0.3072158
32	WP_000524059.1	Extracelular	0,35	0.53347796
33	WP_000769137.1	Extracelular	0.74	0.69353966
34	WP_000784031.1	Extracelular	0.79	-0.22609434
35	WP_001038860.1	Extracelular	0.49	2.4947067
36	WP_011382306.1	Extracelular	0.38	1.7320975
37	WP_000736801.1	Extracelular	0.84	0.96250669
38	WP_001225809.1	Extracelular	0.64	1.1508346
39	WP_001240826.1	Extracelular	0,53	0.75432697
40	WP_000749386.1	Extracelular	0.39	0.65840104
41	WP_000236411.1	Extracelular	0,46	0.68875073
42	WP_000725232.1	Extracelular	0.78	0.69715495
43	WP_000769839.1	Extracelular	0.59	0.76968486
44	WP_000751995.1	Extracelular	0.89	0.45909052
45	WP_000737689.1	Extracelular	0.77	0.27511375
46	WP_011382184.1	Extracelular	0.58	1.1394139
47	WP_001063564.1	Extracelular	0.51	0.55798191
48	WP_000509673.1	Extracelular	0.49	1.4591758
49	WP_000717390.1	Extracelular	0.69	0.57180381
50	WP_001032833.1	Extracelular	0.44	1.2988089
51	WP_000746599.1	Extracelular	0.78	0.81725677
52	WP_000713850.1	Extracelular	0.62	1.528861
53	WP_001044560.1	Extracelular	0.54	1.7475964
54	WP_000791585.1	Extracelular	0.51	0.60872611
55	WP_000739534.1	Extracelular	0.42	0.3878042
56	WP_011382189.1	Extracelular	0.49	1.1159722
57	WP_011382190.1	Extracelular	0.64	-0.031650407
58	WP_000751271.1	Extracelular	0.83	0.69321702
59	WP_000782623.1	Extracelular	0.61	1.3408005
60	WP_001038856.1	Extracelular	0.38	0.90723975
61	WP_000935742.1	Extracelular	0.74	1.5003263
62	WP_000877368.1	Extracelular	0.21	1.1290507
63	WP_000734805.1	Extracelular	1,18	0.063121157
64	WP_000455860.1	Extracelular	0,39	0.1311344
65	WP_000739227.1	Extracelular	0.41	1.2468353
66	WP_000737705.1	Extracelular	0,60	1.1290507
67	WP_000764692.1	Extracelular	0.67	1.1055156
68	WP_001236366.1	Extracelular	0.64	1.3155267
69	WP_011382313.1	Extracelular	0.63	0.79854474
70	WP_000863556.1	Extracelular	0.59	0.96493836
71	WP_000733430.1	Extracelular	1,19	-0.22720683
72	WP_000673053.1	Extracelular	0.65	= 1.035856
73	WP_000673032.1	Extracelular	0.64	1.0351747
74	WP_000166905.1	Extracelular	0.65	0.96227907
75	WP_000874683.1	Extracelular	0.51	0.76354715

76	WP_000841349.1	Extracelular	0,49	0.94236107
77	WP_000201397.1	Extracelular	0,75	0.24643374
78	WP_000948331.1	Extracelular	0,53	-0.053846146
79	WP_011382299.1	Extracelular	0,86	-0.31411624
80	WP_000031105.1	Extracelular	0,74	0.093494448
81	WP_011382342.1	Extracelular	0,84	-0.38744128
82	WP_011382242.1	Extracelular	1,09	0.98429966
83	WP_011382308.1	Extracelular	0,28	-0.45243987
84	WP_011382228.1	Extracelular	0,32	-57284888
85	WP_001790667.1	Extracelular	1,03	-0.035892645

Legenda: Ant Antigenicidade; Aler Alergenicidade. **Fonte:** Autores (2025).

Conforme indicado pelo resultado imunogênico, a avaliação da isoforma (WP_001035596.1) selecionada cumpriu os critérios de antigenicidade e não alergenidade, o que sugere que a proteína selecionada neste estudo pode ser uma opção ideal para o desenvolvimento de um construto vacinal contra a mastite bovina (Figura 1).

Figura 1: Estrutura 3D da toxina-1 da síndrome do choque tóxico (TSST-1) de *Staphylococcus aureus*. PDB DOI: 10.2210/pdb4OHJ/pdb.

A



B 1- MNKKLLMKFFIVSPLLLATIATGFTVPVPLSSNQIHKTAkastNDNIKDLLDWYSSGSDAFT
NSEVLDNSLGSMRIKNTDGSISLIIFSPSY SPTFSKGEKVDLNTKRTKKSQHTSEGTWIHF
QISGVTINTEKLPTPIELPLKVKVHKGDSPLKYWPKFDDKKQLAISTLDFEIRHQLTIQHCL
YRSSDKTGGYWKITMNDGSTYQSDLSKKFEYNTEKPPINIDEIKTIEAEIN - 234

Fonte: Autoria própria (2025).

• Obtenção dos epítopos de Células B

Um total de oito epítopos de células B foram previstos a partir da proteína WP_001035596.1 usando a ferramenta IEDB. Os epítopos de células B são capazes de se

ligar a imunoglobulinas ou anticorpos e são conhecidos como antigênicos. Dos oito epítomos encontrados, apenas três foram distinguidos como imunogênicos e antigênicos (Tabela 2), exibindo a capacidade de ativar o sistema imunológico. Esses epítomos são Pep1 (LMKFFIVSPLLLATI), Pep5 (HFQISG) e Pep8 (RHQLTQIHGLY), sendo selecionados para as etapas subsequentes.

A avaliação da construção prevista da vacina usando o servidor VaxiJen v.2.0 apresentou antigenicidade para as sequências dos Pep1, Pep5 e Pep8, com valores de 1.0369, 1.6775 e 0.0416, respectivamente. Com base nos resultados da previsão do AllerTOP e do ToxinPred, as sequências foram submetidas a uma análise toxicológica para identificar possíveis agentes tóxicos. Ambas as sequências foram confirmadas como naturalmente alergênicas e não tóxicas (Tabela 2).

Tabela 2: Dados referentes aos peptídeos selecionados, ID de identificação, comprimento da cadeia peptídica, pontuação da imunogenicidade e de antigenicidade.

Peptídeo	Sequência	Comprimento	Ant*	Aler*	Tox*
Pep1	LMKFFIVSPLLLATI	15	1.03	NA*	NT*
Pep2	FTPVPLSSN	9	0.63	A*	NT*
Pep3	KDLLDW	6	-1.56	A*	NT*
Pep4	ISLIIFSPYYST	14	0.90	A*	NT*
Pep5	HFQISG	6	1.67	NA*	NT*
Pep6	PTPIELPLKVKVHGKD	16	1.07	A*	NT*
Pep7	KQLAISTLD	9	0.17	A*	NT*
Pep8	RHQLTQIHGLY	11	0.04	NA*	NT*

Fonte: Autoria própria (2025).

As vacinas multiepítomo oferecem uma vantagem significativa ao induzirem tanto imunidade humoral quanto celular, sendo mais eficazes do que vacinas monovalentes. Por meio da abordagem de imunoinformática, epítomos potenciais foram analisados com base em critérios como imunogenicidade, antigenicidade, ausência de toxicidade e características físico-químicas (Hajighahramani *et al.*, 2017). Nesse contexto, epítomos de células B foram previstos e utilizados para o desenvolvimento da construção vacinal. Esses epítomos foram projetados para se ligar ao receptor TLR-2, um passo essencial para avaliar a expressão, estabilização e dobramento do peptídeo vacinal (Noli Truant *et al.*, 2022; Vaillant, Sabir e Jan, 2024).

Adicionalmente, o uso do ligante KK em conjunto com os epítomos de células B desempenhou um papel fundamental na apresentação eficiente de cada epítomo aos anticorpos. Esse processo ocorre por meio do processamento pela protease lisossomal cathepsina B, promovendo a apresentação do antígeno pelo MHC-II. Essa interação é

amplamente reconhecida como uma etapa crucial no desenvolvimento de vacinas multiepitópicas *in silico* e tem sido empregada com sucesso em projetos de vacinologia reversa (Tarrahimofrad *et al.*, 2021).

A previsão das propriedades físico-químicas do construto vacinal desempenha um papel relevante no desenvolvimento de candidatos promissores, uma vez que essas características impactam diretamente sua estabilidade e funcionalidade (Martinelli, 2022). Por exemplo, o II revelou que a proteína possui uma natureza estável após a expressão, favorecendo sua aplicabilidade. Adicionalmente, o IA e o escore GRAVY confirmaram a termoestabilidade e a hidrofílica do construto, respectivamente, indicando propriedades fundamentais para o desenho vacinal (Martinelli, 2022; Kumar *et al.*, 2023).

Com base nesses parâmetros, a proteína CP27 foi selecionada devido às suas propriedades físico-químicas aceitáveis, incluindo pI, II e hidropaticidade. Para analisar essas características, a sequência de aminoácidos da proteína foi submetida ao servidor ProtParam, que indicou um valor de pI igual a 10,68, evidenciando a natureza básica do construto vacinal (Motamedi *et al.*, 2023). Os II, determinados com base em uma escala onde valores menores que 40 indicam proteínas estáveis, apresentaram um resultado de 24 para o construto, corroborando sua estabilidade após a expressão (Asghari *et al.*, 2022). Em contrapartida, o IA foi calculado como 119, sugerindo que a proteína apresenta alta termoestabilidade, o que pode ser considerado uma consistência em condições de temperatura corpórea humana (Takano e Yutani, 2001).

No que diz respeito à hidropaticidade, o valor médio do GRAVY foi estimado em 0,175, indicando uma natureza predominantemente hidrofóbica. Esse resultado sugere que a proteína CP27 possui maior afinidade para interagir com outras proteínas, em vez de moléculas de água, favorecendo seu potencial como componente vacinal (Dzayee *et al.*, 2022; Hajighahramani *et al.*, 2017; Noli Truant *et al.*, 2022).

Adicionalmente, o pI teórico e o II reforçam a estabilidade e a natureza básica da proteína após a expressão, evidenciando melhorias em sua capacidade de uso. Por outro lado, os dados relacionados à termoestabilidade, representados pelo IA, apontam um resultado positivo para aplicação em estudos *in vitro*. Esses resultados destacam como a combinação de termoestabilidade e hidrofílicidade, representada pelo IA e pelo GRAVY, pode fornecer insights fundamentais para o refinamento e a viabilidade de vacinas multiepitótopo (Dzayee *et al.*, 2022).

O projeto de construção da vacina levou em consideração os seguintes fatores principais: a vacina deve ser não tóxica, não alergênica, imunogênica e antigênica. Com

base nesses critérios, um construto vacinal foi gerado, ligando os três epítomos de células B aos respectivos ligantes KK correspondentes. Ao utilizar esses ligantes KK juntamente com os três epítomos, foi projetada uma construção vacinal linear com 36 aminoácidos (Tabela 3).

Diante disso, por meio da ferramenta ProtParam, observou-se que o construto vacinal atendeu a todos os parâmetros relevantes, como pI de 10.68, II de 24, IA de 119, GRAVY de 0.175, além de apresentarem antigenicidade (0.80), não serem alergênicos, não tóxicos e possuírem expressividade em sistema heterólogo (>50).

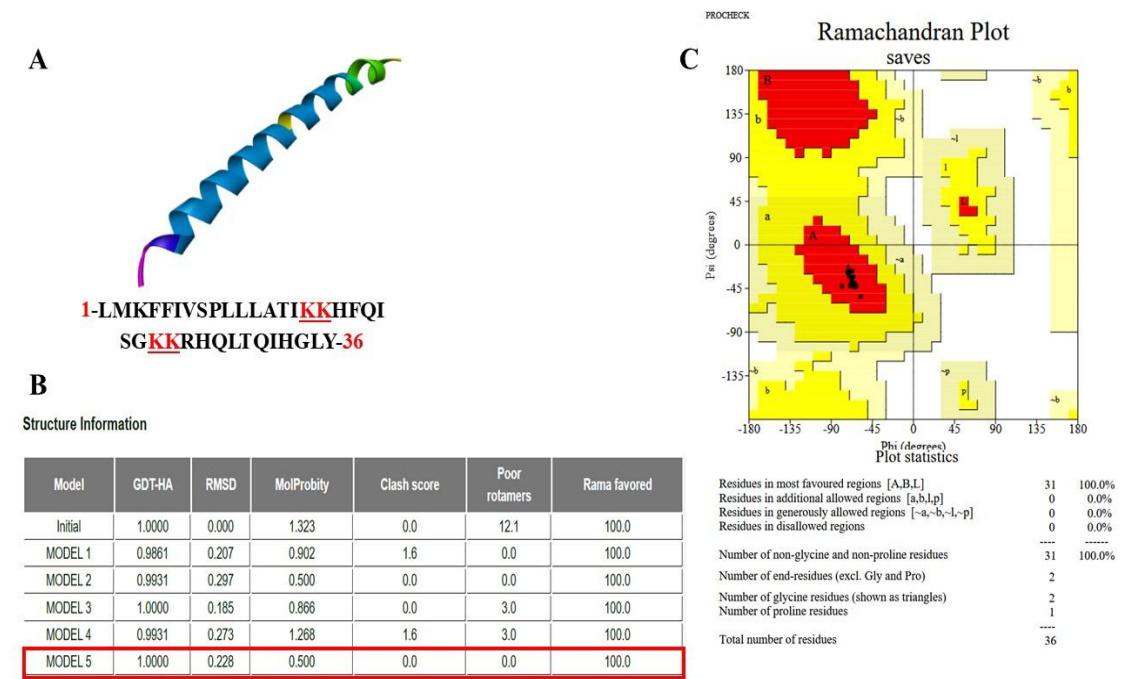
Tabela 3: Dados referentes ao construto vacinal e seus parâmetros físico-química.

Construto vacinal								
LMKFFIVSPLLLATIKKHFQISGKKRHQLTQIHGLY								
PM	pI	II	IA	GRAVY	Ant*	Aler*	Tox*	SOLpro
4236.18	10.68	24	119	0.175	0.80	Na*	NT*	0.77

Legenda: PM Peso molecular; pI Ponto isoelétrico; II Índice de instabilidade; IA Índice Alifático; GRAVY Grande média de hidropatia; Tox Toxicidade. **Fonte:** Autoria própria (2025).

A estrutura 3D da vacina gerada pelo *AlphaFold 3* e refinada pelo *GalaxyRefine* é apresentada juntamente com sua validação por meio do gráfico de *Ramachandran*. Essas análises de refinamento e validação foram empregadas para avaliar a qualidade da estrutura prevista (Figura 2). O melhor modelo selecionado entre os cinco previstos pelo *GalaxyRefine* foi o modelo 5, indicando alta similaridade em relação ao modelo inicial, conforme evidenciado pela pontuação do Teste de Distância Global – Alta Precisão (*GDT-HA*), que alcançou valores elevados, bem como pelas pontuações do *MolProbity*, que foram inferiores aos modelos iniciais, demonstrando que o processo de refinamento reduziu os erros críticos nos modelos 3D. A análise do gráfico de *Ramachandran* revelou que 100,0% dos resíduos estavam na zona favorecida, 0,0% na região permitida generosa e 0,0% na região discrepante, indicando que a qualidade geral da vacina é ótima (Figura 2).

Figura 2: Estratégias utilizadas na concepção de uma vacina multi-epítopo: (A) Estrutura da construção da vacina projetada em modelo tridimensional e sequência de aminoácidos do modelo 5 da vacina previsto pelo GalaxyRefine. Os alfabetos pretos indicam os epítopos imunodominantes de células B, enquanto os alfabetos vermelhos sublinhados representam os ligantes KK. (B) Análise do GalaxyRefine e seus cinco modelos previstos, sendo o modelo 5 selecionado. (C) Análise do gráfico de Ramachandran do modelo 5.



Fonte: Autoria própria (2025).

• **Docking molecular da vacina com receptor TLR-2**

O modelo 5 submetido ao HADDOCK 2.4 gerou cinco clusters. Após análise da menor pontuação HADDOCK, apenas o cluster 2 foi selecionado para avaliações posteriores (Tabela 4).

Tabela 4: Análise dos clusters fornecidos pelo HADDOCK seguida de sua pontuação de acoplamento.

Cluster	1	2	3	4	5	6	7	8	9
Pontuação HADDOCK	-72.1	-95.5	-81.8	-87.1	-50.0	-61.0	-66.0	-70.8	-77.9

Fonte: Autoria própria (2025).

Dando sequência à análise das propriedades físico-químicas e imunológicas do construto vacinal, a interação com os receptores Toll-like (TLRs) desempenha um papel crucial no reconhecimento de proteínas bacterianas e na ativação de respostas imunológicas. Esses receptores, especialmente o TLR2, são essenciais para a indução de mediadores pró-inflamatórios, como interferons e interleucinas, fortalecendo a imunidade inata contra infecções bacterianas (Hajighahramani *et al.*, 2017).

Os TLRs interagem com estruturas específicas, como as regiões ricas em leucina (LRR, do inglês leucine-rich repeat), que conferem uma configuração estrutural em forma de fechadura, crucial para o reconhecimento molecular. Esse domínio conservado é responsável pela interação com estruturas proteicas e desempenha um papel chave em infecções bacterianas, conforme evidenciado em estudos prévios (Dey *et al.*, 2021).

No presente estudo, a vacina foi projetada para se acoplar ao TLR2, um dos receptores mais bem caracterizados. Essa escolha baseia-se na sua capacidade de ativar vias de sinalização inflamatória e mediar a expressão de peptídeos antimicrobianos, citocinas e quimiocinas, direcionando células imunológicas ao local da infecção e promovendo a resposta imune adaptativa (Hajighahramani *et al.*, 2017). A utilização do TLR2 como molécula receptora tem se mostrado eficaz em abordagens *in silico* para o desenvolvimento de vacinas contra *Staphylococcus aureus*, reforçando sua relevância no contexto deste estudo (Dey *et al.*, 2021; Hajighahramani *et al.*, 2017).

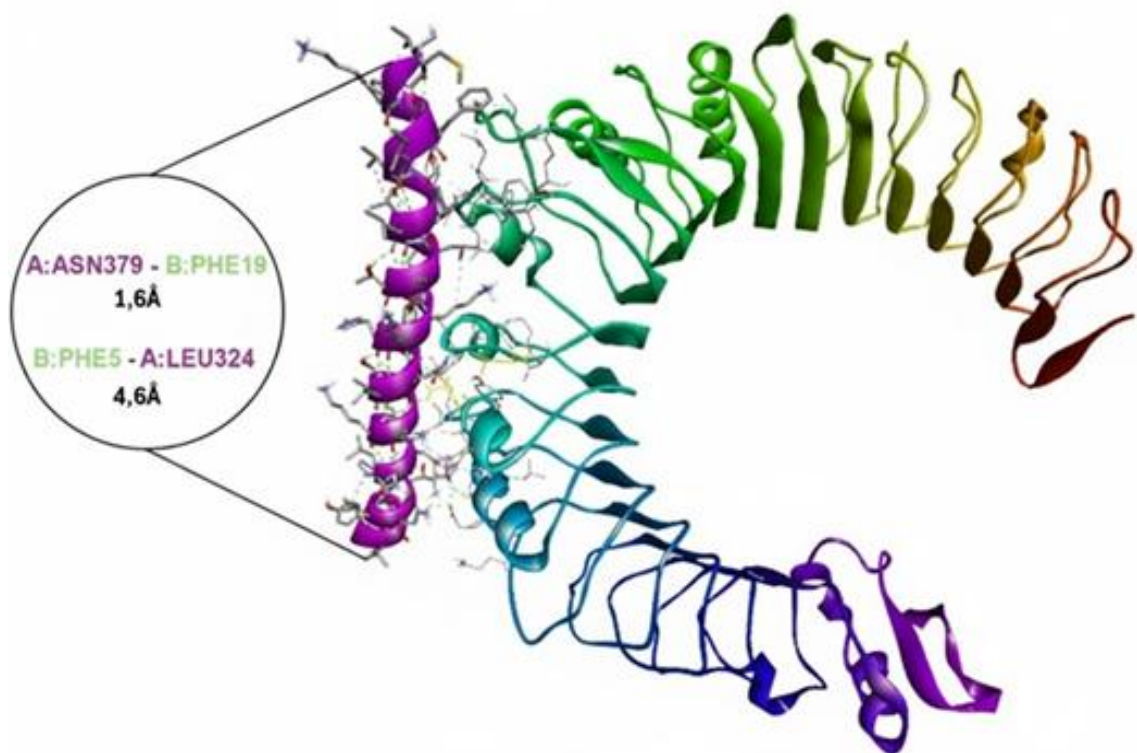
Para modelar o complexo vacina-receptor, utilizou-se o servidor HADDOCK, que revelou interações específicas entre a vacina e os resíduos Leu324 e Asn379 do TLR2. Observou-se uma ligação de hidrogênio entre os resíduos A:ASN379 e B:PHE19, com distância de 1,6 Å, e uma interação hidrofóbica entre os resíduos B:PHE5 e A:LEU324, com distância de 4,6 Å. Esses resíduos também foram identificados em estudos anteriores, mostrando interações no patch hidrofóbico contínuo do TLR2, próximo à hélice H5 e ao loop $\beta 2$ – $\beta 3$ (Koymans *et al.*, 2015).

As interações observadas reforçam a similaridade estrutural entre os complexos SSL3-TLR2 de humanos e camundongos, evidenciando a robustez do modelo *in silico* e

a potencial eficácia do construto vacinal no desencadeamento de respostas imunológicas (Koymans *et al.*, 2018). Esses achados destacam o papel do TLR2 como uma âncora estratégica no desenho de vacinas multiepitopo e abrem novas perspectivas para o desenvolvimento de imunoterapias contra infecções bacterianas.

Para verificar uma possível interação entre o TLR-2 e a vacina, a Figura 3 apresenta o docking molecular do cluster 2, que foi escolhido por apresentar uma ligação de hidrogênio entre os resíduos de aminoácidos A:ASN379 e B:PHE19, com distância de 1,6 Å, e uma interação hidrofóbica entre os resíduos B:PHE5 e A:LEU324, com distância de 4,6 Å.

Figura 3: Docking molecular entre a vacina com o receptor TLR-2. A predição do docking molecular foi realizado usando o software HADDOCK 2.4 e a proteína TLR-2 foi obtida no PDB (5D3I) e foi preparada, anteriormente, para retirada do ligante pelo Discovery Studio. As imagens de interação foram obtidas pelo PyMOL. Previsão de interação da vacina (em roxo) com o TLR-2 (policromática).



Fonte: Autoria própria (2025).

Considerações Finais

Atualmente, o avanço das ferramentas de imunoinformática permite explorar plataformas para o desenvolvimento de vacinas multiepítomos altamente imunogênicas, econômicas e de baixo risco, especialmente no combate à mastite bovina. No presente projeto, foram utilizados sucessivamente diferentes softwares e ferramentas de imunoinformática para o desenvolvimento de uma nova vacina multiepítomo candidata que pode ser eficaz contra infecções por *S. aureus*. Os epítomos de células B identificados mostraram elevada antigenicidade, com forte ligação ao receptor TLR2. Além disso, análises detalhadas, como a caracterização físico-química e o *docking* molecular com a molécula receptora, confirmaram o potencial da vacina para ser avaliada experimentalmente quanto à resposta imune inata.

A construção da vacina não apenas atenderá à necessidade de combater os casos crescentes de mastite bovina a nível mundial, mas também se espera que seja eficaz contra a linhagem *S. aureus* RF122. Adicionalmente, junto à eficácia, as vacinas desenvolvidas a partir desses métodos foram consideradas seguras; no entanto, estudos laboratoriais adicionais continuam sendo essenciais. A aprovação, seguida da produção em massa e da disponibilização dessas vacinas, pode proporcionar uma solução acessível e eficaz para a prevenção global da mastite bovina.

As características das vacinas desenvolvidas neste estudo estão alinhadas com esse objetivo. Os três epítomos, Pep1 (LMKFFIVSPLLLATI), Pep5 (HFQISG) e Pep8 (RHQLTQIHGLY), dos construtos projetados têm o potencial de induzir respostas imunes eficazes, e as sequências adjuvantes podem desempenhar um papel relevante no aumento da imunogenicidade dos construtos. Este estudo contribui para o entendimento do potencial das vacinas baseadas em epítomos como abordagens importantes tanto protetoras quanto terapêuticas para a imunização bacteriana. Estudos futuros avaliarão se as vacinas multiepítomos podem induzir respostas imunes e proporcionar proteção contra *Staphylococcus aureus* RF122 e outras linhagens.

Referências

ABRIL, A. G.; VILLA, T. G.; BARROS-VELÁZQUEZ, J.; CAÑAS, B.; SÁNCHEZ-PÉREZ, A.; CALO-MATA, P.; CARRERA, M. *Staphylococcus aureus* Exotoxins and Their Detection in the Dairy Industry and Mastitis. **Toxins** **2020**, Vol. 12, Page 537, v. 12, n. 9, p. 537, 20 ago. 2020.

- ASGHARI, A.; KORDI, B.; MALEKI, B.; MAJIDIANI, H.; SHAMS, M.; NASERIFAR, R. Neosporea caninum SRS2 protein: essential vaccination targets and biochemical features for next-generation vaccine design. *BioMed Research international* 2022; 2022.
- BIBI, S.; ULLAH, I.; ZHU, B.; ADNAN, M.; LIAQAT, R.; KONG, W. B.; NIU, S. In silico analysis of epitope-based vaccine candidate against tuberculosis using reverse vaccinology. **Scientific Reports** 2021 **11:1**, v. 11, n. 1, p. 1–16, 13 jan. 2021.
- CLEMENTS, B. W.; CASANI, J. A. P. Emerging and Reemerging Infectious Disease Threats. **Disasters and Public Health**, p. 245, 2016.
- COBIRKA, M.; TANCIN, V.; SLAMA, P. Epidemiology and Classification of Mastitis. **Animals** 2020, **Vol. 10**, **Page 2212**, v. 10, n. 12, p. 2212, 26 nov. 2020.
- DEY, J.; MAHAPATRA, S. R.; SINGH, P.; PATRO, S.; KUSHWAHA, G. S.; MISRA, N.; SUAR, M. B and T cell epitope-based peptides predicted from clumping factor protein of *Staphylococcus aureus* as vaccine targets. **Microbial Pathogenesis**, v. 160, p. 105171, 1 nov. 2021.
- DZAYEE, S. A.; KHUDHUR, P. K.; MAHMOOD, A.; MARKOV, A.; MASELENO, A.; EBRAHIMPOUR GORJI, A. Computational design of a new multi-epitope vaccine using immunoinformatics approach against mastitis disease. **Animal Biotechnology**, v. 33, n. 6, p. 1359–1370, 18 nov. 2022.
- EL-SAYED, A.; KAMEL, M. Bovine mastitis prevention and control in the post-antibiotic era. **Tropical Animal Health and Production** 2021 **53:2**, v. 53, n. 2, p. 1–16, 31 mar. 2021.
- EWIDA, R. M.; AL-HOSARY, A. A. T. Prevalence of enterotoxins and other virulence genes of *Staphylococcus aureus* caused subclinical mastitis in dairy cows. **Veterinary World**, v. 13, n. 6, p. 1193, 1 jun. 2020.
- GEBRE, M. S.; BRITO, L. A.; TOSTANOSKI, L. H.; EDWARDS, D. K.; CARFI, A.; BAROUCH, D. H. Novel approaches for vaccine development. **Cell**, v. 184, n. 6, p. 1589, 18 mar. 2021.
- HAJIGHAHRAMANI, N.; NEZAFAT, N.; ESLAMI, M.; NEGAHDARIPOUR, M.; RAHMATABADI, S. S.; GHASEMI, Y. Immunoinformatics analysis and in silico designing of a novel multi-epitope peptide vaccine against *Staphylococcus aureus*. **Infection, Genetics and Evolution**, v. 48, p. 83–94, 1 mar. 2017.
- HASHEMZADEH, P.; GHORBANZADEH, V.; LASHGARIAN, H. E.; KHEIRANDISH, F.; DARIUSHNEJAD, H. Harnessing Bioinformatic Approaches to Design Novel Multi-epitope Subunit Vaccine Against *Leishmania infantum*. **International Journal of Peptide Research and Therapeutics**, v. 26, n. 3, p. 1417–1428, 1 set. 2020.
- KOYMANS, K. J.; FEITSMA, L. J.; BISSCHOP, A.; HUIZINGA, E. G.; STRIJP, J. A. G. VAN; HAAS, C. J. C. DE; MCCARTHY, A. J. Molecular basis determining species specificity for TLR2 inhibition by staphylococcal superantigen-like protein 3 (SSL3). **Veterinary Research**, v. 49, n. 1, p. 115, 28 nov. 2018.

KOYMANS, K. J.; FEITSMA, L. J.; BRONDIJK, T. H. C.; AERTS, P. C.; LUKKIEN, E.; LÖSSL, P.; KESSEL, K. P. M. VAN; HAAS, C. J. C. DE; STRIJP, J. A. G. VAN; HUIZINGA, E. G. Structural basis for inhibition of TLR2 by staphylococcal superantigen-like protein 3 (SSL3). **Proceedings of the National Academy of Sciences of the United States of America**, v. 112, n. 35, p. 11018–11023, 1 set. 2015.

KUMAR, K. M. et al. Immunoinformatic exploration of a multi-epitope-based peptide vaccine candidate targeting emerging variants of SARS-CoV-2. **Frontiers in Microbiology**, v. 14, p. 1251716, 2023.

MA, Q.; WEI, Y.; MENG, Z.; CHEN, Y.; ZHAO, G. Effects of Water Extract from *Artemisia argyi* Leaves on LPS-Induced Mastitis in Mice. **Animals** 2022, Vol. 12, Page 907, v. 12, n. 7, p. 907, 1 abr. 2022.

MARTINELLI, Dominic D. In silico vaccine design: A tutorial in immunoinformatics. **Healthcare Analytics**, v. 2, p. 100044, 2022.

MUKHERJEE, S.; HUDA, S.; SINHA BABU, S. P. Toll-like receptor polymorphism in host immune response to infectious diseases: A review. **Scandinavian Journal of Immunology**, v. 90, n. 1, p. e12771, 1 jul. 2019.

MOTAMEDI, H.; ALVANDI, A.; FATHOLLAHI, M.; ARI, M. M.; MORADI, S.; MORADI, J.; ABIRI, R. In silico designing and immunoinformatics analysis of a novel peptide vaccine against metallo-beta-lactamase (VIM and IMP) variants. **PloS One** 2023; 18(7): e0275237.

NOLI TRUANT, S.; MARZI, M. C. DE; SARRATEA, M. B.; ANTONOGLOU, M. B.; MEO, A. P.; IANNANTUONO LÓPEZ, L. V.; FERNÁNDEZ LYNCH, M. J.; TODONE, M.; MALCHIODI, E. L.; FERNÁNDEZ, M. M. egc Superantigens Impair Monocytes/Macrophages Inducing Cell Death and Inefficient Activation. **Frontiers in Immunology**, v. 10, p. 474389, 15 jan. 2020.

NOLI TRUANT, S.; REDOLFI, D. M.; SARRATEA, M. B.; MALCHIODI, E. L.; FERNÁNDEZ, M. M. Superantigens, a Paradox of the Immune Response. **Toxins**, v. 14, n. 11, p. 800, 18 nov. 2022.

NOOR, F.; AHMAD, S.; SALEEM, M.; ALSHAYA, H.; QASIM, M.; REHMAN, A.; EHSAN, H.; TALIB, N.; SALEEM, H.; JARDAN, Y. A. BIN; ASLAM, S. Designing a multi-epitope vaccine against *Chlamydia pneumoniae* by integrating the core proteomics, subtractive proteomics and reverse vaccinology-based immunoinformatics approaches. **Computers in Biology and Medicine**, v. 145, p. 105507, 1 jun. 2022.

OLIVER, S. P.; MURINDA, S. E. Antimicrobial resistance of mastitis pathogens. **The Veterinary clinics of North America. Food animal practice**, v. 28, n. 2, p. 165–185, jul. 2012.

PARVIZPOUR, S.; POURSEIF, M. M.; RAZMARA, J.; RAFI, M. A.; OMIDI, Y. Epitope-based vaccine design: a comprehensive overview of bioinformatics approaches. **Drug Discovery Today**, v. 25, n. 6, p. 1034–1042, 1 jun. 2020.

- RAS-CARMONA, A.; LEHMANN, A. A.; LEHMANN, P. V.; RECHE, P. A. Prediction of B cell epitopes in proteins using a novel sequence similarity-based method. **Scientific Reports** **2022 12:1**, v. 12, n. 1, p. 1–9, 12 ago. 2022.
- RONCO, T.; KLAAS, I. C.; STEGGER, M.; SVENNESEN, L.; ASTRUP, L. B.; FARRE, M.; PEDERSEN, K. Genomic investigation of *Staphylococcus aureus* isolates from bulk tank milk and dairy cows with clinical mastitis. **Veterinary Microbiology**, v. 215, p. 35–42, 1 fev. 2018.
- ROSTAMTABAR, M.; RAHMANI, A.; BAEI, M.; KARKHAH, A.; PRAJAPATI, V. K.; EBRAHIMPOUR, S.; NOURI, H. R. Development a multi-epitope driven subunit vaccine for immune response reinforcement against Serogroup B of *Neisseria meningitidis* using comprehensive immunoinformatics approaches. **Infection, Genetics and Evolution**, v. 75, p. 103992, 1 nov. 2019.
- SETTE, A.; RAPPUOLI, R. Reverse Vaccinology: Developing Vaccines in the Era of Genomics. **Immunity**, v. 33, n. 4, p. 530, 29 out. 2010.
- TAKANO, K.; YUTANI, K. A new scale for side-chain contribution to protein stability based on the empirical stability analysis of mutant proteins. **Protein Engineering, Design and Selection**, v. 14, n. 8, p. 525–528, 1 ago. 2001.
- TARRAHIMOFRAD, H.; RAHIMNAHAL, S.; ZAMANI, J.; JAHANGIRIAN, E.; AMINZADEH, S. Designing a multi-epitope vaccine to provoke the robust immune response against influenza A H7N9. **Scientific Reports** **2021 11:1**, v. 11, n. 1, p. 1–22, 29 dez. 2021.
- VAILLANT, A. A. J.; SABIR, S.; JAN, A. Physiology, Immune Response. **StatPearls**, 27 jul. 2024.
- ZAATOUT, N.; AYACHI, A.; KECHA, M. Epidemiological investigation of subclinical bovine mastitis in Algeria and molecular characterization of biofilm-forming *Staphylococcus aureus*. **Tropical Animal Health and Production**, v. 52, n. 1, p. 283–292, 1 jan. 2020.

CAPÍTULO 5

IN SILICO IDENTIFICATION AND CHARACTERIZATION OF CALEOSIN GENES IN CASHEW (*Anacardium occidentale* L.)

IDENTIFICAÇÃO E CARACTERIZAÇÃO IN SILICO DOS GENES DA CALEOSINA EM CAJU (*Anacardium occidentale* L.)

DOI: <https://doi.org/10.56001/26.9786501906539.05>

Submetido em: 19/10/2026

Revisado em: 19/10/2026

Publicado em: 24/10/2026

Maria Cidinaria Silva Alves

Universidade Federal de Pernambuco, Centro de Biociências, Departamento de Genética,
Recife-PE, Brasil

<https://orcid.org/0000-0003-0625-416X>

Rousilândia de Araujo Silva

Universidade Federal de Pernambuco, Departamento de Ciências Farmacêuticas,
Recife-PE, Brasil

<https://orcid.org/0000-0002-9944-0684>

Gleyce Hellen de Almeida de Souza

Universidade Federal da Grande Dourados, Programa de Pós-Graduação em Ciências da
Saúde, Dourados-MS, Brasil

<https://orcid.org/0000-0001-8860-9109>

José Eduardo Souza Echeverria

Universidade Federal da Grande Dourados, Programa de Pós-Graduação em Ciências da
Saúde, Dourados-MS, Brasil

<https://orcid.org/0009-0006-6311-260X>

Fabíola Lucini

Universidade Federal da Grande Dourados, Programa de Pós-Graduação em Ciências da
Saúde, Dourados-MS, Brasil

<https://orcid.org/0000-0001-6658-4043>

Ruana Carolina Cabral da Silva

Universidade Federal da Grande Dourados, Programa de Pós-Graduação em Ciências da

Saúde, Dourados-MS, Brasil

<https://orcid.org/0000-0003-0435-1871>

Abstract

Caleosins are lipid-associated proteins involved in plant lipid metabolism, development, and environmental responses. This study aimed to identify and characterize candidate caleosin proteins in *Anacardium occidentale* L. using an integrated in silico approach. Candidate sequences were identified by similarity searches and conserved-domain analysis, followed by physicochemical characterization, subcellular localization prediction, conserved motif and gene structure analyses, phylogenetic reconstruction, functional annotation, and examination of three-dimensional models. Six candidates, designated AoCLO1–AoCLO6, were identified. Five proteins contained 176 amino acids, whereas AoCLO4 contained 113 amino acids. Molecular weights ranged from 12.60 to 20.19 kDa, theoretical isoelectric points ranged from 5.52 to 9.40, and all candidates exhibited negative GRAVY values. Predicted subcellular localizations included the chloroplast, nucleus, and cytoplasm. Ten sequence motifs were detected, three of which were shared by all candidates, while exon–intron organization varied among the corresponding genes. The AoCLO proteins were distributed between two of the three groups highlighted in the phylogenetic analysis. Functional annotations included terms associated with abscisic acid responses, seed dormancy, lipid droplet organization, and peroxxygenase activity. Structural models displayed α -helices connected by loops and heterogeneous surface electrostatic potentials. Structural validation revealed differences in stereochemical and global model quality, supporting cautious interpretation of the predicted structures. These findings establish a comparative framework for investigating cashew caleosins and identify candidates for experimental studies of lipid metabolism, development, and stress-related processes. The proposed functional associations remain computational predictions requiring biological validation.

Keywords: *Anacardium occidentale*; Caleosin; Conserved motifs; Functional annotation; Protein modeling.

Introduction

Anacardium occidentale L. (Anacardiaceae), commonly known as cashew, is a tropical tree crop cultivated for its edible kernels and cashew apples. Understanding the genetic basis of economically relevant traits is important for advancing research and breeding in this species. The publication of a draft cashew genome has expanded the resources available for investigating gene families and their potential contributions to plant development and environmental responses. These genomic resources provide a foundation for identifying candidate genes and examining their sequence characteristics and evolutionary relationships (Savadi *et al.*, 2022).

Caleosins, also referred to as caleosin/peroxygenases (CLO/PXGs), constitute a family of lipid-associated proteins found in land plants and some green algae and fungi. Initially characterized as components of plant lipid droplets, these proteins combine

structural properties with enzymatic functions associated with the metabolism of lipid-derived substrates. Typical plant caleosins contain an EF-hand calcium-binding motif and a central hydrophobic region involved in membrane or lipid-droplet association. Some members also exhibit heme-dependent peroxygenase activity, linking this protein family to lipid oxidation and the formation of signaling molecules (Hanano *et al.*, 2023).

Experimental studies have demonstrated that caleosins participate in processes beyond lipid-droplet organization. In *Arabidopsis thaliana*, Liu *et al.* (2022) showed that multiple caleosins have overlapping functions in seed oil accumulation and embryo development. Their findings indicate that individual family members can contribute to related biological processes, highlighting the importance of considering functional redundancy when interpreting caleosin gene families. Furthermore, Miklaszewska *et al.* (2023) demonstrated that CALEOSIN 1 interacts with AUTOPHAGY-RELATED PROTEIN 8 and facilitates lipid-droplet microautophagy in seedlings, providing a mechanistic connection between caleosins and the mobilization of lipid reserves.

Caleosins have also been implicated in the integration of environmental responses and cellular signaling. In *A. thaliana*, the stress-induced caleosin RD20/CLO3 interacts with GPA1, the alpha subunit of the heterotrimeric G-protein complex, and acts as a negative regulator of this signaling component. The induction of RD20/CLO3 by drought, salt, and abscisic acid supports its involvement in pathways associated with abiotic stress responses (Brunetti *et al.*, 2021). These experimentally established functions provide a rationale for investigating caleosin homologs in other plant species, although their conservation in cashew requires specific evaluation.

Comparative genomic analyses have revealed variation in the distribution and diversification of caleosin/peroxygenase genes across the Viridiplantae. Differences in family composition among plant lineages indicate that conserved sequence features coexist with evolutionary diversification, making comparative analyses useful for investigating relationships among family members (Rahman *et al.*, 2018).

In this context, the identification of caleosin candidates in the cashew genome, together with analyses of conserved domains, gene organization, and phylogenetic relationships, can establish a framework for studying this protein family in *A. occidentale*. Structural modeling can complement these analyses by enabling the examination of predicted conformations and conserved residues, while functional annotations provide hypotheses to guide subsequent experimental investigation.

Therefore, this study aimed to identify and characterize candidate caleosin genes in the genome of *A. occidentale* using an integrated in silico approach. The analyses encompassed physicochemical properties, predicted subcellular localization, conserved domains and motifs, gene structure, phylogenetic relationships, functional annotation, and three-dimensional protein modeling. By integrating sequence-based and structural evidence, this study seeks to establish a molecular framework for investigating cashew caleosins and to identify candidates for future studies of lipid metabolism, plant development, and stress responses.

Materials and Methods

- **Identification and physicochemical characterization of caleosin proteins**

Candidate caleosin proteins were identified in the annotated proteome of *A. occidentale* through BLASTP searches using previously characterized plant caleosin sequences as queries. The presence of the caleosin domain was verified using the NCBI Conserved Domain Database through the CD-Search tool (Wang *et al.*, 2023). The identified candidates were designated AoCLO1–AoCLO6. Protein length, theoretical molecular weight, isoelectric point, and grand average of hydropathicity (GRAVY) were calculated using ProtParam, available through the ExPASy platform. Subcellular localization was predicted using WoLF PSORT.

- **Conserved motifs and gene structure**

Conserved amino acid motifs were analyzed using MEME Suite version 5.5.3. Motif distribution was compared among the AoCLO proteins to identify shared sequence features and differences among family members. Exon–intron organization was examined using Gene Structure Display Server version 2.0, based on the corresponding gene annotation and sequence data.

- **Sequence alignment and phylogenetic analysis**

Caleosin protein sequences from representative plant species were retrieved from UniProt and analyzed together with the AoCLO sequences. Multiple sequence alignment was performed using ClustalW within MEGA version 7.0. The phylogenetic tree was reconstructed using the Neighbor-Joining method with the p-distance model, and branch support was evaluated using 1,000 bootstrap replicates (Kumar *et al.*, 2016). The resulting

tree was visualized and annotated using the Interactive Tree of Life platform (iTOL) (Letunic and Bork, 2016).

- **Functional annotation**

Functional annotation of the AoCLO proteins was performed using eggNOG-mapper version 2. Gene Ontology terms, KEGG Orthology identifiers, and KEGG pathway assignments were retrieved from the annotation output to investigate potential functions associated with the candidate proteins. These assignments were interpreted as computational predictions based on sequence similarity and orthology.

- **Structural analysis and model validation**

Homologous structures available in the AlphaFold Protein Structure Database were selected based on sequence similarity to the AoCLO proteins and used as structural references for modeling (Varadi *et al.*, 2022). The supplied AoCLO structural models were evaluated using MolProbity Ramachandran analysis implemented in CCTBX (version 2025.11), ProSA-web, and QMEANDisCo (Williams *et al.*, 2018; Wiederstein and Sippl, 2007; Studer *et al.*, 2020). Ramachandran distributions were classified into favored, allowed, and outlier regions using residue-specific reference distributions. Terminal residues lacking complete backbone ϕ/ψ angles were excluded from this analysis. ProSA-web Z-scores and QMEANDisCo global scores were obtained for each model. Structural representations were generated using PyMOL.

- **Electrostatic surface analysis**

Structural models were prepared using PDB2PQR version 3.7.1, with protonation states assigned using PROPKA version 3.5.1 at pH 7.0 and atomic charges and radii assigned using CHARMM parameters (Brooks *et al.*, 2009). Electrostatic potentials were calculated using APBS version 3.4.1 by solving the linearized Poisson–Boltzmann equation with multigrid focusing (Jurrus *et al.*, 2018). Calculations used solute and solvent dielectric constants of 4.0 and 78.54, respectively, a temperature of 298.15 K, and monovalent positive and negative ions at concentrations of 0.15 M each. Electrostatic potentials were mapped onto molecular surfaces in PyMOL using a common scale ranging from -5 to $+5$ kBT/e.

Results and Discussion

• Identification and physicochemical properties of AoCLO proteins

Six candidate caleosin proteins were identified in *A. occidentale* and designated AoCLO1–AoCLO6 (Table 1). All candidates contained a caleosin domain. Five proteins consisted of 176 amino acids, whereas AoCLO4 contained 113 amino acids. Predicted molecular weights ranged from 12.60 kDa for AoCLO4 to 20.19 kDa for AoCLO5.

The theoretical isoelectric points ranged from 5.52 to 9.40. AoCLO4 exhibited the highest value (9.40), whereas the remaining proteins presented values between 5.52 and 6.59. All proteins had negative GRAVY values, ranging from –0.488 for AoCLO3 to –0.232 for AoCLO5. Subcellular localization predictions assigned AoCLO2, AoCLO5, and AoCLO6 to the chloroplast, AoCLO1 and AoCLO4 to the nucleus, and AoCLO3 to the cytoplasm.

Table 1. Physicochemical properties and predicted subcellular localization of candidate caleosin proteins in *Anacardium occidentale*.

ID	ID_gene	Domain	Length (aa)	pI	MW (Kda)	GRAVY	Subcellular Localization
AoCLO1	gene_15413	Caleosin	176	6.59	19,92	-0.469	Nucleus
AoCLO2	gene_21080	Caleosin	176	5.52	19,83	-0.237	Chloroplast
AoCLO3	gene_21081	Caleosin	176	6.59	20,03	-0.488	Cytoplasm
AoCLO4	gene_21082	Caleosin	113	9.40	12,60	-0.251	Nucleus
AoCLO5	gene_21725	Caleosin	176	6.54	20,19	-0.232	Chloroplast
AoCLO6	gene_27423	Caleosin	176	5.63	19,85	-0.411	Chloroplast

The identification of six candidate caleosin proteins in *A. occidentale* provides a starting point for investigating this family in cashew. Comparative analyses across the Viridiplantae have shown that caleosin/peroxygenase families vary in composition among plant lineages and include conserved features alongside sequence diversification (Rahman *et al.*, 2018). The AoCLO candidates therefore fit within a broader pattern of multiple caleosin-related sequences in plant genomes. Nevertheless, the number identified here should be interpreted in relation to the genome annotation and search procedure used, rather than as evidence of family expansion or contraction in cashew.

The physicochemical differences among AoCLO proteins indicate variation in amino acid composition despite the identical length of five candidates. AoCLO4 was particularly distinct, combining a shorter sequence with the highest theoretical isoelectric point. Its reduced length, absence of several motifs found in the other candidates, and shorter modeled structure make it a priority for examining sequence completeness and gene annotation. These features may represent biological variation, but the available data do not distinguish a naturally shorter protein from an incomplete predicted sequence. Similarly, the negative GRAVY values describe average sequence hydrophathy and do not exclude the presence of local hydrophobic segments involved in membrane association. This distinction is relevant because caleosins possess a central hydrophobic region associated with their interaction with membranes and lipid droplets (Hanano *et al.*, 2023).

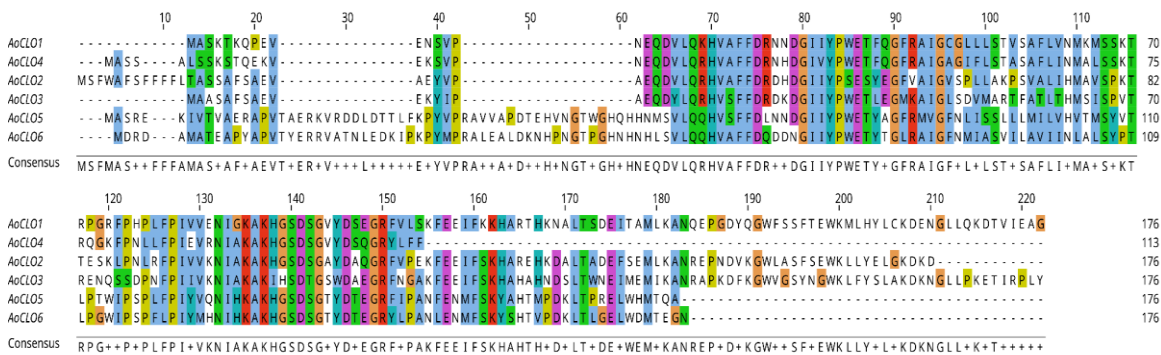
Subcellular localization predictions suggested differences among the AoCLO candidates, with chloroplast, nuclear, and cytoplasmic assignments. However, these predictions were not fully concordant with the cellular component annotations obtained through functional annotation. For example, the two approaches did not assign chloroplast localization to the same set of proteins. This discrepancy reflects the different evidence used by localization predictors and annotation transfer methods and prevents a definitive assignment of cellular distribution. The predicted nuclear localization of AoCLO1 and AoCLO4 should therefore be treated as a testable hypothesis. Fluorescent protein fusion experiments would help determine whether the candidates occupy the predicted compartments or associate them with lipid droplets and other membranes.

- **Multiple sequence alignment**

The multiple sequence alignment revealed conserved regions interspersed with variable segments among the six AoCLO proteins (Figure 1). Sequence conservation was evident in the central portion of the alignment, whereas greater variation was observed in the N-terminal and C-terminal regions. Alignment gaps reflected differences in sequence length and composition.

AoCLO4 shared conserved residues with the other candidates but lacked the C-terminal region present in the longer sequences. AoCLO5 and AoCLO6 contained an additional aligned segment preceding the conserved central region, whereas AoCLO1, AoCLO2, and AoCLO3 showed gaps across much of this segment.

Figure 1. Multiple sequence alignment of the six predicted AoCLO proteins. Colors indicate amino acid classes, dashes represent alignment gaps, and the consensus sequence is shown below each alignment block.



• **Conserved motifs and gene structure**

Ten conserved motifs were detected across the AoCLO proteins (Figure 2A–B). Motifs 1, 5, and 2 were present in all six candidates and occurred in the same relative order. AoCLO1 and AoCLO3 displayed identical motif arrangements, comprising motifs 6, 1, 5, 2, 3, 4, and 9. AoCLO2 contained motifs 10, 6, 1, 5, 2, 3, and 4. AoCLO4 presented four motifs: 6, 1, 5, and 2. AoCLO5 and AoCLO6 shared an arrangement consisting of motifs 8, 7, 1, 5, 2, and 3. Motif 9 was restricted to AoCLO1 and AoCLO3, whereas motif 10 occurred only in AoCLO2. Motifs 7 and 8 were detected exclusively in AoCLO5 and AoCLO6.

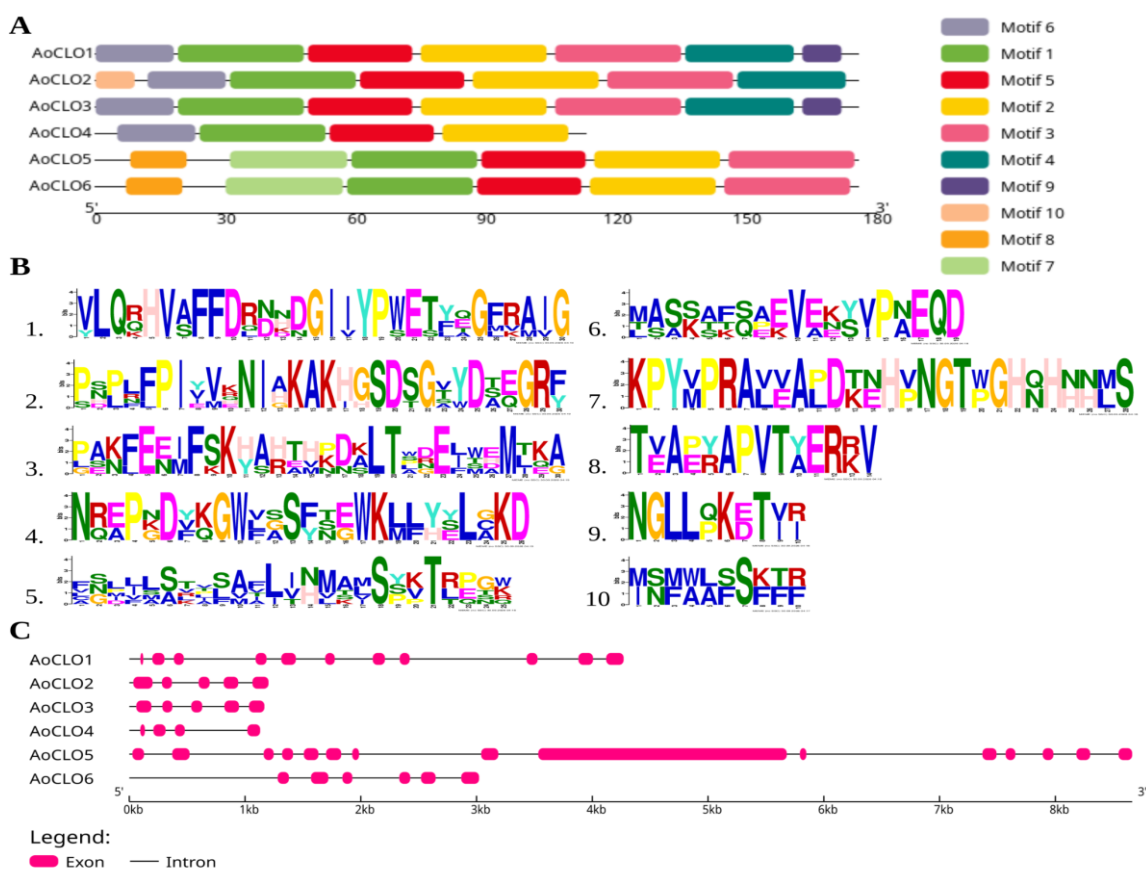
The exon–intron representations showed differences in gene length and organization among the six candidates (Figure 2C). AoCLO2 and AoCLO3 exhibited similar structural patterns, whereas AoCLO4 presented a shorter genomic span and fewer displayed exonic segments. AoCLO5 occupied the largest genomic span, exceeding 8 kb, and contained a comparatively long exon in the central portion of the gene. AoCLO1 extended approximately 4 kb, while the displayed AoCLO6 structure was located within approximately 3 kb.

The alignment and motif analyses revealed a shared sequence framework together with differences among AoCLO members. Motifs 1, 5, and 2 occurred in all six candidates, whereas other motifs were restricted to subsets of proteins. The matching arrangements of AoCLO1 and AoCLO3, and of AoCLO5 and AoCLO6, provide evidence of similarities within these pairs. Conserved motif organization has also been reported in caleosins from *Carya cathayensis*, supporting the usefulness of this approach for describing family architecture (Cao *et al.*, 2024). However, motif numbers are specific to each MEME analysis: similarly numbered motifs in different studies are not necessarily

equivalent, and differences in the number detected may reflect both sequence composition and analytical settings.

The differences in exon–intron organization add another level of variation to the AoCLO family. Similar structures in AoCLO2 and AoCLO3 are consistent with their shared sequence features, whereas the larger genomic span of AoCLO5 contrasts with its protein length of 176 amino acids. This relationship highlights the importance of distinguishing genomic span from coding sequence length. Before attributing the long displayed exonic region to biological divergence, its correspondence with coding and untranslated regions should be checked. Gene structure and motif similarity can guide comparative interpretation, but neither establishes the mechanism or timing of gene duplication without additional genomic evidence.

Figure 2. Conserved motifs and gene structures of the AoCLO family. (A) Distribution of conserved motifs in the protein sequences. (B) Sequence logos of the ten identified motifs. (C) Exon–intron organization, with exons represented by pink blocks and introns by black lines.

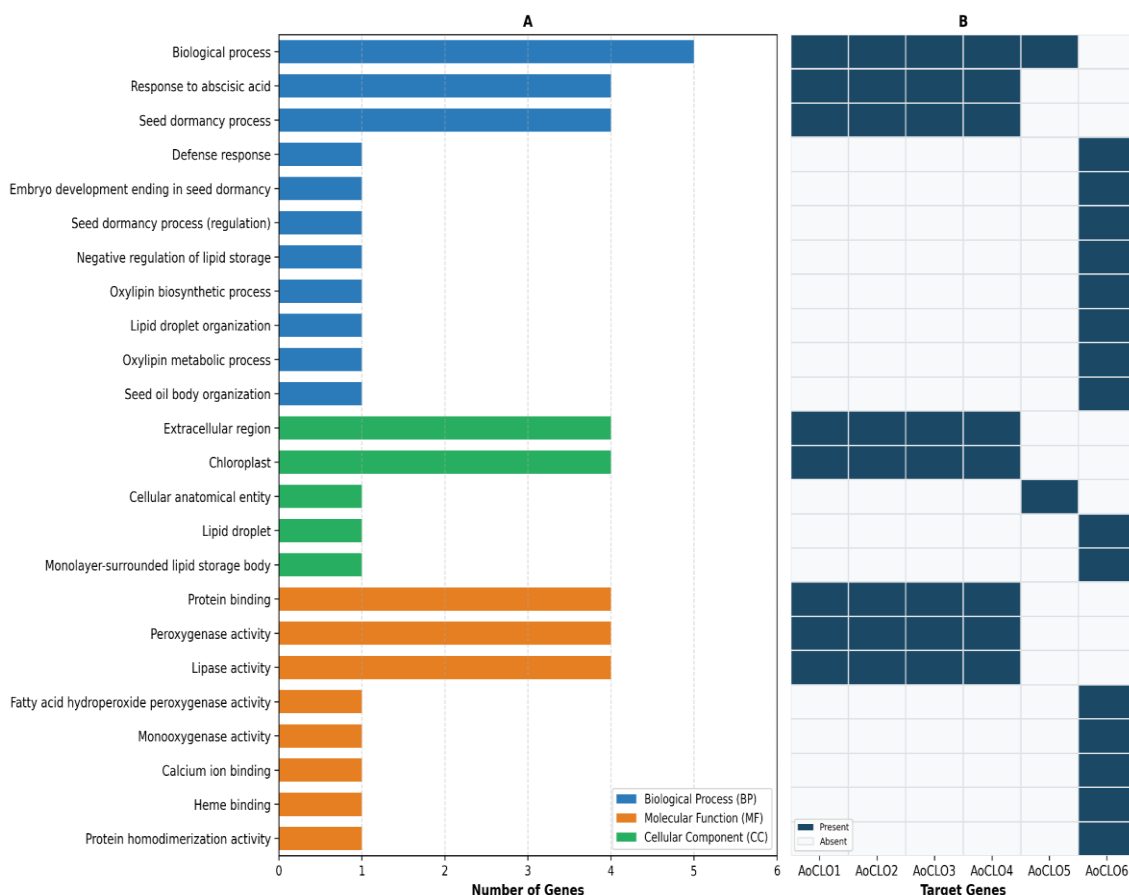


- **Functional annotation**

Gene Ontology annotations were distributed among biological process, cellular component, and molecular function categories (Figure 3). Within biological processes, response to abscisic acid and seed dormancy process were assigned to AoCLO1–AoCLO4. AoCLO6 received annotations related to defense response, embryo development ending in seed dormancy, regulation of seed dormancy, negative regulation of lipid storage, oxylipin biosynthesis and metabolism, lipid droplet organization, and seed oil body organization. Within the cellular component category, extracellular region and chloroplast were assigned to AoCLO1–AoCLO4. AoCLO5 received the general annotation cellular anatomical entity, whereas AoCLO6 was associated with lipid droplet and monolayer-surrounded lipid storage body.

Within molecular functions, protein binding, peroxygenase activity, and lipase activity were assigned to AoCLO1–AoCLO4. AoCLO6 received annotations for fatty acid hydroperoxide peroxygenase activity, monooxygenase activity, calcium ion binding, heme binding, and protein homodimerization activity. The eggNOG-mapper output also included the KEGG Orthology identifier K17991 and the KEGG pathway identifier 00073.

Figure 3. Gene Ontology annotations of AoCLO proteins. (A) Number of candidates assigned to each annotation, grouped into biological process, cellular component, and molecular function categories. (B) Presence–absence matrix showing the distribution of annotations among individual AoCLO candidates.



The biological process annotations associated with abscisic acid responses and seed dormancy suggest candidate links between AoCLO proteins and developmental or environmental signaling. Experimental evidence from *Arabidopsis* provides a relevant context: Brunetti *et al.* (2021) demonstrated that RD20/CLO3 interacts with the heterotrimeric G-protein alpha subunit GPA1 and acts as a negative regulator of this component. They also reported that RD20/CLO3 is induced by drought, salt, and abscisic acid. These findings support investigating AoCLO expression under comparable treatments, but they do not establish that the cashew candidates perform the same signaling functions or enhance stress tolerance.

Annotations related to lipid storage, seed development, and lipid droplet organization are also consistent with experimentally established functions of some plant caleosins. Liu *et al.* (2022) demonstrated overlapping roles of multiple *Arabidopsis* caleosins in oil accumulation and embryo development. Their results show that functional redundancy can occur within this family, which is relevant when interpreting the shared

motifs and annotations observed among AoCLO proteins. In addition, Miklaszewska *et al.* (2023) showed that CALEOSIN 1 interacts with ATG8 and facilitates lipid droplet microautophagy in seedlings. Together, these studies identify reserve accumulation and mobilization as useful directions for future investigation in cashew, without establishing equivalent activities for individual AoCLO members.

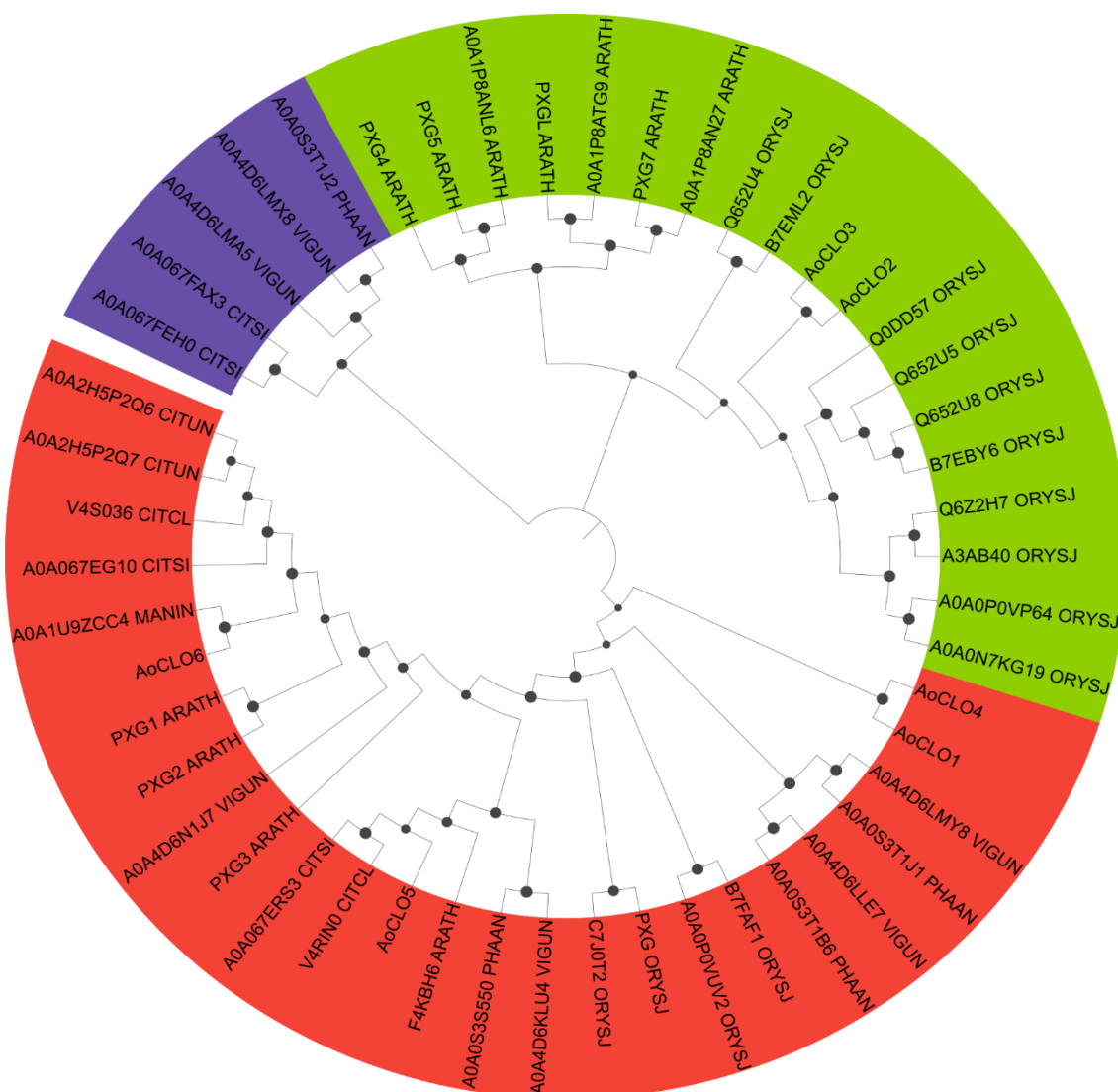
The annotation profile of AoCLO6 was broader than those of the other candidates, including terms associated with oxylipin metabolism, calcium binding, heme binding, and lipid droplets. This pattern may reflect the functional information available for its inferred homologs rather than greater biological multifunctionality. Likewise, the lipase and monooxygenase annotations require individual evaluation and should not be presented as measured enzymatic activities. The KEGG assignment K17991 corresponds to peroxygenase, whereas pathway 00073 represents cutin, suberine, and wax biosynthesis. These assignments suggest a connection with lipid-modifying reactions, but membership in a reference pathway does not demonstrate that the candidates participate in all processes represented in that pathway.

• **Phylogenetic analysis**

The phylogenetic representation included the six AoCLO proteins and caleosin sequences from other plant species (Figure 4). Three groups were highlighted in green, red, and purple. AoCLO2 and AoCLO3 were assigned to the green group, together with reference sequences from *Arabidopsis thaliana* and *Oryza sativa*. AoCLO1, AoCLO4, AoCLO5, and AoCLO6 were assigned to the red group, which also included reference sequences from *Arabidopsis*, *Oryza*, *Phaseolus*, *Vigna*, *Citrus*, and *Mangifera*. No AoCLO protein was assigned to the purple group.

The distribution of AoCLO proteins between two of the highlighted phylogenetic groups indicates sequence heterogeneity within the cashew family. This observation is compatible with the diversification described in broader caleosin analyses (Rahman *et al.*, 2018). However, the groups highlighted here should not automatically be equated with the H and L isoform classes described in that study. Such classification requires examination of the diagnostic sequence features and supported relationships with characterized reference proteins. Similarly, the placement of AoCLO proteins alongside sequences from other species provides comparative context but does not, by itself, establish identical biological functions.

Figure 4. Phylogenetic representation of AoCLO proteins and reference caleosin sequences from other plant species. Colored sectors indicate the three groups highlighted in the analysis.



- **Three-dimensional structural analysis**

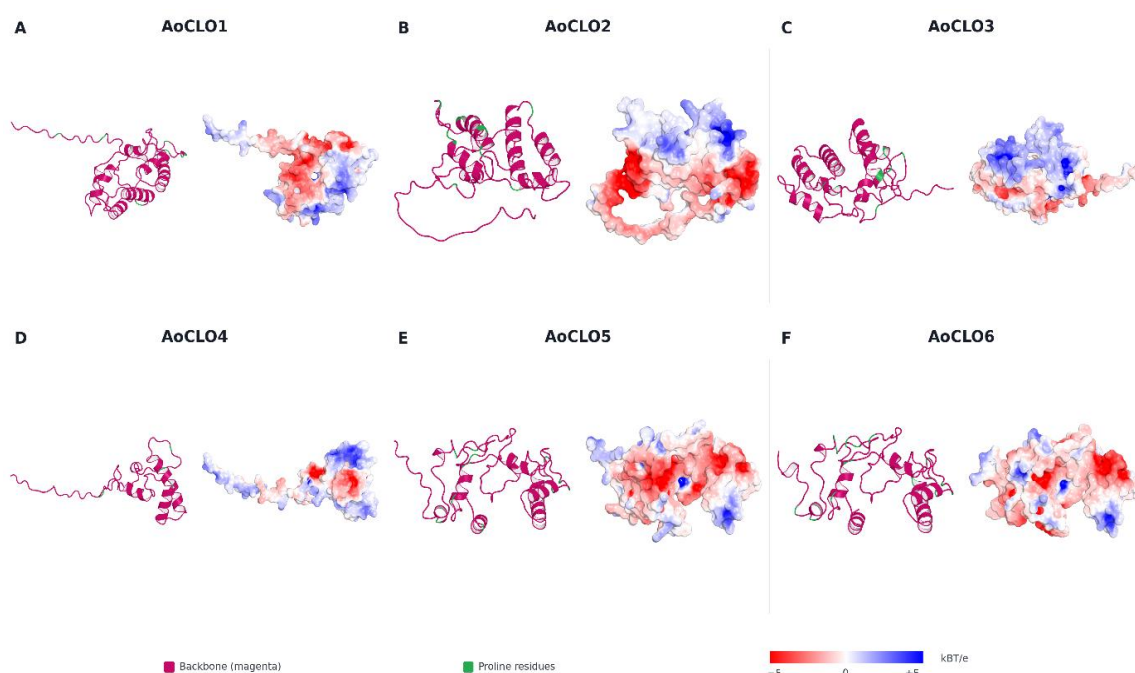
The three-dimensional models of AoCLO1–AoCLO6 exhibited differences in their overall conformation, including the arrangement of α -helices and connecting loop regions (Figure 5). Helical elements were visible in all six models, whereas no prominent β -sheet structures were observed in the representations displayed.

AoCLO1, AoCLO2, and AoCLO3 presented a folded region containing several α -helices, accompanied by extended terminal segments. AoCLO4, the shortest protein in the dataset, exhibited fewer visible helical elements and an extended segment projecting from the folded region. AoCLO5 and AoCLO6 displayed multiple α -helices interconnected by loops, with differences in the orientation and extension of peripheral segments.

Proline residues, highlighted in green, were distributed across different positions in the models, including loop regions and positions adjacent to helical elements. Molecular surface representations showed differences in overall shape and surface contours among the six proteins, with protrusions corresponding to extended segments of the structural models.

Structural validation revealed differences in stereochemical quality and global model quality among the six AoCLO proteins. Ramachandran analysis showed variation in the distribution of residues across favored, allowed, and outlier regions, while ProSA-web and QMEANDisCo provided complementary assessments of the predicted structures. Electrostatic surface representations displayed regions of negative and positive potential in all six models under a common visualization scale (Figure 5).

Figure 5. Structural and electrostatic representations of AoCLO proteins. Panels A–F correspond to AoCLO1–AoCLO6, respectively. Each panel shows a cartoon representation, with the backbone in magenta and proline residues highlighted in green, alongside an electrostatic molecular surface calculated using APBS. Surface potentials are displayed on a common scale from -5 to $+5$ kBT/e, with red indicating negative potential, white indicating values near zero, and blue indicating positive potential.



Structural validation indicated heterogeneous quality among the AoCLO models. AoCLO5 and AoCLO6 exhibited the highest proportions of residues in favored Ramachandran regions, whereas AoCLO4 showed the largest proportion of outliers. However, favorable backbone stereochemistry did not consistently coincide with higher QMEANDisCo scores. The global scores therefore support cautious interpretation of the predicted structures, particularly for comparisons involving flexible regions or specific residue contacts (Williams *et al.*, 2018; Studer *et al.*, 2020). The electrostatic maps provide an exploratory comparison of surface potential under a common computational protocol, but they do not establish ligand affinity, membrane association, or biological activity.

Taken together, the analyses identify shared sequence features and distinguishable characteristics among the six AoCLO candidates. These findings support targeted investigations rather than definitive functional assignments. Confirming sequence completeness, examining expression during seed development and stress treatments, testing subcellular localization, and measuring peroxygenase activity would connect the

computational observations to biological mechanisms. Such experiments would clarify whether the differences observed among AoCLO members correspond to redundancy, distinct expression patterns, or specialization in cashew lipid metabolism and signaling.

Conclusion

This study identified six candidate caleosin proteins in *A. occidentale* and characterized their sequence features, physicochemical properties, gene organization, predicted functional associations, and structural models. Shared motifs revealed conserved features across the candidates, while differences in motif composition, gene structure, and predicted conformation distinguished individual members. These findings provide a comparative basis for investigating caleosins in cashew and selecting candidates for experimental studies of lipid metabolism, development, and environmental responses. Functional assignments and subcellular localization remain predictions requiring validation.

Acknowledgment

The authors acknowledge the use of ChatGPT (OpenAI) for assistance with manuscript organization, English language editing, and preparation of scripts for structural visualization in PyMOL. All AI-assisted content and outputs were critically reviewed and verified by the authors, who take full responsibility for the scientific accuracy, interpretation, and final content of the manuscript.

References

- BROOKS, B. R. *et al.* CHARMM: The biomolecular simulation program. **Journal of Computational Chemistry**, v. 30, p. 1545-1614, 2009. DOI: <https://doi.org/10.1002/jcc.21287>
- BRUNETTI, S. C.; ARSENEAULT, M. K. M.; WRIGHT, J. A.; WANG, Z.; EHDAEIVAND, M.-R.; LOWDEN, M. J.; RIVOAL, J.; KHALIL, H. B.; GARG, G.; GULICK, P. J. The stress induced caleosin, RD20/CLO3, acts as a negative regulator of GPA1 in Arabidopsis. **Plant Molecular Biology**, v. 107, p. 159-175, 2021. DOI: <https://doi.org/10.1007/s11103-021-01189-x>
- CAO, Y.; ZHAN, Y.; LIU, J.; TANG, T.; LI, J.; ZHAO, R.; ZHANG, Q.; HU, S.; CAO, W.; GAO, Y. Identification and Analysis of Expression Patterns of the Caleosin Genes in Hickory (*Carya cathayensis* Sarg.). **Forests**, v. 15, artigo 609, 2024. DOI: <https://doi.org/10.3390/f15040609>

- HANANO, A.; BLÉE, E.; MURPHY, D. J. Caleosin/peroxygenases: multifunctional proteins in plants. **Annals of Botany**, v. 131, p. 387-409, 2023. DOI: <https://doi.org/10.1093/aob/mcad001>
- JURRUS, E. *et al.* Improvements to the APBS biomolecular solvation software suite. **Protein Science**, v. 27, p. 112-128, 2018. DOI: <https://doi.org/10.1002/pro.3280>.
- KUMAR, S.; STECHER, G.; TAMURA, K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. **Molecular Biology and Evolution**, v. 33, p. 1870-1874, 2016. DOI: <https://doi.org/10.1093/molbev/msw054>
- LETUNIC, I.; BORK, P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. **Nucleic Acids Research**, v. 44, p. W242-W245, 2016. DOI: <https://doi.org/10.1093/nar/gkw290>
- LIU, X.; YANG, Z.; WANG, Y.; SHEN, Y.; JIA, Q.; ZHAO, C.; ZHANG, M. Multiple caleosins have overlapping functions in oil accumulation and embryo development. **Journal of Experimental Botany**, v. 73, p. 3946-3962, 2022. DOI: <https://doi.org/10.1093/jxb/erac153>
- MIKLASZEWSKA, M.; ZIENKIEWICZ, K.; KLUGIER-BOROWSKA, E.; RYGIELSKI, M.; FEUSSNER, I.; ZIENKIEWICZ, A. CALEOSIN 1 interaction with AUTOPHAGY-RELATED PROTEIN 8 facilitates lipid droplet microautophagy in seedlings. **Plant Physiology**, v. 193, p. 2361-2380, 2023. DOI: <https://doi.org/10.1093/plphys/kiad471>
- RAHMAN, F.; HASSAN, M.; ROSLI, R.; ALMOUSALLY, I.; HANANO, A.; MURPHY, D. J. Evolutionary and genomic analysis of the caleosin/peroxygenase (CLO/PXG) gene/protein families in the Viridiplantae. **PLOS ONE**, v. 13, artigo e0196669, 2018. DOI: <https://doi.org/10.1371/journal.pone.0196669>
- SAVADI, S.; MURALIDHARA, B. M.; GODWIN, J.; ADIGA, J. D.; MOHANA, G. S.; ERADASAPPA, E.; SHAMSUDHEEN, M.; KARUN, A. De novo assembly and characterization of the draft genome of the cashew (*Anacardium occidentale* L.). **Scientific Reports**, v. 12, artigo 18187, 2022. DOI: <https://doi.org/10.1038/s41598-022-22600-7>
- STUDER, G.; REMPFER, C.; WATERHOUSE, A. M.; GUMIENNY, R.; HAAS, J.; SCHWEDE, T. QMEANDisCo—distance constraints applied on model quality estimation. **Bioinformatics**, v. 36, p. 1765-1771, 2020. DOI: <https://doi.org/10.1093/bioinformatics/btz828>
- VARADI, M. *et al.* AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. **Nucleic Acids Research**, v. 50, p. D439-D444, 2022. DOI: <https://doi.org/10.1093/nar/gkab1061>
- WANG, J. *et al.* The conserved domain database in 2023. **Nucleic Acids Research**, v. 51, p. D384-D388, 2023. DOI: <https://doi.org/10.1093/nar/gkac1096>

WIEDERSTEIN, M.; SIPPL, M. J. ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. **Nucleic Acids Research**, v. 35, p. W407-W410, 2007. DOI: <https://doi.org/10.1093/nar/gkm290>

WILLIAMS, C. J. *et al.* MolProbity: More and better reference data for improved all-atom structure validation. **Protein Science**, v. 27, p. 293-315, 2018. DOI:<https://doi.org/10.1002/pro.3330>.

CAPÍTULO 6

BIOINFORMÁTICA: A BÚSSOLA PARA VACINAS PERSONALIZADAS CONTRA O MELANOMA

BIOINFORMATICS: THE COMPASS FOR PERSONALIZED VACCINES AGAINST MELANOMA

DOI: <https://doi.org/10.56001/26.9786501906539.06>

Submetido em: 19/10/2026

Revisado em: 19/10/2026

Publicado em: 24/10/2026

Maria Cidinaria Silva Alves

Laboratório de Parasitologia Molecular, Faculdade de Medicina de Ribeirão Preto,
Universidade de São Paulo, Ribeirão Preto, São Paulo, Brasil

<http://lattes.cnpq.br/1633031262017450>

Rousilândia de Araujo Silva

Programa de Pós-Graduação em Ciências Farmacêuticas, Universidade Federal de
Pernambuco, Recife, Pernambuco, Brasil

<http://lattes.cnpq.br/8194570266422228>

Jakeline Moreira da Silva

Escola Mínima Antônio de Moraes Andrade, Macaparana, Pernambuco, Brasil

<http://lattes.cnpq.br/7884741057259883>

José Lucas de Araújo

Escola Municipal em Tempo Integral Severino Francisco da Silva
Macaparana, Pernambuco, Brasil

<http://lattes.cnpq.br/3760156602428318>

José Cláudio Gomes de Araújo

Programa de Pós-Graduação em Ciência e Engenharia de Materiais, Universidade
Federal da Paraíba, João Pessoa, Paraíba, Brasil

<http://lattes.cnpq.br/0757182882367448>

José Eduardo Souza Echeverria

Programa de Pós-Graduação em Ciências da Saúde, Universidade Federal da Grande
Dourados, Dourados, Mato Grosso do Sul, Brasil

<http://lattes.cnpq.br/1188506108074331>

Gleyce Hellen de Almeida de Souza

Programa de Pós-Graduação em Ciências da Saúde, Universidade Federal da Grande
Dourados, Dourados, Mato Grosso do Sul, Brasil

<http://lattes.cnpq.br/8163519978837401>

Fabíola Lucini

Programa de Pós-Graduação em Ciências da Saúde, Universidade Federal da Grande
Dourados, Dourados, Mato Grosso do Sul, Brasil

<http://lattes.cnpq.br/1608714419112836>

Ruana Carolina Cabral da Silva

Programa de Pós-Graduação em Ciências da Saúde, Universidade Federal da Grande
Dourados, Mato Grosso do Sul, Brasil

<http://lattes.cnpq.br/8634356371128001>

Resumo

O desenvolvimento de vacinas personalizadas contra o melanoma tem sido impulsionado pelos avanços da bioinformática e das tecnologias de sequenciamento genético. Este capítulo tem como objetivo discutir a contribuição das ferramentas bioinformáticas para a identificação e seleção de candidatos a neoantígenos tumorais e sua aplicação no desenvolvimento de vacinas terapêuticas personalizadas. São abordadas as principais etapas desse processo, incluindo a identificação de mutações somáticas por sequenciamento de nova geração, a análise da expressão gênica, a predição computacional da afinidade de ligação de peptídeos às moléculas do complexo principal de histocompatibilidade (MHC) e a avaliação de sua potencial imunogenicidade. Também são discutidas as aplicações da imunopeptidômica na validação de candidatos e as plataformas utilizadas na formulação vacinal, como peptídeos sintéticos e RNA mensageiro. Além disso, são apresentadas perspectivas relacionadas à inteligência artificial, à análise de células individuais e à associação dessas vacinas com outras estratégias imunoterapêuticas. Apesar dos avanços, persistem desafios relacionados à heterogeneidade tumoral, à precisão das predições computacionais e à validação clínica. Conclui-se que a integração entre bioinformática, imunologia e tecnologias moleculares constitui uma abordagem relevante para o desenvolvimento de vacinas personalizadas contra o melanoma.

Palavras-Chave: Bioinformática; Melanoma; Vacinas personalizadas; Neoantígenos; Imunoterapia.

Abstract

The development of personalized vaccines against melanoma has been driven by advances in bioinformatics and genetic sequencing technologies. This chapter aims to discuss the contribution of bioinformatics tools to the identification and selection of candidate tumor neoantigens and their application in the development of personalized therapeutic vaccines. The main stages of this process are addressed, including the identification of somatic mutations through next-generation sequencing, gene expression analysis, computational prediction of peptide-binding affinity to major histocompatibility complex (MHC) molecules, and assessment of their potential immunogenicity. The applications of immunopeptidomics in candidate validation and the platforms used for vaccine formulation, such as synthetic peptides and messenger RNA, are also discussed. In addition, perspectives related to artificial intelligence, single-cell

analysis, and the combination of these vaccines with other immunotherapeutic strategies are presented. Despite these advances, challenges remain regarding tumor heterogeneity, the accuracy of computational predictions, and clinical validation. It is concluded that the integration of bioinformatics, immunology, and molecular technologies represents a relevant approach to the development of personalized vaccines against melanoma.

Keywords: Bioinformatics; Melanoma; Personalized vaccines; Neoantigens; Immunotherapy.

Introdução

O melanoma representa a forma mais agressiva de câncer de pele, caracterizando-se por uma capacidade notória de sofrer mutações genéticas devido à exposição prolongada à radiação ultravioleta (Moni *et al.*, 2025). Historicamente, o tratamento baseava-se em intervenções cirúrgicas e quimioterapia citotóxica, com resultados limitados em estágios avançados. No entanto, a compreensão do melanoma como um tumor altamente imunogênico mudou o paradigma terapêutico, colocando o sistema imunológico no centro da estratégia de combate à doença (Knight *et al.*, 2023).

A eficácia dessa resposta imune depende da capacidade das células T em distinguir as células cancerosas das saudáveis, um processo mediado pelo reconhecimento de antígenos específicos na superfície tumoral. É aqui que os neoantígenos, derivados de alterações genéticas específicas das células tumorais, surgem como alvos relevantes para a imunoterapia, pois sua expressão é predominantemente restrita às células tumorais, reduzindo o risco de reconhecimento de tecidos normais. A identificação desses alvos, contudo, é um desafio de proporções monumentais devido à vasta heterogeneidade genômica entre diferentes pacientes (Silva *et al.*, 2021; Zhang *et al.*, 2021).

Neste contexto, a bioinformática emerge como uma ferramenta indispensável, atuando como a ponte entre o sequenciamento genético e a aplicação clínica. Através de algoritmos avançados e modelos computacionais, é possível filtrar milhares de mutações e prever quais delas resultarão em epítomos capazes de desencadear uma resposta imune robusta. Este artigo explora como a integração da bioinformática e da imunologia está viabilizando a criação de vacinas personalizadas, transformando o tratamento do melanoma em uma medicina de precisão absoluta.

Desenvolvimento

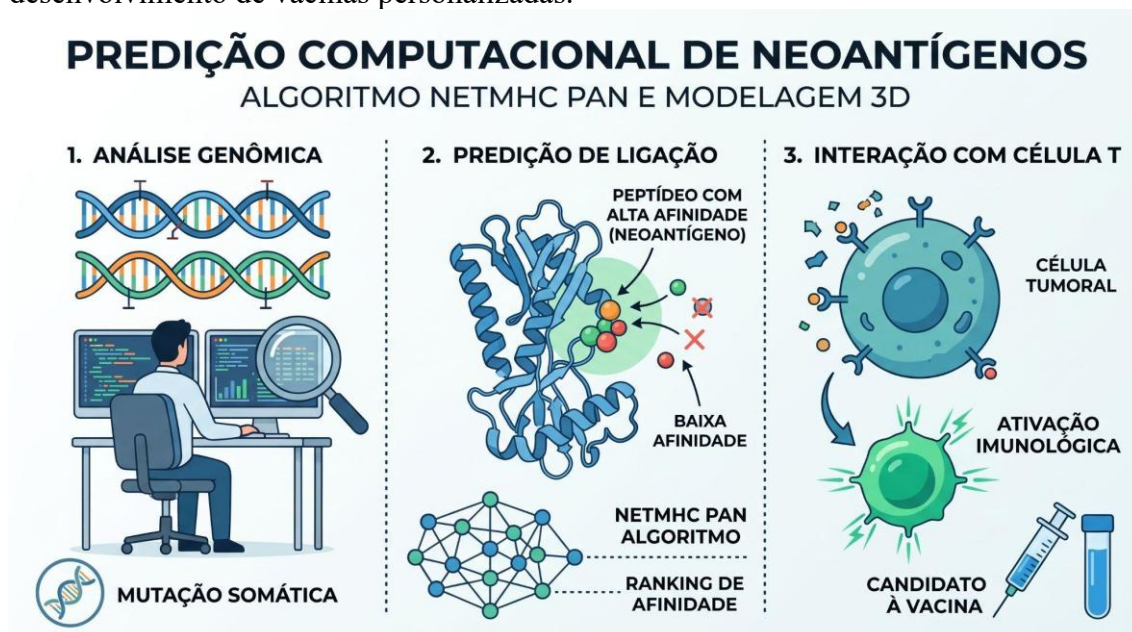
Identificação de Neoantígenos via Sequenciamento de Nova Geração (NGS)

O primeiro passo na construção de uma vacina personalizada é o mapeamento completo do genoma do paciente, utilizando o Sequenciamento de Nova Geração (NGS) para comparar o DNA tumoral com o DNA de células saudáveis. Esse processo gera uma quantidade massiva de dados brutos que precisam ser processados por *pipelines* bioinformáticas para identificar as chamadas variantes de nucleotídeo único (SNVs) e inserções/deleções (*indels*). A bioinformática permite que os pesquisadores foquem apenas nas mutações que alteram a sequência de aminoácidos das proteínas, descartando mutações silenciosas, conforme a Figura 1.

Além do sequenciamento de DNA, a análise de RNA-seq é crucial para confirmar se o gene mutado está, de fato, sendo expresso pela célula tumoral em níveis suficientes. Não basta que a mutação exista no DNA; ela deve ser transcrita e traduzida para que a proteína resultante possa ser processada e apresentada ao sistema imune. Ferramentas computacionais de quantificação de transcritos ajudam a priorizar candidatos que tenham alta abundância intracelular, aumentando as chances de reconhecimento.

A complexidade aumenta quando consideramos a heterogeneidade intratumoral, onde diferentes áreas do mesmo tumor podem apresentar mutações distintas. A bioinformática utiliza algoritmos de inferência clonal para identificar mutações "tronco", aquelas presentes em todas as células do tumor, que são os alvos mais seguros para a vacinação. Sem o poder de processamento computacional, seria impossível distinguir essas variantes fundamentais do ruído biológico, tornando a vacina ineficaz contra a massa tumoral total.

Figura 1. Fluxo de trabalho para predição computacional de neoantígenos e desenvolvimento de vacinas personalizadas.



Legenda: 1. Análise Genômica: Identificação de mutações somáticas a partir do sequenciamento e comparação entre o genoma tumoral e o tecido normal via ferramentas bioinformáticas. 2. Predição de Ligação: Modelagem tridimensional da fenda do MHC e aplicação de algoritmos de aprendizado de máquina (como o NetMHCpan) para ranquear e selecionar os peptídeos mutados com alta afinidade de ligação em detrimento dos de baixa afinidade. 3. Interação com Célula T: Neoantígenos validados são selecionados como candidatos à formulação vacinal, promovendo a ativação imunológica de linfócitos T específicos para o reconhecimento e eliminação direcionada das células tumorais. **Fonte:** Elaborada pelos autores (2026).

Predição Computacional de Afinidade ao MHC

Uma vez identificados os potenciais neoantígenos, o maior desafio imunológico é prever quais deles serão apresentados pelas moléculas do Complexo de Histocompatibilidade Principal (MHC). No ser humano, o sistema HLA (MHC humano) é extremamente polimórfico, o que significa que cada paciente possui um conjunto único de "suportes" para apresentar antígenos. A bioinformática utiliza redes neurais artificiais, treinadas em vastos bancos de dados de ligação peptídica, para calcular a afinidade de cada peptídeo mutado com o alelo específico do paciente (Shen *et al.*, 2022).

Algoritmos como o NetMHCpan revolucionaram essa etapa ao prever com alta precisão a estabilidade do complexo peptídeo-MHC (Nielsen *et al.*, 2007; Xie *et al.*, 2023). Esses modelos levam em conta as propriedades físico-químicas dos aminoácidos e as forças de ligação dentro da fenda do MHC, algo que levaria anos para ser testado manualmente em laboratório. A predição computacional reduz a lista de milhares de candidatos a apenas uma dezena de peptídeos com alta probabilidade de serem exibidos na superfície da célula.

Entretanto, a ligação ao MHC é apenas metade da batalha; o peptídeo também precisa ser reconhecido pelo receptor de células T (TCR) (Lang *et al.*, 2022; Szeto *et al.*, 2021). A bioinformática estrutural e a modelagem molecular ajudam a prever a imunogenicidade, ou seja, se a conformação do peptídeo no MHC será "atraente" o suficiente para ativar um linfócito T (Richters *et al.*, 2019). Esse refinamento algorítmico é o que permite que as vacinas de melanoma sejam desenhadas com alta especificidade, focando nos alvos que realmente provocam uma expansão clonal de células de defesa (Xie *et al.*, 2023).

O Design da Vacina: da Predição à Plataforma Terapêutica

Após a seleção bioinformática dos melhores epítomos, o foco muda para a engenharia da vacina, que pode ser entregue via peptídeos sintéticos, DNA ou, mais recentemente, RNA mensageiro (mRNA). A bioinformática auxilia no design da sequência da vacina, otimizando os códons para garantir a máxima expressão proteica após a administração no paciente. Além disso, é necessário prever a estabilidade da molécula e a sua interação com adjuvantes que potencializam a resposta inflamatória necessária (Richters *et al.*, 2019).

A integração de dados de imunopeptidômica, obtidos por meio da análise de peptídeos associados às moléculas de MHC por espectrometria de massa, contribui para a validação das predições computacionais iniciais (Estephan *et al.*, 2025). Esse ciclo de retroalimentação entre a bioinformática e a experimentação direta permite o aprimoramento constante dos modelos de predição. O resultado é um produto biotecnológico único para cada indivíduo, projetado para educar o sistema imune a reconhecer a "assinatura" específica do seu próprio melanoma (Xie *et al.*, 2023).

A tabela abaixo resume as etapas principais deste fluxo de trabalho integrado:

Tabela 1. Fluxo de trabalho para o desenvolvimento de vacinas personalizadas.

Etapa	Ferramenta/Técnica	Objetivo Imunológico
Sequenciamento	NGS (WES + RNA-seq)	Identificar mutações exclusivas e níveis de expressão.
Filtro Bioinformático	Chamada de Variantes (GATK)	Distinguir mutações somáticas de germinativas.
Predição de Ligação	NetMHCpan / IEDB	Selecionar peptídeos que se ligam ao HLA do paciente.
Validação	Espectrometria de Massa	Confirmar que o neoantígeno é apresentado na célula.
Formulações	Síntese de mRNA / Peptídeos	Criar o imunógeno para ativação de Células T.

Fonte: Elaborada pelos autores (2026).

Perspectivas Futuras

O futuro das vacinas de melanoma reside na integração de Inteligência Artificial ainda mais sofisticada, capaz de prever não apenas a ligação ao MHC, mas a dinâmica de escape do tumor em tempo real. Acredita-se que, em breve, seremos capazes de realizar o sequenciamento e o design da vacina em questão de dias, tornando o tratamento viável mesmo para pacientes com doença rapidamente progressiva. A combinação de vacinas personalizadas com inibidores de *checkpoint* (como anti-PD-1) promete aumentar drasticamente as taxas de cura (Boussiotis, 2016; Knight *et al.*, 2023).

Além disso, a análise de célula única (*single-cell*) começará a desempenhar um papel crucial, permitindo que a bioinformática identifique neoantígenos em subpopulações raras de células-tronco tumorais que sobrevivem às terapias convencionais. Isso permitirá a criação de vacinas "polivalentes" que atacam o tumor de múltiplos ângulos genéticos simultaneamente (Han *et al.*, 2014; Lu *et al.*, 2018). O custo das tecnologias de sequenciamento continua a cair, o que é fundamental para a democratização desse acesso, permitindo que a medicina personalizada deixe de ser um luxo de centros de pesquisa e chegue à rotina oncológica.

Por fim, a expansão dessas tecnologias para outros tipos de câncer, além do melanoma, dependerá da nossa capacidade computacional de entender ambientes tumorais menos imunogênicos. A bioinformática terá o papel de encontrar formas de

"aquecer" tumores frios, possivelmente através da identificação de neoantígenos compartilhados ou modificações epigenéticas. O horizonte é promissor: a transição de um tratamento de "tamanho único" para uma terapia onde o código genético do próprio tumor dita a cura.

Considerações Finais

A convergência entre a bioinformática e a imunologia representa um dos marcos mais significativos da oncologia moderna, especialmente no combate ao melanoma. A capacidade de traduzir a complexidade genômica em alvos terapêuticos precisos transformou a percepção de que o câncer é um inimigo invencível e indetectável. As vacinas de neoantígenos não são apenas uma promessa; elas são a materialização do potencial de processamento de dados servindo diretamente à sobrevivência humana.

Apesar dos avanços, a jornada ainda enfrenta desafios técnicos e regulatórios, exigindo uma colaboração interdisciplinar constante entre bioinformatas, biólogos moleculares e oncologistas. A padronização dos *pipelines* de análise e a validação clínica de larga escala são os próximos passos essenciais para consolidar essa tecnologia. A bioinformática provou ser muito mais do que um suporte estatístico; ela é o cérebro por trás da nova geração de imunoterapias, garantindo que o sistema imune tenha o mapa correto para sua missão.

Em última análise, o melanoma serviu como o laboratório ideal para essa revolução tecnológica, e as lições aprendidas aqui estão pavimentando o caminho para o tratamento de diversos outros tumores sólidos. A vacina personalizada simboliza o ápice da medicina de precisão, onde a singularidade biológica de cada paciente é respeitada e utilizada como a ferramenta principal para a sua própria recuperação. O futuro do tratamento oncológico está escrito em código, mas seu resultado é medido em vidas salvas.

Declaração de Uso de Inteligência Artificial Generativa

Durante a elaboração deste capítulo, utilizou-se o ChatGPT (OpenAI) como ferramenta auxiliar na organização e redação preliminar do texto, na revisão linguística e na elaboração da figura ilustrativa do fluxo de identificação de neoantígenos e desenvolvimento de vacinas personalizadas. Os conteúdos gerados foram submetidos à análise crítica e à revisão dos autores, incluindo a verificação das informações e das referências bibliográficas. A interpretação dos estudos, as decisões editoriais e a

aprovação da versão final permaneceram sob responsabilidade dos autores, que assumem integral responsabilidade pela precisão, originalidade e integridade científica do trabalho.

Referências

- BOUSSIOTIS, V. A. Molecular and Biochemical Aspects of the PD-1 Checkpoint Pathway. **The New England Journal of Medicine**, v. 375, p. 1767-1778, 2016. DOI: 10.1056/NEJMr1514296.
- ESTEPHAN, H.; HAMMOND, E. M.; ADAMOPOULOU, E. Immunopeptidomics Workflow for Isolation and LC-MS/MS Analysis of MHC Class I-Bound Peptides Under Hypoxic Conditions. **Bio-protocol**, v. 15, e5505, 2025. DOI: 10.21769/BioProtoc.5505.
- HAN, A.; GLANVILLE, J.; HANSMANN, L.; DAVIS, M. M. Linking T-cell receptor sequence to functional phenotype at the single-cell level. **Nature Biotechnology**, v. 32, p. 684-692, 2014. DOI: 10.1038/nbt.2938.
- KNIGHT, A.; KARAPETYAN, L.; KIRKWOOD, J. M. Immunotherapy in Melanoma: Recent Advances and Future Directions. **Cancers**, v. 15, art. 1106, 2023. DOI: 10.3390/cancers15041106.
- LANG, F. *et al.* Identification of neoantigens for individualized therapeutic cancer vaccines. **Nature Reviews Drug Discovery**, v. 21, p. 261-282, 2022. DOI: 10.1038/s41573-021-00387-y.
- LU, Y.-C. *et al.* An Efficient Single-Cell RNA-Seq Approach to Identify Neoantigen-Specific T Cell Receptors. **Molecular Therapy**, v. 26, p. 379-389, 2018. DOI: 10.1016/j.ymthe.2017.10.018.
- MONI, S. S. *et al.* Advances in Materials Science for Precision Melanoma Therapy: Nanotechnology-Enhanced Drug Delivery Systems. **Pharmaceutics**, v. 17, art. 296, 2025. DOI: 10.3390/pharmaceutics17030296.
- NIELSEN, M. *et al.* NetMHCpan, a Method for Quantitative Predictions of Peptide Binding to Any HLA-A and -B Locus Protein of Known Sequence. **PLOS ONE**, v. 2, e796, 2007. DOI: 10.1371/journal.pone.0000796.
- RICHTERS, M. M. *et al.* Best practices for bioinformatic characterization of neoantigens for clinical utility. **Genome Medicine**, v. 11, art. 56, 2019. DOI: 10.1186/s13073-019-0666-2.
- SHEN, Y. *et al.* Neoantigen vaccine and neoantigen-specific cell adoptive transfer therapy in solid tumors: Challenges and future directions. **Cancer Innovation**, v. 1, p. 168-182, 2022. DOI: 10.1002/cai2.26.
- SILVA, E. V. de S.; NASCENTE, E. de P.; MIGUEL, M. P.; ALVES, C. E. F.; MOURA, V. M. B. D. de. Elucidando a imunovigilância e imunoedição tumoral: uma revisão abrangente. **Ciência Animal Brasileira**, v. 22, 2021. DOI: 10.1590/1809-6891v22e-68544.

SZETO, C.; LOBOS, C. A.; NGUYEN, A. T.; GRAS, S. TCR Recognition of Peptide–MHC-I: Rule Makers and Breakers. **International Journal of Molecular Sciences**, v. 22, art. 68, 2021. DOI: 10.3390/ijms22010068.

XIE, N.; SHEN, G.; GAO, W.; HUANG, Z.; HUANG, C.; FU, L. Neoantigens: promising targets for cancer therapy. **Signal Transduction and Targeted Therapy**, v. 8, art. 9, 2023. DOI: 10.1038/s41392-022-01270-x.

ZHANG, Z. *et al.* Neoantigen: A New Breakthrough in Tumor Immunotherapy. **Frontiers in Immunology**, v. 12, art. 672356, 2021. DOI: 10.3389/fimmu.2021.672356.

CAPÍTULO 7

CARACTERIZAÇÃO GENÔMICA DE DETERMINANTES DE VIRULÊNCIA E RESISTÊNCIA EM *ESCHERICHIA COLI* ASSOCIADA À PIELONEFRITE: ESTUDO DA CEPA CFT073

GENOMIC CHARACTERIZATION OF VIRULENCE AND RESISTANCE DETERMINANTS IN *ESCHERICHIA COLI* ASSOCIATED WITH PYELONEPHRITIS: A STUDY OF STRAIN CFT073

DOI: <https://doi.org/10.56001/26.9786501906539.07>

Submetido em: 19/10/2026

Revisado em: 19/10/2026

Publicado em: 24/10/2026

Ruana Carolina Cabral da Silva

Programa de Pós-Graduação em Ciências da Saúde, Universidade Federal da Grande
Dourados, Mato Grosso do Sul, Brasil

<http://lattes.cnpq.br/8634356371128001>

Gleyce Hellen de Almeida de Souza

Programa de Pós-Graduação em Ciências da Saúde, Universidade Federal da Grande
Dourados, Dourados, Mato Grosso do Sul, Brasil

<http://lattes.cnpq.br/8163519978837401>

Fabíola Lucini

Programa de Pós-Graduação em Ciências da Saúde, Universidade Federal da Grande
Dourados, Dourados, Mato Grosso do Sul, Brasil

<http://lattes.cnpq.br/1608714419112836>

Jakeline Moreira da Silva

Escola Mínima Antônio de Moraes Andrade, Macaparana, Pernambuco, Brasil

<http://lattes.cnpq.br/7884741057259883>

José Lucas de Araújo

Escola Municipal em Tempo Integral Severino Francisco da Silva

Macaparana, Pernambuco, Brasil

<http://lattes.cnpq.br/3760156602428318>

José Cláudio Gomes de Araújo

Programa de Pós-Graduação em Ciência e Engenharia de Materiais, Universidade

Federal da Paraíba, João Pessoa, Paraíba, Brasil

<http://lattes.cnpq.br/0757182882367448>

José Eduardo Souza Echeverria

Programa de Pós-Graduação em Ciências da Saúde, Universidade Federal da Grande

Dourados, Dourados, Mato Grosso do Sul, Brasil

<http://lattes.cnpq.br/1188506108074331>

Lucas Dionizio Leite de Abreu

Programa de Pós-Graduação em Ciências da Saúde, Universidade Federal da Grande

Dourados, Dourados, Mato Grosso do Sul, Brasil

<http://lattes.cnpq.br/0406653660316372>

Rousilândia de Araujo Silva

Programa de Pós-Graduação em Ciências Farmacêuticas, Universidade Federal de

Pernambuco, Recife, Pernambuco, Brasil

<http://lattes.cnpq.br/8194570266422228>

Maria Cidinaria Silva Alves

Laboratório de Parasitologia Molecular, Faculdade de Medicina de Ribeirão Preto,

Universidade de São Paulo, Ribeirão Preto, São Paulo, Brasil

<http://lattes.cnpq.br/1633031262017450>

Resumo

A caracterização genômica de *Escherichia coli* uropatogênica permite investigar componentes associados à interação com o hospedeiro e à resistência aos antimicrobianos. Este estudo teve como objetivo caracterizar o perfil molecular e identificar determinantes associados à virulência e à resistência no genoma público de *E. coli* CFT073, cepa isolada de paciente com pielonefrite aguda. Foi realizado um estudo bioinformático descritivo e exploratório, utilizando a montagem completa GCF_000007445.1, disponível no National Center for Biotechnology Information. A tipagem foi conduzida com MLST, enquanto as triagens de resistência e virulência foram realizadas com ABRicate nas bases NCBI e VFDB, respectivamente, aplicando limiares mínimos de 90% de identidade nucleotídica e cobertura. Os registros foram organizados por nome, frequência e coordenadas cromossômicas, com processamento em Python. A cepa foi identificada como ST73. Na VFDB, foram detectadas 139 correspondências, distribuídas entre 128 nomes distintos, todas cromossômicas, abrangendo grupos relacionados à adesão, aquisição de ferro, cápsula, hemolisina, autotransportadores e região *clb*. Onze nomes do grupo *pap* foram encontrados em

duas regiões distintas. A triagem de resistência identificou uma correspondência à referência blaEC-19, com cobertura de 100% e identidade de 99,47%. Concluiu-se que a abordagem integrada permitiu descrever o repertório genético e sua organização cromossômica, oferecendo uma base reprodutível para comparações e investigações funcionais de isolados associados à pielonefrite.

Palavras-chave: *Escherichia coli*; Pielonefrite; Virulência; Resistência antimicrobiana; Bioinformática.

Abstract

Genomic characterization of uropathogenic *Escherichia coli* enables the investigation of components associated with host interactions and antimicrobial resistance. This study aimed to characterize the molecular profile and identify virulence- and resistance-associated determinants in the publicly available genome of *E. coli* CFT073, a strain isolated from a patient with acute pyelonephritis. A descriptive and exploratory bioinformatics study was conducted using the complete assembly GCF_000007445.1, available from the National Center for Biotechnology Information. Sequence typing was performed using MLST, while resistance and virulence screening was carried out using ABRicate with the NCBI and VFDB databases, respectively, applying minimum thresholds of 90% nucleotide identity and coverage. Records were organized by name, frequency, and chromosomal coordinates and processed using Python. The strain was identified as ST73. VFDB screening detected 139 matches corresponding to 128 distinct names, all chromosomal, encompassing groups associated with adhesion, iron acquisition, capsule, hemolysin, autotransporters, and the *clb* region. Eleven names belonging to the *pap* group were found in two distinct regions. Resistance screening identified one match to the blaEC-19 reference, with 100% coverage and 99.47% identity. The integrated approach enabled characterization of the genetic repertoire and its chromosomal organization, providing a reproducible basis for comparisons and functional investigations of isolates associated with pyelonephritis.

Keywords: *Escherichia coli*; Pyelonephritis; Virulence; Antimicrobial resistance; Bioinformatics.

Introdução

As infecções do trato urinário constituem um campo relevante de investigação das relações entre adaptação bacteriana e adoecimento humano. Entre seus agentes etiológicos, destaca-se *Escherichia coli* uropatogênica (UPEC), cuja capacidade de estabelecer infecção envolve diferentes características genéticas e funcionais. Estudos com a cepa CFT073 demonstram que a colonização do trato urinário depende de fatores que favorecem a sobrevivência e a multiplicação bacteriana nesse ambiente, abrangendo funções que ultrapassam a presença de determinantes de virulência tradicionalmente reconhecidos (Shea *et al.*, 2020). Essa perspectiva permite compreender a patogenicidade como resultado da interação entre múltiplos componentes bacterianos e as condições encontradas no hospedeiro.

No contexto da pielonefrite, a cepa CFT073 representa um modelo de estudo com origem clínica documentada. Isolada do sangue de uma paciente com pielonefrite aguda, teve seu genoma completo descrito por Welch *et al.* (2002). A comparação com outros genomas de *E. coli* evidenciou uma organização em mosaico e diferenças expressivas no conteúdo genético, contribuindo para a investigação das bases moleculares da adaptação e da patogenicidade. A disponibilidade dessa sequência possibilita examinar a

distribuição de determinantes no cromossomo e relacionar os achados de análises computacionais ao conhecimento experimental acumulado sobre a cepa.

Entre os componentes envolvidos na interação com o hospedeiro, destacam-se as estruturas de adesão. Snyder *et al.* (2005), ao investigarem a expressão de fimbrias em CFT073, identificaram regulação coordenada entre os sistemas de fimbrias do tipo 1 e fimbrias P. Os resultados demonstraram que a expressão de um sistema pode se relacionar à alteração da expressão de outro, evidenciando a complexidade regulatória das estratégias de adesão. Assim, a identificação dos grupos gênicos associados a essas estruturas oferece informações sobre o repertório disponível, enquanto sua interpretação deve considerar que a atividade depende de processos regulatórios e do contexto de interação com o hospedeiro.

A aquisição de ferro também participa da adaptação das bactérias uropatogênicas. Garcia, Brumbaugh e Mobley (2011) investigaram a contribuição de receptores de sideróforos e heme em modelos experimentais de infecção urinária, utilizando mutantes de CFT073 e da cepa 536. Os autores observaram tanto redundância funcional quanto especificidade relacionada aos sítios de colonização, demonstrando que diferentes sistemas podem contribuir de maneira distinta para a aptidão bacteriana. Esses achados fundamentam o interesse em caracterizar conjuntamente os componentes de aquisição de ferro, em vez de interpretar a presença de cada sequência como um indicador isolado de capacidade patogênica.

Além da virulência, a caracterização de determinantes associados à resistência aos antimicrobianos amplia a investigação de isolados clínicos. Stoesser *et al.* (2013) avaliaram a predição de suscetibilidade em isolados de *E. coli* e *Klebsiella pneumoniae* a partir de sequências genômicas, comparando os resultados com testes fenotípicos. O estudo demonstrou a aplicabilidade das informações genéticas à investigação da resistência, mediante critérios definidos de análise e comparação. Nesse sentido, integrar a pesquisa de determinantes de resistência à caracterização de virulência permite descrever dimensões complementares do genoma, sem pressupor que o número de correspondências detectadas expresse, diretamente, a gravidade da infecção ou o comportamento frente aos antimicrobianos.

A análise de genomas públicos oferece uma oportunidade de organizar essas informações por procedimentos computacionais reprodutíveis. Para a cepa CFT073, a combinação entre tipagem molecular, pesquisa de sequências de referência e inspeção de coordenadas permite relacionar o perfil alélico à distribuição cromossômica das

correspondências. Essa abordagem também possibilita distinguir nomes distintos de registros presentes em mais de uma região, conferindo maior precisão à descrição do repertório identificado.

Diante desse contexto, este estudo teve como objetivo caracterizar o perfil de tipagem molecular e identificar determinantes associados à virulência e à resistência aos antimicrobianos no genoma público de *E. coli* CFT073, de origem associada à pielonefrite. Buscou-se descrever as correspondências detectadas, seus parâmetros de similaridade e sua organização cromossômica, com destaque para a distribuição dos principais grupos gênicos e para os registros encontrados em regiões distintas.

Metodologia

Foi realizado um estudo bioinformático descritivo e exploratório, utilizando o genoma público de *E. coli* CFT073. A seleção considerou a disponibilidade de uma montagem completa e a documentação da cepa na literatura como proveniente de paciente com pielonefrite aguda. A sequência foi obtida no National Center for Biotechnology Information (NCBI), sob o acesso RefSeq GCF_000007445.1, em formato FASTA.

As análises foram executadas em ambiente Linux, por comandos Bash, com ferramentas instaladas em ambiente Conda. Inicialmente, os cabeçalhos do arquivo FASTA foram inspecionados para conferir a identificação das sequências. A tipagem molecular foi realizada com o programa MLST, utilizando o esquema *ecoli_achtman_4*, baseado nos loci *adk*, *fumC*, *gyrB*, *icd*, *mdh*, *purA* e *recA*. Foram registrados o perfil alélico e o tipo de sequência atribuído pela ferramenta.

A pesquisa de determinantes associados à resistência aos antimicrobianos e à virulência foi conduzida com o ABRicate, utilizando, respectivamente, as bases NCBI e Virulence Factor Database (VFDB). Em ambas as buscas, foram aplicados limiares mínimos de 90% de identidade nucleotídica e 90% de cobertura da sequência de referência. As informações das bases instaladas foram consultadas pelo comando `abricate --list`.

As saídas foram armazenadas em arquivos tabulados no formato TSV. Para cada correspondência, foram examinados o nome reportado, a sequência de localização, as coordenadas inicial e final e os percentuais de identidade e cobertura. A inspeção das coordenadas permitiu avaliar a distribuição das correspondências e distinguir registros com o mesmo nome em posições genômicas diferentes.

O processamento dos arquivos foi realizado com scripts em Python, utilizando os módulos padrão csv e collections. Foram contabilizados os registros, os nomes distintos e suas frequências. As correspondências de virulência foram organizadas descritivamente por grupos gênicos, preservando as denominações e os sinônimos apresentados na saída original.

Foi desenvolvido um script para representar as correspondências em um mapa cromossômico linear. O procedimento utilizou o comprimento da sequência, calculado diretamente do FASTA, e as coordenadas fornecidas pelo ABRicate.

A interpretação foi descritiva, considerando a tipagem molecular, a similaridade com as referências e a organização genômica das correspondências. Foram utilizados exclusivamente dados públicos, sem coleta de amostras biológicas ou acesso a informações individuais identificáveis.

Resultados e Discussão

Caracterização molecular e panorama das correspondências identificadas

A tipagem molecular identificou *E. coli* CFT073 como pertencente ao tipo de sequência ST73, pelo esquema *ecoli_achtman_4*. O perfil alélico foi constituído por *adk*(36), *fumC*(24), *gyrB*(9), *icd*(13), *mdh*(17), *purA*(11) e *recA*(25). A pesquisa com ABRicate identificou uma correspondência na base de resistência NCBI e 139 correspondências na VFDB, distribuídas entre 128 nomes distintos. Todos os registros estavam localizados no cromossomo NC_004431.1. A Tabela 1 sintetiza os resultados da caracterização.

Tabela 1: Síntese da caracterização bioinformática do genoma de *Escherichia coli* CFT073.

Parâmetro	Resultado
Montagem analisada	GCF_000007445.1
Sequência cromossômica	NC_004431.1
Esquema MLST	<i>ecoli_achtman_4</i>
Tipo de sequência	ST73
Perfil alélico	<i>adk</i> (36), <i>fumC</i> (24), <i>gyrB</i> (9), <i>icd</i> (13), <i>mdh</i> (17), <i>purA</i> (11), <i>recA</i> (25)

Correspondências na base NCBI de resistência	1
Correspondências na VFDB	139
Nomes distintos reportados na VFDB	128
Nomes identificados em duas posições	11, pertencentes ao grupo pap
Localização das correspondências	Cromossômica

A identificação de ST73 permite situar a cepa em um perfil de tipagem já investigado em infecções extraintestinais. Alhashash *et al.* (2016), ao analisarem isolados ST73 relacionados a bacteremia e infecção urinária, observaram diversidade genômica e plasmidial, acompanhada de conservação de parte do conteúdo associado à virulência. Essa observação evidencia que o compartilhamento de um ST não implica identidade entre genomas ou uniformidade dos perfis de resistência. No presente estudo, o ST73 funciona como referência de classificação, complementada pela descrição das correspondências e de sua localização.

Repertório de sequências associadas à virulência

As 139 correspondências na VFDB abrangeram diferentes grupos relacionados à interação com o hospedeiro, incluindo sistemas de adesão, aquisição de ferro, componentes capsulares, hemolisina, autotransportadores e região *clb*. O agrupamento apresentado na Tabela 2 organiza os principais conjuntos identificados, preservando a distinção entre componentes de um sistema e fatores de virulência independentes.

Tabela 2: Principais grupos de correspondências identificados na triagem de virulência de *Escherichia coli* CFT073.

Grupo ou região	Correspondências representativas	Característica descritiva
Fímbrias P	<i>papA, papB, papC, papD, papE, papF, papG, papH, papI, papJ, papK e papX</i>	Dois conjuntos cromossômicos para 11 nomes; <i>papX</i> identificado em uma região
Fímbrias do tipo 1	<i>fimA, fimB, fimC, fimD, fimE, fimF, fimG, fimH e fimI</i>	Conjunto localizado próximo de 5,14 Mb
Outros grupos de adesão	<i>ecp, foc, sfa, csg e fdeC</i>	Correspondências distribuídas em diferentes regiões
Enterobactina e transporte associado	<i>ent, fep e fes</i>	Componentes de biossíntese, transporte e processamento
Salmochelina	<i>iroB, iroC, iroD, iroE e iroN</i>	Conjunto localizado próximo de 1,20 Mb
Aerobactina	<i>iucA, iucB, iucC, iucD e iutA</i>	Conjunto localizado próximo de 3,46 Mb
Região relacionada à yersiniabactina	<i>ybt, irp1 e fyuA/psn</i>	Correspondências a componentes da região
Utilização de heme	<i>chuA, chuS, chuT, chuU, chuV, chuW, chuX e chuY</i>	Conjunto localizado próximo de 4,09 Mb
Sistema de hemolisina	<i>hlyA, hlyB, hlyC e hlyD</i>	Conjunto localizado próximo de 3,42 Mb
Autotransportadores	<i>pic, vat e sat</i>	Regiões distintas ao longo do cromossomo
Região relacionada à colibactina	<i>clbA–clbQ e clbS</i>	Dezoito nomes identificados em região próxima de 2,26–2,32 Mb
Componentes capsulares	<i>kpsC, kpsD, kpsE, kpsF, kpsM, kpsS e kpsU</i>	Correspondências próximas de 3,51–3,53 Mb

A variedade dos grupos identificados demonstra que o repertório anotado na VFDB contempla funções distintas. Liu *et al.* (2022) propuseram uma classificação abrangente dos fatores de virulência bacterianos, incorporando categorias que permitem

organizar esse conteúdo para além de uma lista de genes. Essa perspectiva auxilia a interpretação dos resultados de CFT073: o total de correspondências expressa registros de similaridade com referências da base, enquanto o agrupamento descritivo permite reconhecer sistemas compostos por múltiplos componentes. Assim, os 139 registros não devem ser interpretados como 139 mecanismos independentes de patogenicidade.

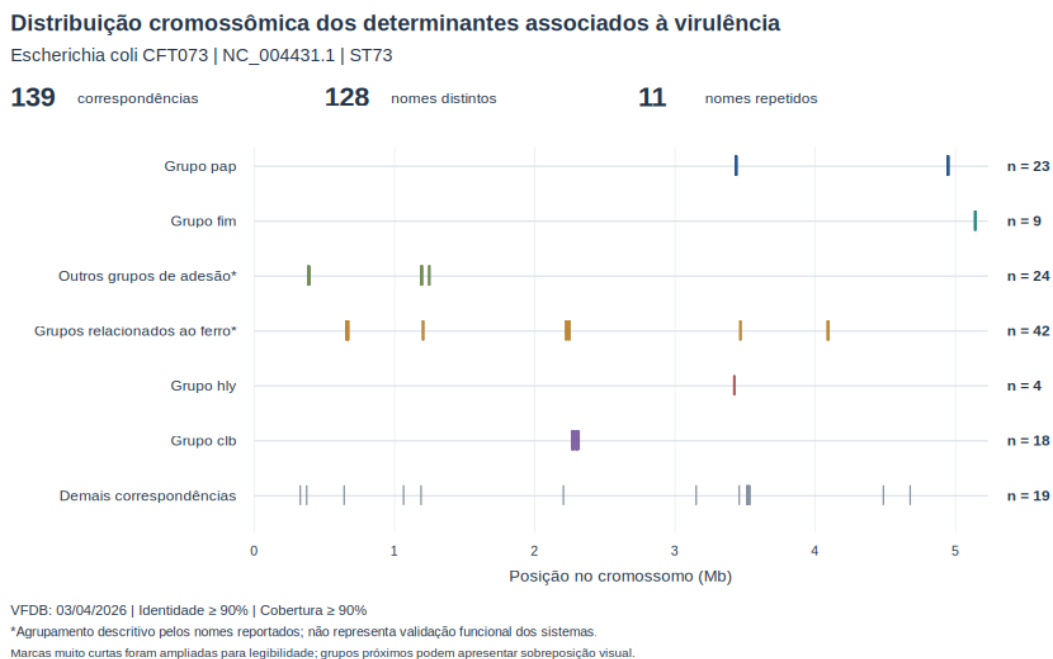
Organização cromossômica e conjuntos repetidos do grupo pap

A inspeção das coordenadas identificou dois conjuntos de correspondências do grupo pap. O primeiro estava localizado aproximadamente entre 3,43 e 3,44 Mb, e o segundo entre 4,94 e 4,95 Mb. Onze nomes: *papA*, *papB*, *papC*, *papD*, *papE*, *papF*, *papG*, *papH*, *papI*, *papJ* e *papK*, apareceram duas vezes, em posições diferentes. O registro de *papX* foi identificado somente próximo ao primeiro conjunto.

Essa organização explica a diferença entre 139 correspondências e 128 nomes distintos na VFDB. Os registros repetidos não derivam da contabilização de uma mesma posição, mas da identificação de regiões cromossômicas separadas com correspondências aos mesmos nomes de referência. A análise espacial acrescenta, portanto, informação que não seria recuperada apenas pela contagem de nomes únicos.

Também foram observados agrupamentos de correspondências fim, hly, iro, iuc, chu e clb em regiões delimitadas. A Figura 1 apresenta sua distribuição no cromossomo, permitindo visualizar os dois conjuntos pap e a posição relativa dos demais grupos.

Figura 1: Distribuição cromossômica das correspondências associadas à virulência em *Escherichia coli* CFT073.



Legenda: Mapa linear das correspondências identificadas na VFDB, com identidade e cobertura mínimas de 90%. As faixas representam agrupamentos descritivos pelos nomes reportados. As posições são proporcionais às coordenadas cromossômicas; marcas muito curtas foram ampliadas para legibilidade. O agrupamento não equivale à validação funcional dos sistemas.

Aquisição de ferro e regiões associadas à interação com o hospedeiro

As correspondências relacionadas à enterobactina, salmochelina, aerobactina e utilização de heme evidenciam a presença de diferentes componentes genéticos envolvidos na obtenção de ferro. Porcheron *et al.* (2014), em estudo experimental com CFT073, demonstraram a participação do pequeno RNA RyhB na produção de sideróforos e na virulência durante infecção urinária. Essa contribuição permite discutir os conjuntos identificados como parte de uma rede sujeita à regulação, cuja atividade não é determinada exclusivamente pela presença das sequências. Neste estudo, a distribuição de diferentes grupos descreve o repertório cromossômico disponível para investigações funcionais.

A identificação de componentes da região *clb*, juntamente com correspondências relacionadas a sideróforos, acrescenta outra dimensão à caracterização. Martin *et al.* (2013) demonstraram interação entre vias de biossíntese de sideróforos e colibactina em *E. coli*, destacando a participação de *ClbA* nesse contexto. Na análise de CFT073, a

coexistência dessas regiões fundamenta sua discussão conjunta, mas não representa demonstração de produção de colibactina ou de atividade genotóxica. O resultado obtido é a identificação e a localização de sequências relacionadas a essas funções.

Entre os autotransportadores, a correspondência a *sat* apresentou identidade e cobertura de 100%, localizada entre as posições 3.456.362 e 3.460.249. Guyer *et al.* (2002) investigaram experimentalmente Sat produzida por CFT073 e identificaram efeitos citotóxicos em células epiteliais de origem urinária, além de alterações no epitélio renal em modelo animal. A correspondência exata encontrada neste estudo conecta o resultado computacional a um determinante previamente caracterizado nessa cepa. Entretanto, a análise realizada descreve sua presença genômica, sem mensurar expressão ou atividade.

Determinante associado à resistência aos antimicrobianos

A triagem na base NCBI identificou uma correspondência à sequência de referência *bla_{EC-19}*, localizada no cromossomo. O registro apresentou cobertura de 100% e identidade de 99,47%. A Tabela 3 detalha os parâmetros obtidos.

Tabela 3: Correspondência associada à resistência aos antimicrobianos identificada em *Escherichia coli* CFT073.

Referência reportada	Sequência	Início	Final	Cobertura (%)	Identidade (%)
<i>bla_{EC-19}</i>	NC_004431.1	4.986.532	4.987.665	100,00	99,47

A diferença numérica entre os registros de resistência e virulência caracteriza as saídas das duas buscas, mas não constitui uma medida direta da importância biológica de cada dimensão. As bases possuem conteúdos e critérios próprios, e um mesmo sistema pode reunir várias correspondências. Dessa forma, a interpretação conjunta deve considerar a natureza das referências e a organização dos registros.

O perfil observado também reforça a necessidade de distinguir a caracterização de CFT073 de generalizações sobre ST73. A diversidade de resistência plasmidial identificada por Alhashash *et al.* (2016) em outros isolados desse ST demonstra que a classificação por MLST não determina, isoladamente, o repertório de resistência. No presente estudo, a contribuição foi descrever uma correspondência cromossômica nos

parâmetros empregados e integrá-la ao perfil de virulência e à organização genômica da cepa.

Considerações Finais

A caracterização bioinformática do genoma de *E. coli* CFT073, cepa associada à pielonefrite, identificou o perfil molecular ST73 e um repertório cromossômico diversificado de sequências relacionadas à virulência. A triagem na VFDB revelou 139 correspondências, distribuídas entre 128 nomes distintos, abrangendo componentes de adesão, aquisição de ferro, cápsula, hemolisina, autotransportadores e região *clb*.

A inspeção das coordenadas permitiu reconhecer dois conjuntos de correspondências do grupo *pap* em regiões cromossômicas distintas, explicando a ocorrência repetida de 11 nomes. Esse resultado demonstrou a contribuição da análise espacial para complementar as contagens e descrever a organização dos determinantes identificados. Na pesquisa de resistência aos antimicrobianos, foi detectada uma correspondência cromossômica à referência *bla_{EC-19}*, com cobertura de 100% e identidade de 99,47%.

Conclui-se que a integração entre tipagem molecular, triagem em bases de referência e mapeamento cromossômico permitiu organizar informações complementares sobre a composição genética de CFT073. A abordagem estabeleceu uma base reprodutível para comparações com outros isolados e para investigações sobre os componentes associados à infecção urinária, preservando a distinção entre presença de sequências, atividade funcional e expressão fenotípica.

Declaração de Uso de Inteligência Artificial

Durante a elaboração deste capítulo, utilizou-se o ChatGPT (OpenAI) como ferramenta auxiliar na organização e revisão textual, na tradução do resumo para o inglês e no desenvolvimento de comandos e scripts para as análises bioinformáticas. As análises foram executadas pelos autores em ambiente computacional local, utilizando o genoma público e as ferramentas descritas na metodologia. Os resultados foram conferidos e interpretados pelos autores, que assumem integral responsabilidade pela precisão das informações, pela verificação das referências bibliográficas e pela integridade científica do trabalho.

Referências

- ALHASHASH, Fahad *et al.* Increase in bacteraemia cases in the East Midlands region of the UK due to MDR *Escherichia coli* ST73: high levels of genomic and plasmid diversity in causative isolates. **Journal of Antimicrobial Chemotherapy**, v. 71, n. 2, p. 339–343, 2016. DOI: 10.1093/jac/dkv365.
- GARCIA, Erin C.; BRUMBAUGH, Ariel R.; MOBLEY, Harry L. T. Redundancy and specificity of *Escherichia coli* iron acquisition systems during urinary tract infection. **Infection and Immunity**, v. 79, n. 3, p. 1225–1235, 2011. DOI: 10.1128/IAI.01222-10.
- GUYER, Debra M. *et al.* Sat, the secreted autotransporter toxin of uropathogenic *Escherichia coli*, is a vacuolating cytotoxin for bladder and kidney epithelial cells. **Infection and Immunity**, v. 70, n. 8, p. 4539–4546, 2002. DOI: 10.1128/IAI.70.8.4539-4546.2002.
- LIU, Bo *et al.* VFDB 2022: a general classification scheme for bacterial virulence factors. **Nucleic Acids Research**, v. 50, n. D1, p. D912–D917, 2022. DOI: 10.1093/nar/gkab1107.
- MARTIN, Patricia *et al.* Interplay between siderophores and colibactin genotoxin biosynthetic pathways in *Escherichia coli*. **PLOS Pathogens**, v. 9, n. 7, e1003437, 2013. DOI: 10.1371/journal.ppat.1003437.
- PORCHERON, Gaëlle *et al.* The small RNA RyhB contributes to siderophore production and virulence of uropathogenic *Escherichia coli*. **Infection and Immunity**, v. 82, n. 12, p. 5056–5068, 2014. DOI: 10.1128/IAI.02287-14.
- SHEA, Allyson E. *et al.* *Escherichia coli* CFT073 fitness factors during urinary tract infection: identification using an ordered transposon library. **Applied and Environmental Microbiology**, v. 86, n. 13, e00691-20, 2020. DOI: 10.1128/AEM.00691-20.
- SNYDER, Jennifer A. *et al.* Coordinate expression of fimbriae in uropathogenic *Escherichia coli*. **Infection and Immunity**, v. 73, n. 11, p. 7588–7596, 2005. DOI: 10.1128/IAI.73.11.7588-7596.2005.
- STOESSER, Nicole *et al.* Predicting antimicrobial susceptibilities for *Escherichia coli* and *Klebsiella pneumoniae* isolates using whole genomic sequence data. **Journal of Antimicrobial Chemotherapy**, v. 68, n. 10, p. 2234–2244, 2013. DOI: 10.1093/jac/dkt180.
- WELCH, Rodney A. *et al.* Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. **Proceedings of the National Academy of Sciences**, v. 99, n. 26, p. 17020–17024, 2002. DOI: 10.1073/pnas.252529799.

CAPÍTULO 8

DA SEQUÊNCIA À VARIANTE: FUNDAMENTOS E ETAPAS DA ANÁLISE BIOINFORMÁTICA DE VARIANTES GENÉTICAS

FROM SEQUENCE TO VARIANT: FOUNDATIONS AND STAGES OF BIOINFORMATIC ANALYSIS OF GENETIC VARIANTS

DOI: <https://doi.org/10.56001/26.9786501906539.08>

Submetido em: 19/10/2026

Revisado em: 19/10/2026

Publicado em: 24/10/2026

Ruana Carolina Cabral da Silva

Programa de Pós-Graduação em Ciências da Saúde, Universidade Federal da Grande
Dourados, Mato Grosso do Sul, Brasil

<http://lattes.cnpq.br/8634356371128001>

Gleyce Hellen de Almeida de Souza

Programa de Pós-Graduação em Ciências da Saúde, Universidade Federal da Grande
Dourados, Dourados, Mato Grosso do Sul, Brasil

<http://lattes.cnpq.br/8163519978837401>

Fabíola Lucini

Programa de Pós-Graduação em Ciências da Saúde, Universidade Federal da Grande
Dourados, Dourados, Mato Grosso do Sul, Brasil

<http://lattes.cnpq.br/1608714419112836>

Lais Albuquerque de Oliveira

Programa de Pós-Graduação em Ciências da Saúde, Universidade Federal da Grande
Dourados, Dourados, Mato Grosso do Sul, Brasil

<http://lattes.cnpq.br/2161403193721501>

Jakeline Moreira da Silva

Escola Mínima Antônio de Moraes Andrade, Macaparana, Pernambuco, Brasil

<http://lattes.cnpq.br/7884741057259883>

José Lucas de Araújo

Escola Municipal em Tempo Integral Severino Francisco da Silva

Macaparana, Pernambuco, Brasil

<http://lattes.cnpq.br/3760156602428318>

José Cláudio Gomes de Araújo

Programa de Pós-Graduação em Ciência e Engenharia de Materiais, Universidade

Federal da Paraíba, João Pessoa, Paraíba, Brasil

<http://lattes.cnpq.br/0757182882367448>

José Eduardo Souza Echeverria

Programa de Pós-Graduação em Ciências da Saúde, Universidade Federal da Grande

Dourados, Dourados, Mato Grosso do Sul, Brasil

<http://lattes.cnpq.br/1188506108074331>

Rousilândia de Araujo Silva

Programa de Pós-Graduação em Ciências Farmacêuticas, Universidade Federal de

Pernambuco, Recife, Pernambuco, Brasil

<http://lattes.cnpq.br/8194570266422228>

Maria Cidinaria Silva Alves

Laboratório de Parasitologia Molecular, Faculdade de Medicina de Ribeirão Preto,

Universidade de São Paulo, Ribeirão Preto, São Paulo, Brasil

<http://lattes.cnpq.br/1633031262017450>

Resumo

A análise bioinformática de variantes genéticas compreende procedimentos que transformam leituras de DNA em registros passíveis de interpretação biológica. Este capítulo teve como objetivo apresentar os fundamentos e as principais etapas desse percurso, com ênfase na identificação de variantes de nucleotídeo único e pequenas inserções e deleções por alinhamento a um genoma de referência. Foi realizada uma revisão narrativa da literatura, de caráter descritivo e abordagem didático-metodológica, mediante consultas ao PubMed, a periódicos científicos e à documentação oficial de ferramentas bioinformáticas. As fontes foram selecionadas intencionalmente, considerando sua pertinência e complementaridade para a construção de um roteiro conceitual. A síntese foi organizada em obtenção dos dados, controle de qualidade, seleção da referência, alinhamento, avaliação dos alinhamentos, chamada de variantes, normalização, filtragem, anotação e interpretação. A discussão destacou a necessidade de adequar ferramentas e parâmetros às características dos dados, ao organismo, à ploidia e ao objetivo da investigação. Também evidenciou a distinção entre detecção de variantes, previsão de consequências moleculares e atribuição de significado biológico ou clínico. Concluiu-se que a organização integrada dessas etapas oferece uma base didática para o planejamento de análises documentadas e reprodutíveis. O roteiro constitui uma proposta fundamentada na literatura, sem execução experimental ou avaliação de desempenho.

Palavras-chave: Bioinformática; Sequenciamento de DNA; Variantes genéticas; Chamada de variantes; Reprodutibilidade.

Abstract

Bioinformatic analysis of genetic variants comprises procedures that transform DNA sequencing reads into records suitable for biological interpretation. This chapter aimed to present the foundations and main stages of this process, emphasizing the identification of single-nucleotide variants and small insertions and deletions through alignment to a reference genome. A descriptive narrative literature review with an educational and methodological approach was conducted through searches in PubMed, scientific journals, and official bioinformatics tool documentation. Sources were purposively selected according to their relevance and complementary contributions to a conceptual workflow. The synthesis addressed data acquisition, quality control, reference selection, alignment, alignment assessment, variant calling, normalization, filtering, annotation, and interpretation. The discussion highlighted the need to adapt tools and parameters to data characteristics, organism, ploidy, and research objectives. It also emphasized the distinction between variant detection, prediction of molecular consequences, and assignment of biological or clinical significance. The integrated organization of these stages provides an educational foundation for planning documented and reproducible analyses. The workflow is a literature-based proposal and was not experimentally implemented or evaluated for performance.

Keywords: Bioinformatics; DNA sequencing; Genetic variants; Variant calling; Reproducibility.

Introdução

O sequenciamento de DNA permite investigar diferenças genéticas entre indivíduos, populações e organismos, produzindo informações relevantes para a pesquisa biológica e a saúde. Entretanto, a transformação das leituras geradas pelo sequenciador em um conjunto de variantes interpretáveis depende de procedimentos computacionais articulados. A identificação de diferenças em relação a uma sequência de referência exige organizar os dados, avaliar sua qualidade e aplicar métodos capazes de distinguir sinais de variação de possíveis artefatos. O desenvolvimento de estruturas como o Genome Analysis Toolkit contribuiu para sistematizar o processamento de dados de sequenciamento e a implementação de métodos de análise genômica (McKenna *et al.*, 2010).

Nesse percurso, a qualidade dos dados de entrada constitui uma dimensão central. As leituras podem apresentar bases de baixa qualidade e sequências de adaptadores que precisam ser avaliadas antes das etapas subsequentes. Chen *et al.* (2018), ao apresentarem o fastp, reuniram procedimentos de avaliação de qualidade, filtragem e processamento de arquivos FASTQ, incluindo a remoção de adaptadores. Essas operações permitem examinar as características das leituras e documentar as alterações introduzidas durante o processamento. A escolha dos procedimentos deve acompanhar as propriedades dos dados e o objetivo da investigação, preservando informações relevantes para a análise.

Após a avaliação inicial, uma estratégia amplamente utilizada envolve o alinhamento das leituras a um genoma de referência e o processamento dos arquivos

resultantes. A organização dos alinhamentos permite examinar a distribuição das leituras e preparar os dados para a identificação de variantes. Danecek *et al.* (2021) descrevem o desenvolvimento de SAMtools e BCFtools, conjuntos de ferramentas voltados à manipulação de alinhamentos e de arquivos de variantes, incluindo procedimentos de chamada, filtragem e organização dos resultados. Essas ferramentas ilustram a importância de formatos padronizados e operações reprodutíveis na integração das etapas de um fluxo bioinformático.

A chamada de variantes corresponde à inferência de diferenças genéticas a partir das evidências disponíveis nos dados. Seu resultado depende das características das leituras, da referência utilizada e dos modelos empregados na análise. Entre as classes frequentemente investigadas estão as variantes de nucleotídeo único e as pequenas inserções e deleções, cuja identificação deve ser acompanhada da avaliação dos indicadores de qualidade. Nesse contexto, a filtragem integra o processo de seleção dos registros que atendem aos critérios definidos para o estudo, exigindo que parâmetros e decisões sejam explicitados para permitir a compreensão e a reprodução do procedimento (McKenna *et al.*, 2010; Danecek *et al.*, 2021).

A obtenção de um arquivo de variantes, contudo, representa uma etapa intermediária da investigação. A anotação acrescenta informações sobre a localização das alterações e suas possíveis consequências em genes, transcritos e outras regiões genômicas. McLaren *et al.* (2016), ao apresentarem o Ensembl Variant Effect Predictor, demonstraram uma abordagem para análise, anotação e priorização de variantes em regiões codificantes e não codificantes. A associação entre coordenadas, alelos e informações funcionais amplia as possibilidades de investigação, permitindo organizar os achados de acordo com a pergunta biológica e os recursos de anotação disponíveis.

A interpretação requer uma distinção adicional entre consequência molecular prevista e significado clínico. No contexto das doenças mendelianas humanas, Richards *et al.* (2015) propuseram recomendações para a classificação de variantes mediante integração de diferentes tipos de evidência, incluindo informações populacionais, funcionais, computacionais e de segregação. Essa contribuição evidencia que a identificação de uma alteração ou sua anotação como potencialmente relevante não estabelece, isoladamente, uma relação causal com uma doença. A passagem da variante detectada à interpretação depende, portanto, do contexto de aplicação e da avaliação crítica das evidências.

Diante desse cenário, este capítulo tem como objetivo apresentar os fundamentos e as principais etapas da análise bioinformática de variantes genéticas, desde a obtenção das sequências até a anotação e a interpretação dos registros. O texto adota uma abordagem metodológica e didática, com ênfase em fluxos baseados no alinhamento a um genoma de referência e na identificação de variantes de nucleotídeo único e pequenas inserções e deleções. Busca-se esclarecer a finalidade de cada etapa e as decisões necessárias para construir análises documentadas e reprodutíveis, sem restringir a discussão a um cromossomo ou organismo específico.

Metodologia

Este capítulo foi elaborado por meio de revisão narrativa da literatura, de caráter descritivo e abordagem didático-metodológica, voltada à apresentação dos fundamentos e das etapas da análise bioinformática de variantes genéticas. O recorte contemplou fluxos baseados no alinhamento de leituras de DNA a um genoma de referência, com ênfase na identificação de variantes de nucleotídeo único e pequenas inserções e deleções.

O levantamento bibliográfico foi realizado mediante consultas ao PubMed, às páginas dos periódicos científicos e à documentação oficial de ferramentas bioinformáticas. Foram utilizadas combinações de termos em inglês relacionados a sequenciamento de DNA, controle de qualidade, alinhamento, chamada de variantes, filtragem, anotação e interpretação, incluindo *DNA sequencing*, *quality control*, *read alignment*, *variant calling*, *variant filtering* e *variant annotation*. A seleção foi intencional, considerando a pertinência dos materiais para a explicação das etapas e sua complementaridade na construção do roteiro.

Foram incluídos artigos de apresentação de ferramentas, publicações metodológicas e recomendações de interpretação relacionadas ao tema. A leitura considerou a finalidade dos procedimentos, os tipos de dados de entrada, os arquivos produzidos e os critérios envolvidos nas decisões analíticas. As contribuições específicas da interpretação clínica de variantes humanas foram contextualizadas de acordo com seu campo de aplicação.

A síntese foi organizada em etapas: obtenção e identificação dos dados; avaliação e processamento das leituras; seleção da referência e alinhamento; processamento e avaliação dos alinhamentos; chamada de variantes; normalização e filtragem; anotação; e interpretação. Para cada etapa, foram considerados seu objetivo, sua relação com as demais operações e os elementos necessários à documentação do procedimento, incluindo

versões das ferramentas, referência utilizada e parâmetros de execução. A partir dessa organização, foi estruturado um roteiro conceitual para orientar a compreensão do percurso entre as sequências e as variantes anotadas.

Resultados e Discussão

Organização do percurso entre sequências e variantes

A síntese da literatura permitiu estruturar um roteiro metodológico que conecta a obtenção das sequências à interpretação das variantes. O principal produto desta revisão foi a organização das etapas, de seus arquivos e das decisões necessárias à continuidade da análise. O roteiro concentra-se em variantes de nucleotídeo único e pequenas inserções e deleções identificadas por alinhamento a uma referência, sem representar a execução ou a validação de um protocolo experimental.

A organização proposta está apresentada no Quadro 1. As etapas constituem uma estrutura orientadora, cuja aplicação exige adequação à tecnologia de sequenciamento, ao organismo, à ploidia e ao objetivo do estudo.

Quadro 1: Etapas e decisões do roteiro bioinformático para análise de variantes genéticas.

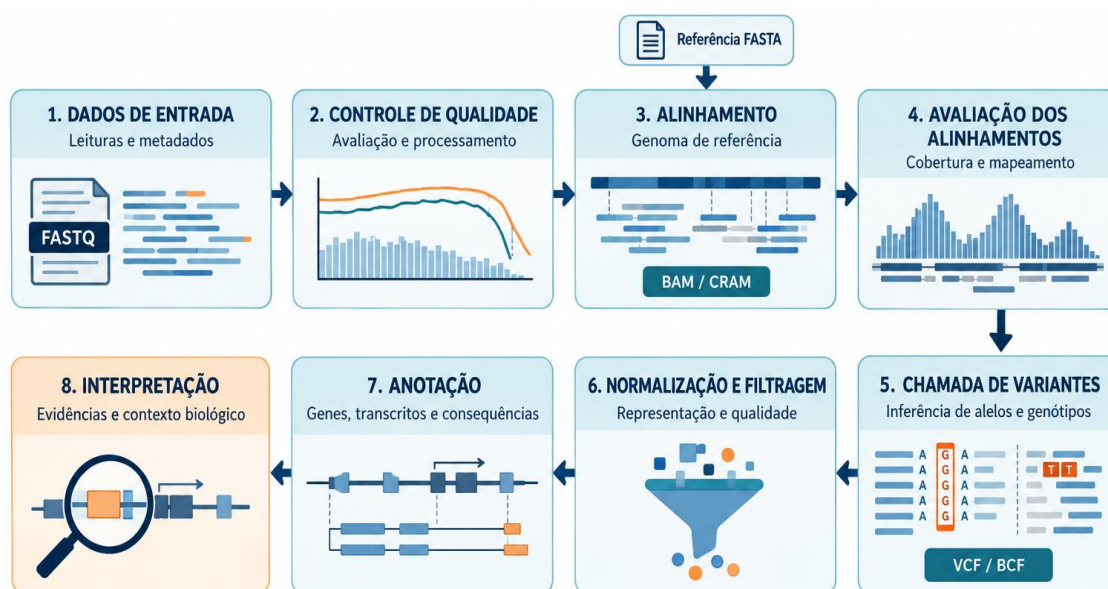
Etapas	Finalidade	Arquivos ou produtos	Decisões principais
Obtenção dos dados	Identificar sequências e informações da amostra	FASTQ e metadados	Origem, tecnologia, desenho e ploidia
Controle de qualidade	Examinar características das leituras	Relatórios e FASTQ processados	Necessidade de remoção de adaptadores e filtragem
Seleção da referência	Definir a sequência utilizada na comparação	FASTA e índices	Montagem, versão e compatibilidade
Alinhamento	Posicionar as leituras na referência	SAM, BAM ou CRAM	Alinhador e parâmetros apropriados
Processamento e avaliação	Organizar e examinar os alinhamentos	Arquivos ordenados, indexados e relatórios	Cobertura, qualidade de mapeamento e duplicatas
Chamada de variantes	Inferir alelos e genótipos	VCF ou BCF	Modelo, ploidia e tipo de variante
Normalização e filtragem	Padronizar registros e selecionar chamadas	VCF ou BCF processados	Representação dos alelos e critérios de qualidade

Anotação	Associar variantes a informações genômicas	Arquivos anotados e tabelas	Referência, transcritos e versões das bases
Interpretação	Relacionar os achados à pergunta do estudo	Síntese dos registros priorizados	Evidências e contexto de aplicação

Fonte: Elaboração dos autores a partir da síntese da literatura.

DePristo *et al.* (2011) apresentaram uma estrutura integrada para descoberta e genotipagem de variantes, articulando processamento dos dados, inferência e avaliação de artefatos. A contribuição desse trabalho para o roteiro está no reconhecimento de que a qualidade da chamada depende de operações anteriores e posteriores ao algoritmo de detecção. Seus procedimentos devem, entretanto, ser contextualizados historicamente: etapas específicas daquele fluxo não constituem exigências universais para todas as ferramentas e tecnologias. A aplicação prática requer compatibilidade entre os métodos escolhidos e os dados disponíveis.

Figura 1: Fluxo conceitual da análise bioinformática: das leituras de DNA à interpretação de variantes.



Referência e alinhamento como bases da comparação

A referência estabelece o sistema de coordenadas utilizado para posicionar as leituras e descrever as variantes. Sua seleção precisa anteceder as operações de alinhamento e anotação, mantendo coerência entre os arquivos empregados. No roteiro

proposto, a identificação da montagem e de sua versão integra a documentação mínima da análise.

Li e Durbin (2009) apresentaram o BWA como uma abordagem para alinhamento de leituras curtas a sequências de referência, considerando diferenças e lacunas durante o mapeamento. O estudo evidencia que o alinhamento envolve uma decisão computacional sobre a posição das leituras, e não apenas uma comparação literal entre sequências. Essa contribuição fundamenta a necessidade de avaliar a adequação do alinhador ao tipo de dado e de considerar a confiança atribuída ao posicionamento antes da chamada de variantes.

Como implicação metodológica, as diferenças observadas entre leitura e referência devem ser examinadas em conjunto com as evidências do alinhamento. Regiões com posicionamento ambíguo exigem atenção porque a interpretação de uma diferença depende de sua localização. Assim, a avaliação dos alinhamentos constitui uma etapa de decisão sobre a qualidade das evidências disponíveis para a inferência.

Normalização e filtragem dos registros

A normalização organiza a representação das variantes para facilitar comparações e integração entre arquivos. Tan, Abecasis e Kang (2015) demonstraram a importância de representar variantes de maneira consistente e concisa, especialmente em situações nas quais uma mesma alteração pode admitir diferentes descrições. Essa contribuição é relevante para a análise de inserções e deleções e para a comparação entre resultados produzidos por métodos distintos.

No roteiro, normalização e filtragem possuem finalidades complementares. A primeira atua sobre a representação dos registros; a segunda aplica critérios de seleção às chamadas. A padronização de uma variante, portanto, não comprova sua qualidade biológica ou técnica. Da mesma forma, um registro que atende a determinado filtro continua dependente das evidências que sustentaram sua identificação.

A definição dos filtros deve considerar os indicadores disponíveis e a pergunta do estudo. Em vez de estabelecer um conjunto único de limiares para qualquer situação, propõe-se explicitar quais características serão examinadas e como orientarão a seleção. O Quadro 2 reúne dimensões que podem ser consideradas nessa avaliação.

Quadro 2: Dimensões para avaliação e documentação das chamadas de variantes.

Dimensão	Questão orientadora	Informação a registrar
Suporte por leituras	Quais evidências sustentam o alelo?	Profundidade e leituras de suporte
Qualidade do mapeamento	O posicionamento das leituras é confiável?	Indicadores de alinhamento
Qualidade da chamada	Qual confiança foi atribuída ao registro?	Escores e filtros aplicados
Genótipo e ploidia	O modelo corresponde à amostra?	Configuração e critérios utilizados
Representação	Os alelos estão descritos consistentemente?	Procedimento de normalização
Contexto genômico	A região apresenta características que exigem atenção?	Regiões repetitivas ou de difícil análise
Anotação	Quais recursos foram associados à variante?	Montagem, transcritos e versões das bases
Avaliação externa	Há materiais adequados para verificar o fluxo?	Conjunto de referência e regiões avaliadas

Fonte: Elaboração dos autores.

Avaliação de desempenho e confiabilidade

A avaliação de um fluxo exige critérios adicionais à contagem de variantes produzidas. Krusche *et al.* (2019) apresentaram recomendações para comparação de chamadas de pequenas variantes germinativas humanas com conjuntos de referência, considerando diferenças de representação e características das regiões avaliadas. O trabalho destaca a importância de métricas como precisão e sensibilidade e da estratificação por contexto genômico.

Essa abordagem permite compreender que uma análise com maior número de registros não é necessariamente superior. A avaliação depende da relação entre chamadas corretas, incorretas e variantes não recuperadas, dentro de regiões apropriadas à comparação. Para o roteiro proposto, o uso de materiais de referência representa uma possibilidade de verificar o desempenho quando houver dados compatíveis. As recomendações voltadas a variantes germinativas humanas devem ser aplicadas dentro desse escopo, sem extensão automática a outros organismos ou tipos de alteração.

Anotação e construção da interpretação

A anotação acrescenta informações sobre a relação das variantes com elementos genômicos. Cingolani *et al.* (2012), ao apresentarem o SnpEff, demonstraram uma abordagem para anotar e prever consequências de alterações em diferentes contextos da sequência. Essa contribuição permite organizar registros conforme sua posição em relação aos modelos gênicos e suas consequências previstas, oferecendo uma etapa intermediária entre a detecção e a interpretação biológica.

A interpretação proposta no capítulo parte dessa distinção. Uma consequência prevista depende da referência e dos modelos de anotação utilizados; sua relevância depende também da pergunta investigada. Por isso, o roteiro recomenda preservar as informações que sustentam a priorização, permitindo compreender por que determinados registros foram selecionados.

Em aplicações clínicas, a interpretação exige procedimentos e evidências específicos, adicionais à chamada e à anotação. Em estudos populacionais, evolutivos ou microbiológicos, outras perguntas orientarão a análise. O percurso técnico pode compartilhar etapas, mas as conclusões precisam acompanhar o contexto em que os dados foram produzidos.

Documentação e reprodutibilidade do fluxo

A documentação constitui um componente transversal do roteiro. Devem ser registrados a origem dos dados, a referência, as versões das ferramentas e bases, os parâmetros e a relação entre arquivos de entrada e saída. Essa organização permite compreender as decisões e repetir o procedimento em condições equivalentes.

Mölder *et al.* (2021) discutiram a sustentabilidade da análise de dados com Snakemake, relacionando a organização de fluxos à transparência, adaptação e reprodutibilidade. A contribuição para este capítulo está na possibilidade de representar dependências entre operações e documentar sua execução de maneira estruturada. A automatização pode apoiar esse processo, mas precisa incorporar verificações e registros que permitam identificar falhas.

O roteiro resultante desta revisão organiza, portanto, uma sequência de decisões documentáveis. Seu valor didático está em tornar explícita a função de cada etapa e sua conexão com as demais. A passagem da sequência à variante interpretada depende dessa

articulação entre qualidade dos dados, inferência computacional, representação consistente e avaliação das evidências.

Considerações Finais

A análise bioinformática de variantes genéticas envolve etapas interdependentes que conectam a qualidade das sequências à interpretação dos registros identificados. A síntese apresentada permitiu organizar esse percurso em um roteiro didático, abrangendo obtenção dos dados, controle de qualidade, alinhamento, avaliação dos alinhamentos, chamada de variantes, normalização, filtragem, anotação e interpretação.

A discussão evidenciou que as decisões sobre referência, ferramentas, parâmetros e critérios de seleção precisam acompanhar as características dos dados e os objetivos da investigação. Também destacou a importância de distinguir a detecção de uma variante, sua consequência molecular prevista e seu significado biológico ou clínico, reconhecendo que cada dimensão exige evidências próprias.

Conclui-se que a principal contribuição do capítulo consiste em explicitar a finalidade e a articulação das etapas, oferecendo uma base para o planejamento de análises documentadas e reproduzíveis. O roteiro proposto pode apoiar a formação em bioinformática e a organização de estudos genéticos, devendo ser adaptado ao organismo, à ploidia e à tecnologia de sequenciamento, com avaliação de desempenho quando aplicado a dados reais.

Declaração de Uso de Inteligência Artificial Generativa

Durante a elaboração deste capítulo, utilizou-se o ChatGPT (OpenAI) como ferramenta auxiliar na organização e redação preliminar do texto, na revisão linguística e na elaboração da figura ilustrativa do fluxo bioinformático. Os conteúdos gerados foram submetidos à análise crítica e à revisão dos autores, incluindo a verificação das informações e das referências bibliográficas. A interpretação dos estudos, as decisões editoriais e a aprovação da versão final permaneceram sob responsabilidade dos autores, que assumem integral responsabilidade pela precisão, originalidade e integridade científica do trabalho.

Referências

CHEN, Shifu *et al.* fastp: an ultra-fast all-in-one FASTQ preprocessor. **Bioinformatics**, v. 34, n. 17, p. i884–i890, 2018. DOI: 10.1093/bioinformatics/bty560.

CINGOLANI, Pablo *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. **Fly**, v. 6, n. 2, p. 80–92, 2012. DOI: 10.4161/fly.19695.

DANECEK, Petr *et al.* Twelve years of SAMtools and BCFtools. **GigaScience**, v. 10, n. 2, giab008, 2021. DOI: 10.1093/gigascience/giab008.

DEPRISTO, Mark A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. **Nature Genetics**, v. 43, n. 5, p. 491–498, 2011. DOI: 10.1038/ng.806.

KRUSCHE, Peter *et al.* Best practices for benchmarking germline small-variant calls in human genomes. **Nature Biotechnology**, v. 37, n. 5, p. 555–560, 2019. DOI: 10.1038/s41587-019-0054-x.

LI, Heng; DURBIN, Richard. Fast and accurate short read alignment with Burrows-Wheeler transform. **Bioinformatics**, v. 25, n. 14, p. 1754–1760, 2009. DOI: 10.1093/bioinformatics/btp324.

MCKENNA, Aaron *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. **Genome Research**, v. 20, n. 9, p. 1297–1303, 2010. DOI: 10.1101/gr.107524.110.

MCLAREN, William *et al.* The Ensembl Variant Effect Predictor. **Genome Biology**, v. 17, art. 122, 2016. DOI: 10.1186/s13059-016-0974-4.

MÖLDER, Felix *et al.* Sustainable data analysis with Snakemake. **F1000Research**, v. 10, art. 33, 2021. Versão 2. DOI: 10.12688/f1000research.29032.2.

RICHARDS, Sue *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. **Genetics in Medicine**, v. 17, n. 5, p. 405–424, 2015. DOI: 10.1038/gim.2015.30.

TAN, Adrian; ABECASIS, Gonçalo R.; KANG, Hyun Min. Unified representation of genetic variants. **Bioinformatics**, v. 31, n. 13, p. 2202–2204, 2015. DOI: 10.1093/bioinformatics/btv112.

CAPÍTULO 9

IDENTIFICAÇÃO E CARACTERIZAÇÃO BIOINFORMÁTICA DO GENE TP53: ESTUDO DE CASO DE UM SUPRESSOR TUMORAL

BIOINFORMATIC IDENTIFICATION AND CHARACTERIZATION OF THE TP53 GENE: A CASE STUDY OF A TUMOR SUPPRESSOR

DOI: <https://doi.org/10.56001/26.9786501906539.09>

Submetido em: 19/10/2026

Revisado em: 19/10/2026

Publicado em: 24/10/2026

Ruana Carolina Cabral da Silva

Programa de Pós-Graduação em Ciências da Saúde, Universidade Federal da Grande
Dourados, Mato Grosso do Sul, Brasil

<http://lattes.cnpq.br/8634356371128001>

Gleyce Hellen de Almeida de Souza

Programa de Pós-Graduação em Ciências da Saúde, Universidade Federal da Grande
Dourados, Dourados, Mato Grosso do Sul, Brasil

<http://lattes.cnpq.br/8163519978837401>

Lais Albuquerque de Oliveira

Programa de Pós-Graduação em Ciências da Saúde, Universidade Federal da Grande
Dourados, Dourados, Mato Grosso do Sul, Brasil

<http://lattes.cnpq.br/2161403193721501>

Mariana Carvalho Sturaro

Programa de Pós-Graduação em Biociências e Biotecnologia, Universidade de São
Paulo, Ribeirão Preto, São Paulo, Brasil

<http://lattes.cnpq.br/3041100311056029>

José Eduardo Souza Echeverria

Programa de Pós-Graduação em Ciências da Saúde, Universidade Federal da Grande
Dourados, Dourados, Mato Grosso do Sul, Brasil

<http://lattes.cnpq.br/1188506108074331>

Aline Andrade Martins

Programa de Pós-Graduação em Ciências da Saúde, Universidade Federal da Grande
Dourados, Dourados, Mato Grosso do Sul, Brasil

<http://lattes.cnpq.br/8716162097729810>

Fabíola Lucini

Programa de Pós-Graduação em Ciências da Saúde, Universidade Federal da Grande
Dourados, Dourados, Mato Grosso do Sul, Brasil

<http://lattes.cnpq.br/1608714419112836>

Resumo

O gene *TP53* codifica a proteína p53, um supressor tumoral envolvido na regulação de processos celulares relacionados à manutenção da integridade genômica. Este estudo teve como objetivo caracterizar a conservação das sequências de referência da proteína p53 humana e murina por meio de uma análise bioinformática comparativa. Foram utilizadas as sequências NP_000537.3, de *Homo sapiens*, e NP_035770.2, de *Mus musculus*, obtidas no National Center for Biotechnology Information. O alinhamento foi realizado localmente com o BLASTP, versão 2.16.0+, considerando identidade, posições positivas, cobertura, lacunas e escore. As proteínas humana e murina apresentaram, respectivamente, 393 e 390 aminoácidos. O alinhamento compreendeu 393 colunas, com cobertura de 100% da sequência humana, identidade de 77,35%, posições positivas de 82,95% e seis lacunas, correspondentes a 1,53% do alinhamento. O escore foi de 598 bits. Observou-se ampla conservação das sequências, com distribuição heterogênea das substituições e concentração das lacunas no segmento correspondente aos primeiros 60 resíduos humanos. Conclui-se que a comparação permitiu caracterizar semelhanças e diferenças entre as proteínas de referência, demonstrando a aplicabilidade de uma estratégia computacional acessível e reprodutível ao estudo da conservação molecular de um supressor tumoral.

Palavras-Chave: TP53; p53; Bioinformática; Alinhamento de sequências; Supressor tumoral.

Abstract

The *TP53* gene encodes p53, a tumor suppressor protein involved in the regulation of cellular processes associated with the maintenance of genomic integrity. This study aimed to characterize the conservation of human and mouse p53 reference sequences through a comparative bioinformatics analysis. Protein sequences NP_000537.3 from *Homo sapiens* and NP_035770.2 from *Mus musculus* were retrieved from the National Center for Biotechnology Information. Local alignment was performed using BLASTP version 2.16.0+, assessing sequence identity, positive-scoring positions, coverage, gaps, and alignment score. The human and mouse proteins contained 393 and 390 amino acids, respectively. The alignment comprised 393 columns, covering 100% of the human sequence, with 77.35% identity, 82.95% positive-scoring positions, and six gaps, corresponding to 1.53% of the alignment. The alignment score was 598 bits. Extensive sequence conservation was observed, with a heterogeneous distribution of substitutions and gaps concentrated within the segment corresponding to the first 60 human residues. The comparison characterized similarities and differences between the reference proteins, demonstrating the applicability of an accessible and reproducible computational strategy to investigate the molecular conservation of a tumor suppressor.

Keywords: TP53; p53; Bioinformatics; Sequence alignment; Tumor suppressor.

Introdução

O estudo dos genes supressores tumorais contribui para compreender os mecanismos celulares que restringem o desenvolvimento do câncer. Entre esses genes, destaca-se *TP53*, responsável pela codificação da proteína p53. Evidências experimentais estabeleceram sua importância na supressão tumoral: Donehower *et al.* (1992) demonstraram que camundongos deficientes em p53 apresentavam maior suscetibilidade ao desenvolvimento espontâneo de tumores. Esse achado contribuiu para consolidar o interesse na investigação das características moleculares da proteína e das consequências de alterações em sua função.

A relevância de *TP53* também foi evidenciada em estudos de tumores humanos. Donehower *et al.* (2019), ao integrarem informações de diferentes plataformas do The Cancer Genome Atlas, examinaram alterações do gene e de sua via em mais de dez mil amostras de diferentes tipos de câncer. A investigação demonstrou relações entre alterações de *TP53*, características genômicas e padrões de expressão dos tumores. Esses resultados situam o gene como um objeto de interesse para análises que aproximam sequência, organização molecular e biologia tumoral, reconhecendo a diversidade de contextos em que suas alterações ocorrem.

A compreensão dessas relações depende, em parte, do conhecimento da estrutura e das propriedades da proteína codificada. Cho *et al.* (1994) determinaram a estrutura do domínio central da p53 humana em complexo com DNA, oferecendo uma base para investigar como alterações em determinados resíduos podem afetar a organização estrutural ou as interações com a molécula de DNA. Essa contribuição demonstra a importância de relacionar a sequência de aminoácidos ao conhecimento funcional, considerando que diferentes posições podem participar de interações e atividades distintas.

Nesse contexto, a bioinformática oferece recursos para recuperar sequências de referência, comparar proteínas e descrever padrões de conservação. O BLAST permite identificar alinhamentos locais e avaliar sua significância estatística, fornecendo medidas como identidade, extensão do alinhamento e pontuação. Altschul *et al.* (1997) apresentaram avanços nessa abordagem com o desenvolvimento do Gapped BLAST e do PSI-BLAST, ampliando as possibilidades de investigação de similaridade entre sequências proteicas. A interpretação dessas medidas permite descrever semelhanças e

diferenças, desde que seja preservada a distinção entre evidência de similaridade e demonstração experimental de função.

A disponibilidade de sequências públicas torna possível desenvolver estudos de caso reprodutíveis com proteínas previamente caracterizadas. Para *TP53* humano, o NCBI disponibiliza o transcrito de referência NM_000546.6, associado à proteína p53 de acesso NP_000537.3. A identificação explícita dos acessos e das versões permite delimitar o material analisado e estabelecer uma referência para comparações. Neste capítulo, a abordagem foi direcionada à comparação da proteína humana com a p53 de camundongo, permitindo examinar a identidade global e sua distribuição ao longo do alinhamento.

A comparação entre espécies oferece uma perspectiva complementar ao estudo das alterações observadas em tumores. As diferenças entre sequências de referência humana e murina representam variação interespecífica e devem ser interpretadas nesse contexto. Sua análise pode revelar regiões com diferentes níveis de conservação, orientando a discussão sobre características da proteína sem atribuir automaticamente significado patogênico às substituições identificadas. Assim, o estudo de caso articula a identificação documental das sequências com sua caracterização comparativa, mantendo o vínculo entre os procedimentos computacionais e as evidências disponíveis na literatura.

Diante desse cenário, este capítulo teve como objetivo realizar a caracterização bioinformática de sequências de referência relacionadas ao gene *TP53*, com ênfase na comparação da proteína p53 humana e murina. Buscou-se descrever os parâmetros do alinhamento, identificar a distribuição de resíduos idênticos, substituições e lacunas e representar a conservação em janelas da sequência humana, contextualizando os achados no estudo de um supressor tumoral.

Metodologia

Foi realizado um estudo bioinformático descritivo e comparativo, baseado em sequências públicas de referência relacionadas ao gene *TP53* e à proteína p53. Foram selecionadas a sequência humana NP_000537.3 e a sequência de *Mus musculus* NP_035770.2, ambas identificadas nos registros como proteína p53, isoforma a. Também foi obtido o transcrito humano NM_000546.6 para identificação documental da sequência codificadora correspondente.

As sequências foram recuperadas do National Center for Biotechnology Information (NCBI), por requisições ao serviço EFetch da interface E-utilities, utilizando o programa curl. Os arquivos foram armazenados em formato FASTA, e seus cabeçalhos foram conferidos para verificar os acessos, as descrições e os organismos de origem. Os procedimentos foram conduzidos em ambiente Linux, por comandos Bash, utilizando o BLAST+ instalado em ambiente Conda.

A comparação proteica foi realizada localmente com BLASTP, versão 2.16.0+, utilizando a p53 humana como consulta e a p53 de camundongo como sequência-alvo, por meio das opções -query e -subject. Foi estabelecido um limiar de E-value de 1×10^{-10} , mantendo os demais parâmetros padrão. O relatório da execução registrou a matriz BLOSUM62, penalidades de abertura e extensão de lacunas de 11 e 1, respectivamente, e ajuste composicional da matriz de pontuação.

Os resultados foram exportados em formato tabulado e em relatório de alinhamento pareado. Foram examinados o percentual de identidade, o número de posições positivas, o comprimento do alinhamento, as coordenadas nas duas sequências, as lacunas, o E-value e o bit score. A identidade global foi considerada como a proporção de resíduos idênticos em relação ao número total de colunas do alinhamento, incluindo aquelas com lacunas.

Para examinar a distribuição da conservação, foi desenvolvido um script em Python que recuperou as sequências alinhadas do relatório textual. O procedimento organizou janelas consecutivas, não sobrepostas, de até 50 aminoácidos, utilizando as coordenadas da proteína humana. Para cada janela, o cálculo da identidade foi definido pela razão entre o número de resíduos idênticos e o número de posições humanas avaliadas, multiplicada por 100. Lacunas na sequência-alvo foram contabilizadas como posições não idênticas.

Resultados e Discussão

Caracterização das sequências e do alinhamento

A análise comparativa foi realizada entre as sequências de referência da proteína p53 humana, identificada pelo acesso NP_000537.3, e da proteína p53 de *Mus musculus*, correspondente ao acesso NP_035770.2. As sequências apresentaram, respectivamente, 393 e 390 aminoácidos. O BLASTP identificou um segmento de alinhamento que abrangeu toda a sequência humana, com identidade de 77,35% e escore de 598 bits. As

características das sequências e os principais indicadores do alinhamento estão apresentados na Tabela 1.

Tabela 1: Características das sequências e parâmetros do alinhamento entre as proteínas p53 humana e murina.

Indicador	Resultado
Proteína humana	NP_000537.3
Proteína murina	NP_035770.2
Comprimento da proteína humana	393 aminoácidos
Comprimento da proteína murina	390 aminoácidos
Intervalo alinhado na sequência humana	1–393
Intervalo alinhado na sequência murina	4–390
Comprimento do alinhamento	393 colunas
Cobertura da sequência humana	100,00%
Cobertura da sequência murina ¹	99,23%
Identidade	304/393 — 77,35%
Substituições positivas, incluindo identidades	326/393 — 82,95%
Lacunas	6/393 — 1,53%
Escore	598 bits
E-value reportado	0,0

¹ Calculada a partir dos 387 resíduos murinos abrangidos pelo alinhamento, em relação aos 390 resíduos da sequência.

A abrangência do alinhamento demonstra que a similaridade entre as proteínas se distribuiu por praticamente toda a sua extensão, sem ficar restrita a um pequeno segmento. Embora as sequências tenham comprimentos próximos, o alinhamento não correspondeu a uma associação direta entre posições de mesma numeração: a sequência humana foi alinhada a partir do primeiro resíduo, enquanto a sequência murina foi alinhada a partir do quarto.

O E-value foi apresentado pelo programa como 0,0, acompanhado de um escore elevado. Esse valor representa a precisão numérica da saída do BLASTP e indica uma correspondência altamente significativa nas condições da comparação; não deve ser interpretado como probabilidade literalmente nula de ocorrência ao acaso.

Identidade e similaridade entre as proteínas

Dos 393 elementos do alinhamento, 304 corresponderam a aminoácidos idênticos. O relatório também apresentou 326 posições positivas, categoria que inclui as identidades e as substituições com pontuação favorável na matriz BLOSUM62. Dessa forma, foram identificadas 22 substituições positivas entre aminoácidos diferentes. A decomposição dos resultados está apresentada na Tabela 2.

Tabela 2 – Distribuição das correspondências entre as proteínas p53 humana e murina.

Categoria no alinhamento	Número de posições	Percentual
Aminoácidos idênticos	304	77,35%
Substituições positivas entre aminoácidos diferentes	22	5,60%
Outras substituições	61	15,52%
Posições com lacunas	6	1,53%
Total	393	100,00%

Nota: as categorias são mutuamente exclusivas. As 22 substituições positivas foram calculadas pela diferença entre as 326 posições positivas e as 304 identidades.

A diferença entre identidade e similaridade é relevante para a interpretação do resultado. Enquanto a identidade contabiliza exclusivamente os resíduos iguais, as posições positivas também consideram substituições favorecidas pela matriz utilizada. Assim, a identidade de 77,35% deve ser interpretada em conjunto com os 82,95% de posições positivas e com a ampla cobertura das sequências.

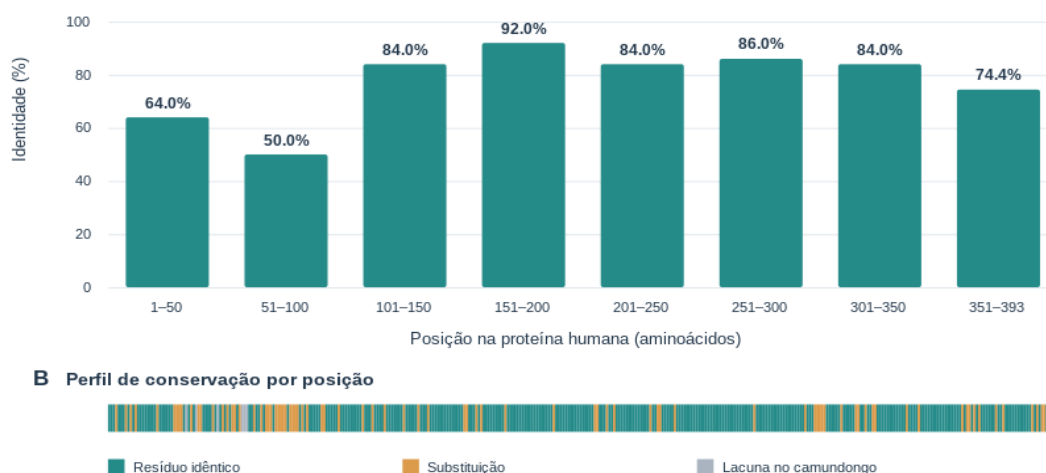
Essa conservação pode ser contextualizada pelo trabalho de Coffill *et al.* (2016), que investigou componentes da regulação de p53 em vertebrados e demonstrou a conservação de interações envolvendo p53, Mdm2 e Mdm4 entre organismos evolutivamente distantes. O estudo oferece uma base experimental para compreender a manutenção de mecanismos regulatórios da proteína, embora a conservação funcional de cada posição não possa ser estabelecida apenas pelo alinhamento apresentado neste capítulo.

Distribuição das diferenças ao longo das sequências

A inspeção do alinhamento revelou uma distribuição heterogênea das identidades e das substituições. No segmento correspondente aos primeiros 60 aminoácidos humanos, foram observadas as seis lacunas reportadas pelo BLASTP, todas na sequência murina. Nos segmentos seguintes, o alinhamento apresentou trechos consecutivos de aminoácidos idênticos, intercalados por substituições.

Esse padrão mostra que a identidade global sintetiza situações distintas ao longo da proteína. Regiões com maior concentração de diferenças coexistem com segmentos mais conservados, justificando a apresentação de uma figura que permita visualizar a distribuição dessas correspondências.

Figura 1. Distribuição da conservação entre as proteínas p53 humana e murina.



Legenda: (A) Percentual de identidade em janelas consecutivas de 50 aminoácidos, utilizando a numeração da sequência humana como referência; a última janela compreende os resíduos 351–393. (B) Representação das posições alinhadas, distinguindo aminoácidos idênticos, substituições e lacunas. Os percentuais de cada janela consideram o número de resíduos humanos abrangidos pelo respectivo intervalo.

A distribuição da conservação também pode ser discutida em relação à organização estrutural da p53. Veprintsev *et al.* (2006), por meio de análises estruturais da proteína completa em solução, demonstraram a participação de interações entre domínios na organização da p53 e em sua associação ao DNA. Esses achados ajudam a contextualizar a relevância biológica da conservação de sequência, sem substituir uma análise estrutural específica dos resíduos comparados neste estudo.

Portanto, a figura complementa os indicadores globais da Tabela 1 ao permitir a identificação visual dos segmentos com diferentes níveis de conservação. A análise por janelas tem caráter descritivo: sua finalidade é localizar a distribuição das diferenças, sem atribuir automaticamente funções ou efeitos biológicos aos intervalos representados.

Implicações para o estudo de um supressor tumoral

A comparação entre as proteínas humana e murina constitui uma abordagem inicial para caracterizar a conservação de um supressor tumoral utilizado em diferentes contextos de pesquisa. Entretanto, as diferenças entre as sequências de referência das duas espécies devem ser compreendidas como variações interespecíficas, e não como mutações tumorais identificadas em pacientes.

A importância de distinguir conservação de sequência e consequência funcional é evidenciada por Kato *et al.* (2003). Ao investigar experimentalmente 2.314 variantes de

p53 com substituições de aminoácidos, os autores identificaram respostas funcionais distintas em ensaios de transativação. Esses resultados mostram que o efeito de uma substituição depende de sua posição e de seu contexto, não podendo ser deduzido exclusivamente do percentual global de identidade entre duas proteínas.

Além disso, a conservação proteica não implica equivalência completa entre as respostas celulares humana e murina. Fischer (2019), ao comparar dados de expressão gênica e ligação de p53 ao DNA, identificou aspectos conservados e divergentes das redes regulatórias nas duas espécies. Esse trabalho reforça que a interpretação de modelos murinos deve considerar tanto a proteína quanto os mecanismos regulatórios e os genes envolvidos em sua resposta.

Em conjunto, os resultados caracterizam uma correspondência extensa entre as proteínas analisadas, com predomínio de aminoácidos idênticos e presença de diferenças distribuídas ao longo do alinhamento. O estudo de caso demonstra a aplicação de uma estratégia computacional simples para comparar sequências de referência e organizar evidências de conservação molecular relevantes para a pesquisa sobre p53.

Considerações Finais

A análise bioinformática comparativa das sequências de referência da proteína p53 humana e murina evidenciou elevada conservação molecular entre as espécies, caracterizada pelo predomínio de aminoácidos idênticos e pela ampla extensão do alinhamento. Entretanto, a distribuição heterogênea das substituições e das lacunas demonstrou que o percentual global de identidade não representa uniformemente a conservação ao longo de toda a sequência proteica.

A abordagem utilizada permitiu caracterizar as principais semelhanças e diferenças entre as proteínas de referência, demonstrando a aplicabilidade de ferramentas computacionais acessíveis à investigação de genes supressores tumorais. Contudo, as variações identificadas correspondem a diferenças interespecíficas e não permitem, isoladamente, estabelecer alterações funcionais ou implicações patológicas.

Conclui-se que a análise de conservação de sequências constitui uma etapa inicial relevante para a caracterização molecular da p53, fornecendo informações que podem subsidiar investigações posteriores. Estudos complementares, envolvendo análises estruturais e funcionais, são necessários para compreender o significado biológico das diferenças observadas.

Declaração de Uso de Inteligência Artificial

Durante a elaboração deste capítulo, utilizou-se o ChatGPT (OpenAI) como ferramenta auxiliar na organização e revisão textual, na tradução do resumo para o inglês e no desenvolvimento de comandos e scripts para as análises bioinformáticas. As análises foram executadas pela autora em ambiente computacional local, utilizando as sequências públicas de referência e as ferramentas descritas na metodologia. Os resultados foram conferidos e interpretados pela autora, que assume integral responsabilidade pela precisão das informações, pela verificação das referências bibliográficas e pela integridade científica do trabalho.

Referências

ALTSCHUL, Stephen F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. **Nucleic Acids Research**, v. 25, n. 17, p. 3389–3402, 1997. DOI: 10.1093/nar/25.17.3389.

CHO, Yunje *et al.* Crystal structure of a p53 tumor suppressor-DNA complex: understanding tumorigenic mutations. **Science**, v. 265, n. 5170, p. 346–355, 1994. DOI: 10.1126/science.8023157.

COFFILL, C. R. *et al.* The p53–Mdm2 interaction and the E3 ligase activity of Mdm2/Mdm4 are conserved from lampreys to humans. **Genes & Development**, v. 30, n. 3, p. 281–292, 2016. DOI: 10.1101/gad.274118.115.

DONEHOWER, Lawrence A. *et al.* Mice deficient for p53 are developmentally normal but susceptible to spontaneous tumours. **Nature**, v. 356, p. 215–221, 1992. DOI: 10.1038/356215a0.

DONEHOWER, L. A. *et al.* Integrated analysis of TP53 gene and pathway alterations in The Cancer Genome Atlas. **Cell Reports**, v. 28, n. 5, p. 1370–1384.e5, 2019. DOI: 10.1016/j.celrep.2019.07.001.

FISCHER, M. Conservation and divergence of the p53 gene regulatory network between mice and humans. **Oncogene**, v. 38, n. 21, p. 4095–4109, 2019. DOI: 10.1038/s41388-019-0706-9.

KATO, S. *et al.* Understanding the function–structure and function–mutation relationships of p53 tumor suppressor protein by high-resolution missense mutation analysis. **Proceedings of the National Academy of Sciences**, v. 100, n. 14, p. 8424–8429, 2003. DOI: 10.1073/pnas.1431692100.

NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION. **Homo sapiens tumor protein p53 (TP53), transcript variant 1, mRNA: NM_000546.6**. Bethesda, MD: NCBI, [s. d.]. Disponível em: https://www.ncbi.nlm.nih.gov/nuccore/NM_000546.6. Acesso em: 22 out. 2026.

VEPRINTSEV, D. B. *et al.* Core domain interactions in full-length p53 in solution. **Proceedings of the National Academy of Sciences**, v. 103, n. 7, p. 2115–2119, 2006. DOI: 10.1073/pnas.0511130103.

CAPÍTULO 10

BASES MOLECULARES DA APOPTOSE MITOCONDRIAL: ORGANIZAÇÃO E DIVERSIDADE ESTRUTURAL DE PROTEÍNAS DA FAMÍLIA BCL-2 E SUAS IMPLICAÇÕES NO CÂNCER

MOLECULAR BASIS OF MITOCHONDRIAL APOPTOSIS: ORGANIZATION AND STRUCTURAL DIVERSITY OF BCL-2 FAMILY PROTEINS AND THEIR IMPLICATIONS IN CANCER

DOI: <https://doi.org/10.56001/26.9786501906539.10>

Submetido em: 24/10/2026

Revisado em: 25/10/2026

Publicado em: 26/10/2026

Lais Albuquerque de Oliveira

Programa de Pós-Graduação em Ciências da Saúde, Universidade Federal da Grande

Dourados, Dourados, Mato Grosso do Sul, Brasil

<http://lattes.cnpq.br/2161403193721501>

Gleyce Hellen de Almeida de Souza

Programa de Pós-Graduação em Ciências da Saúde, Universidade Federal da Grande

Dourados, Dourados, Mato Grosso do Sul, Brasil

<http://lattes.cnpq.br/8163519978837401>

Ruana Carolina Cabral da Silva

Programa de Pós-Graduação em Ciências da Saúde, Universidade Federal da Grande

Dourados, Mato Grosso do Sul, Brasil

<http://lattes.cnpq.br/8634356371128001>

Resumo

A apoptose mitocondrial participa da manutenção da homeostase tecidual, e sua desregulação contribui para a sobrevivência de células tumorais. Este capítulo teve como objetivo discutir os fundamentos moleculares dessa via e sua relação com o câncer, integrando uma análise bioinformática comparativa das proteínas humanas BCL-2, BCL-XL e BAX. Foi realizado um estudo descritivo com sequências de

referência obtidas no UniProtKB, caracterização físico-química com Biopython, anotação de regiões conservadas pelo InterProScan e representação de modelos estruturais do AlphaFold Protein Structure Database no PyMOL. As proteínas apresentaram 239, 233 e 192 aminoácidos, respectivamente, com diferenças de massa molecular, ponto isoelétrico, hidropaticidade média e índice de instabilidade. Foram identificadas correspondências com os motivos BH1, BH2 e BH3 nas três sequências e com BH4 em BCL-2 e BCL-XL, além de segmentos transmembrana preditos próximos à extremidade carboxiterminal. A representação estrutural permitiu contextualizar espacialmente os motivos selecionados. Os achados evidenciam características compartilhadas e diferenças moleculares entre reguladores com funções complementares na apoptose. Conclui-se que a integração de informações de sequência, anotação e estrutura oferece uma base para compreender a organização da família BCL-2 e discutir sua relevância para a biologia do câncer e a investigação de estratégias de modulação da morte celular.

Palavras-Chave: Apoptose mitocondrial; Família BCL-2; Bioinformática; Estrutura proteica; Câncer.

Abstract

Mitochondrial apoptosis contributes to tissue homeostasis, and its dysregulation promotes tumor cell survival. This chapter aimed to discuss the molecular foundations of this pathway and its relationship with cancer, integrating a comparative bioinformatics analysis of human BCL-2, BCL-XL, and BAX proteins. A descriptive study was conducted using reference sequences retrieved from UniProtKB, physicochemical characterization with Biopython, conserved region annotation with InterProScan, and visualization of structural models from the AlphaFold Protein Structure Database using PyMOL. The proteins comprised 239, 233, and 192 amino acids, respectively, and differed in molecular weight, theoretical isoelectric point, average hydropathy, and instability index. Matches to BH1, BH2, and BH3 motifs were identified in all three sequences, whereas BH4 matches were identified in BCL-2 and BCL-XL. Predicted transmembrane segments were located near the carboxyl terminus of each protein. Structural visualization provided spatial context for the selected motifs. The findings highlight shared features and molecular differences among regulators with complementary roles in apoptosis. Integrating sequence, annotation, and structural information provides a foundation for understanding the organization of the BCL-2 family and discussing its relevance to cancer biology and the investigation of strategies to modulate cell death.

Keywords: Mitochondrial apoptosis; BCL-2 family; Bioinformatics; Protein structure; Cancer.

Introdução

A manutenção da integridade dos tecidos depende do equilíbrio entre proliferação, diferenciação e morte celular. Nesse contexto, a apoptose constitui um mecanismo de eliminação regulada de células que se tornaram desnecessárias ou apresentam alterações incompatíveis com a homeostase. Sua relevância para a biologia do câncer decorre da capacidade de restringir a sobrevivência de células potencialmente malignas. Quando os mecanismos de morte são comprometidos, células submetidas a estresse e portadoras de alterações moleculares podem persistir, favorecendo a progressão neoplásica. A resistência à morte celular integra, portanto, o conjunto de capacidades que sustentam o desenvolvimento tumoral, conforme discutido por Hanahan (2022).

A via mitocondrial da apoptose representa um importante ponto de integração de sinais relacionados à sobrevivência celular. Sua regulação envolve proteínas da família BCL-2, que controlam a permeabilização da membrana externa mitocondrial. Esse evento possibilita a liberação de componentes que participam da ativação da cascata de caspases

e da execução da morte celular. A decisão entre sobrevivência e apoptose depende das interações entre membros dessa família, cujas atividades podem favorecer ou restringir a permeabilização mitocondrial (Czabotar; Garcia-Saez, 2023).

Entre esses reguladores, BCL-2 e BCL-XL apresentam atividade predominantemente antiapoptótica, enquanto BAX atua como um dos efetores pró-apoptóticos. Proteínas com atividade de sobrevivência limitam a ação de componentes pró-apoptóticos por meio de interações específicas. Em resposta a estímulos apropriados, BAX pode sofrer mudanças conformacionais, associar-se à membrana mitocondrial e participar de sua permeabilização. Assim, proteínas pertencentes a uma mesma família molecular podem desempenhar funções distintas e complementares no controle do destino celular (Green, 2022).

A compreensão dessas funções exige considerar a relação entre sequência, organização estrutural e interação molecular. As regiões de homologia BCL-2, denominadas BH, participam da caracterização dos membros da família e de seus mecanismos de reconhecimento. Entretanto, a função de cada proteína também depende de seu contexto conformacional, de sua associação com membranas e da disponibilidade de parceiros de interação. A análise das regiões conservadas deve, portanto, ser integrada à organização tridimensional das proteínas, permitindo discutir como características compartilhadas se articulam com comportamentos funcionais diferentes (Czabotar; Garcia-Saez, 2023).

No câncer, alterações no controle da apoptose podem favorecer a manutenção de células que deveriam ser eliminadas e contribuir para a resistência ao tratamento. O papel central da família BCL-2 nesse processo também fundamenta seu interesse como alvo terapêutico. Entre as estratégias desenvolvidas estão os miméticos de BH3, que interferem na atividade de proteínas de sobrevivência e buscam restabelecer a suscetibilidade das células tumorais à apoptose. Essa abordagem evidencia a relevância de compreender as interfaces moleculares que regulam as interações entre os membros da família BCL-2.

Nesse cenário, análises bioinformáticas oferecem recursos para integrar informações sobre sequências proteicas, propriedades físico-químicas, motivos conservados e estruturas preditas. A comparação entre BCL-2, BCL-XL e BAX permite organizar essas características em uma perspectiva que aproxima genética molecular, biologia estrutural e patologia. Neste capítulo, essa abordagem é empregada como recurso para examinar a organização das proteínas e contextualizar suas funções na regulação da

apoptose mitocondrial, sem atribuir às predições computacionais o caráter de validação funcional.

Diante disso, este capítulo tem como objetivo discutir os fundamentos genéticos e moleculares da apoptose mitocondrial e sua relação com o câncer, com ênfase na organização estrutural e nas funções regulatórias da família BCL-2. Complementarmente, apresenta uma análise bioinformática comparativa de BCL-2, BCL-XL e BAX, considerando suas sequências, propriedades físico-químicas, motivos conservados e estruturas tridimensionais preditas.

Metodologia

Foi realizado um estudo bioinformático descritivo e comparativo, com dados públicos das proteínas humanas BCL-2, BCL-XL e BAX, complementado por discussão da literatura sobre apoptose mitocondrial e câncer. As sequências de referência foram obtidas no UniProtKB, pelos acessos P10415, Q07817 e Q07812, respectivamente, e armazenadas nos formatos FASTA e JSON. Para BCL-XL, foi utilizada a sequência canônica correspondente à isoforma longa codificada pelo gene *BCL2L1*. As análises foram conduzidas em ambiente Linux, com scripts em Python e utilização das bibliotecas Biopython, pandas e Matplotlib.

A caracterização físico-química foi realizada com o módulo ProteinAnalysis, do pacote Bio.SeqUtils.ProtParam, considerando comprimento da sequência, massa molecular, ponto isoelétrico teórico, índice de instabilidade, hidropaticidade média, expressa pelo parâmetro GRAVY, e aromaticidade. Os valores foram organizados em tabelas para comparação descritiva entre as proteínas. Esses parâmetros foram tratados como estimativas derivadas das sequências, sem equivalência direta com medidas experimentais de estabilidade, solubilidade ou atividade biológica. O alinhamento múltiplo das sequências foi realizado com o programa MAFFT, utilizando a opção de seleção automática de estratégia, e armazenado em formato FASTA.

A identificação de assinaturas de famílias proteicas, regiões conservadas e motivos foi realizada por meio do InterProScan. Foram examinadas as correspondências provenientes das bases Pfam, SMART e PROSITE, registrando-se a identificação da assinatura e suas posições inicial e final em cada sequência. Para padronizar a representação dos motivos BH, foram utilizadas as coordenadas das assinaturas PROSITE Patterns. As predições de segmentos transmembrana foram obtidas com o Phobius, conforme disponibilizadas na saída da análise. Correspondências de diferentes

métodos que abrangiam uma mesma região foram consideradas evidências complementares, evitando sua contabilização como domínios independentes. A ausência de correspondência com determinada assinatura foi interpretada no âmbito do método empregado.

Os modelos tridimensionais preditos foram obtidos no AlphaFold Protein Structure Database, por consulta à interface de programação de aplicações com os identificadores UniProt das proteínas. Antes da visualização, foram verificadas a correspondência entre as sequências dos modelos e as sequências utilizadas no estudo e a compatibilidade da numeração dos resíduos. As estruturas foram examinadas no PyMOL, em representação cartoon, com aplicação de cores às regiões correspondentes aos motivos BH. Para a figura dedicada a esses motivos, os demais resíduos foram ocultados apenas na visualização, preservando as coordenadas do modelo original. As estruturas foram apresentadas com orientações e escalas de exibição individuais, sem realização de sobreposição estrutural ou cálculo de desvio quadrático médio.

As figuras de arquitetura proteica foram elaboradas com Matplotlib, mantendo uma escala comum de posições de aminoácidos. As imagens estruturais foram renderizadas no PyMOL e organizadas em painéis identificados pelas letras A, B e C, correspondentes a BCL-2, BCL-XL e BAX. Para a representação estrutural, foram exibidos exclusivamente os resíduos com pLDDT igual ou superior a 70, correspondentes às faixas de confiança alta e muito alta, mantendo-se as coordenadas originais dos modelos.

Resultados e Discussão

Caracterização das sequências e propriedades físico-químicas

A análise comparativa evidenciou diferenças de comprimento e propriedades físico-químicas entre BCL-2, BCL-XL e BAX. BCL-2 apresentou a maior sequência, com 239 aminoácidos, seguida de BCL-XL, com 233, e BAX, com 192. As massas moleculares teóricas acompanharam essa distribuição, variando de aproximadamente 21,18 a 26,27 kDa. A Tabela 1 reúne os parâmetros calculados a partir das sequências de referência.

Tabela 1: Características das sequências e propriedades físico-químicas teóricas das proteínas analisadas.

Parâmetro	BCL-2	BCL-XL	BAX
Acesso UniProtKB	P10415	Q07817	Q07812
Comprimento (aminoácidos)	239	233	192
Massa molecular (kDa)	26,266	26,049	21,184
Ponto isoelétrico teórico	6,751	4,859	5,084
Índice de instabilidade	51,632	39,642	36,421
GRAVY	-0,1360	-0,3451	-0,0547
Aromaticidade	0,1088	0,1116	0,0938

Embora BCL-2 e BCL-XL tenham comprimentos próximos, seus pontos isoelétricos apresentaram diferenças. BCL-2 exibiu o maior valor, enquanto BCL-XL apresentou o menor. Esse resultado caracteriza diferenças na composição de resíduos ionizáveis das sequências. Entretanto, o ponto isoelétrico global não permite determinar a distribuição espacial das cargas nas superfícies de interação, cuja avaliação dependeria de análise eletrostática específica.

Os valores de GRAVY foram negativos nas três proteínas, indicando predomínio médio de características hidrofílicas segundo o índice empregado. BCL-XL apresentou o valor mais negativo, enquanto BAX apresentou o mais próximo de zero. A média global, contudo, pode coexistir com regiões localmente hidrofóbicas, como demonstrado pelas predições dos segmentos transmembrana apresentadas adiante.

BCL-2 apresentou o maior índice de instabilidade, enquanto BCL-XL e BAX exibiram valores menores. Esses dados descrevem o comportamento do estimador baseado na sequência e não constituem medidas de estabilidade termodinâmica ou de meia-vida intracelular. Da mesma forma, a aromaticidade ligeiramente maior de BCL-XL não estabelece maior afinidade por parceiros moleculares.

A interpretação funcional deve considerar que a regulação da apoptose envolve a interação entre proteínas, sua abundância e seu contexto celular. Singh, Letai e Sarosiek (2019) discutem esse controle como um equilíbrio dinâmico entre componentes pró-apoptóticos e de sobrevivência. Assim, os parâmetros calculados contribuem para caracterizar as sequências, mas não representam uma classificação direta de sua atividade apoptótica.

Organização das regiões conservadas e dos motivos BH

A análise pelo InterProScan identificou correspondências com a assinatura Pfam PF00452 nas três proteínas. Os intervalos foram 97–195 em BCL-2, 90–188 em BCL-XL e 63–158 em BAX. Também foram identificados motivos BH1, BH2 e BH3 nas três sequências pelas assinaturas PROSITE Patterns. Para BCL-2 e BCL-XL, foram registradas adicionalmente correspondências com a assinatura BH4. As coordenadas estão apresentadas na Tabela 2.

Tabela 2: Regiões e motivos identificados nas sequências de BCL-2, BCL-XL e BAX.

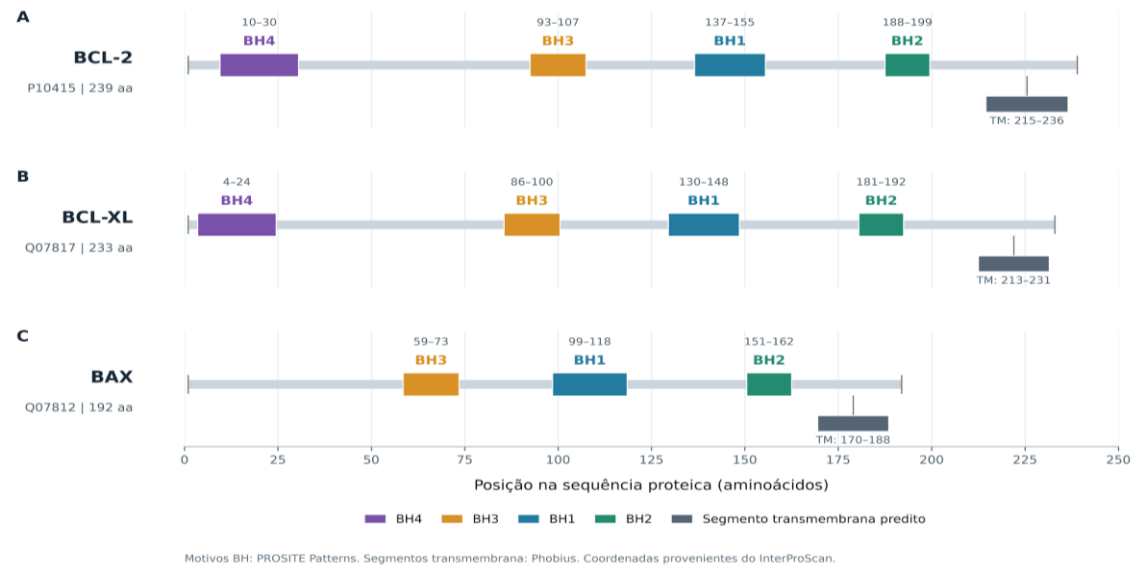
Região ou motivo	Método e assinatura	BCL-2	BCL-XL	BAX
Região da família BCL-2	Pfam PF00452	97–195	90–188	63–158
BH4	PROSITE PS01260	10–30	4–24	—
BH3	PROSITE PS01259	93–107	86–100	59–73
BH1	PROSITE PS01080	137–155	130–148	99–118
BH2	PROSITE PS01258	188–199	181–192	151–162

Nota: posições inclusivas, numeradas a partir do primeiro aminoácido. O traço indica ausência de correspondência com a assinatura indicada no arquivo analisado. A assinatura Pfam e os motivos PROSITE representam níveis distintos de anotação e não devem ser contabilizados como domínios independentes.

A distribuição dos motivos revelou uma organização linear semelhante entre BCL-2 e BCL-XL, com BH4 na região amino-terminal, seguido por BH3, BH1 e BH2. Em BAX, as correspondências BH3, BH1 e BH2 conservaram essa ordem relativa, embora em posições diferentes. A comparação descreve a arquitetura das assinaturas nas sequências, sem estabelecer, por si só, equivalência entre todos os resíduos abrangidos por cada intervalo.

A presença de motivos compartilhados em proteínas com funções distintas destaca que a identidade de uma assinatura não determina isoladamente o efeito sobre a morte celular. Qian *et al.* (2022) descrevem a família BCL-2 como um conjunto de reguladores com atividades pró-apoptóticas e antiapoptóticas, cuja atuação depende de interações moleculares específicas. Nesse sentido, o resultado permite discutir a conservação de regiões de reconhecimento em um sistema que apresenta diversidade funcional.

Figura 1: Organização dos motivos BH e dos segmentos transmembrana preditos nas proteínas analisadas. Representação das sequências de BCL-2, BCL-XL e BAX em escala comum, com indicação dos intervalos identificados por PROSITE Patterns e Phobius. As cores distinguem os motivos e o segmento transmembrana predito.



Segmentos transmembrana preditos e contexto estrutural

O Phobius identificou um segmento transmembrana próximo à extremidade carboxiterminal de cada proteína. Os intervalos compreenderam os resíduos 215–236 em BCL-2, 213–231 em BCL-XL e 170–188 em BAX, conforme a Tabela 3.

Tabela 3: Segmentos transmembrana preditos pelo Phobius.

Proteína	Intervalo de resíduos	Comprimento do segmento (aminoácidos)
BCL-2	215–236	22
BCL-XL	213–231	19
BAX	170–188	19

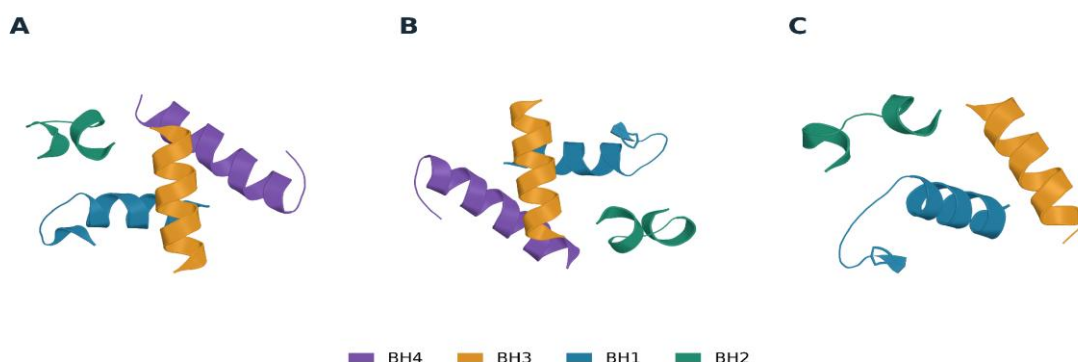
Fonte: elaboração própria, com base nas predições do Phobius disponibilizadas na saída do InterProScan.

Esses resultados demonstram que os valores globais negativos de GRAVY coexistem com segmentos localizados compatíveis com associação à membrana. A análise regional acrescenta, portanto, uma informação que não seria recuperada apenas pela média de hidropaticidade da sequência completa. As predições descrevem características dos segmentos, sem comprovar sua inserção ou orientação em uma membrana biológica.

Essa distinção é relevante para BAX, cuja participação na apoptose envolve mudanças no estado de ativação e na associação mitocondrial. A literatura relaciona a função dos membros da família BCL-2 à integração entre interações proteicas e processos de regulação celular, o que contextualiza a presença de segmentos hidrofóbicos sem transformar a predição em uma demonstração de atividade (Singh; Letai; Sarosiek, 2019).

Para complementar a arquitetura linear, os intervalos dos motivos BH foram selecionados nos modelos tridimensionais preditos. A Figura 2 apresenta exclusivamente esses segmentos, preservando suas coordenadas originais e ocultando os demais resíduos para facilitar a visualização. Essa representação permite localizar os motivos no contexto espacial dos modelos, mas não corresponde a estruturas isoladas experimentalmente.

Figura 2: Localização espacial dos motivos BH nas estruturas preditas de BCL-2 (A), BCL-XL (B) e BAX (C).



Integração dos achados e implicações para o câncer

A comparação reuniu características compartilhadas e diferenças entre os três alvos. As proteínas apresentaram correspondência com a família BCL-2 e motivos BH1, BH2 e BH3, acompanhados por diferenças de comprimento, composição físico-química e posição das assinaturas. Esses resultados sustentam uma abordagem integrada: a sequência informa a composição e a distribuição das regiões, enquanto sua representação estrutural auxilia a contextualizar espacialmente os motivos selecionados.

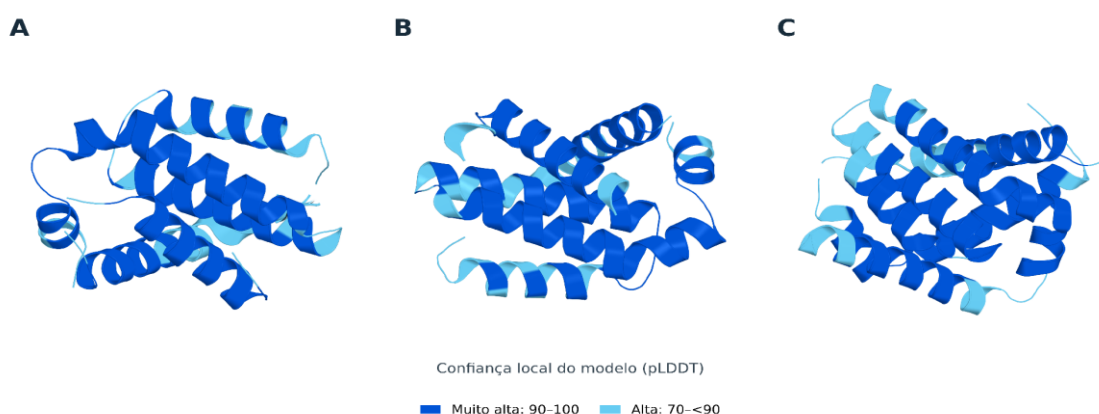
No contexto do câncer, a relevância dessas proteínas está relacionada ao controle da suscetibilidade celular à apoptose. Qian *et al.* (2022) discutem tanto sua participação na tumorigênese e na resposta aos tratamentos quanto o desenvolvimento de estratégias dirigidas a membros antiapoptóticos. As análises apresentadas oferecem uma base

molecular para essa discussão, sem medir expressão tumoral, dependência funcional ou resposta farmacológica.

Complementarmente à localização dos motivos BH, foram representadas as regiões de alta e muito alta confiança dos modelos estruturais de BCL-2, BCL-XL e BAX (Figura 3). A visualização foi restrita aos resíduos com pLDDT ≥ 70 , permitindo destacar os segmentos com maior confiança preditiva local. As coordenadas originais foram preservadas, e os demais resíduos foram ocultados apenas para apresentação.

O principal aporte do estudo consiste em articular informações de sequência, anotação e estrutura para examinar proteínas com funções complementares na apoptose mitocondrial. A organização comparativa favorece a compreensão de como regiões conservadas podem integrar mecanismos regulatórios distintos e oferece um ponto de partida para investigações específicas sobre variantes, superfícies de interação e modulação da morte celular no câncer.

Figura 3: Regiões de alta e muito alta confiança nos modelos estruturais preditos de BCL-2 (A), BCL-XL (B) e BAX (C).



Considerações Finais

A análise comparativa de BCL-2, BCL-XL e BAX permitiu integrar características das sequências, propriedades físico-químicas, motivos conservados e representações estruturais, ampliando a compreensão da organização molecular de reguladores da apoptose mitocondrial. A identificação de motivos BH compartilhados, associada às diferenças de comprimento, composição e distribuição das regiões anotadas, evidencia a diversidade molecular de proteínas que desempenham funções complementares no controle da sobrevivência e da morte celular.

A integração entre arquitetura linear e localização dos motivos nas estruturas preditas constitui um recurso para compreender as relações entre conservação e diversidade funcional na família BCL-2. Nesse contexto, as análises bioinformáticas fornecem uma base para discutir os mecanismos de regulação da apoptose e orientar investigações sobre variantes e interações moleculares.

Conclui-se que o estudo da família BCL-2 estabelece uma conexão relevante entre genética, biologia estrutural e patologia do câncer. A caracterização desses reguladores contribui para fundamentar a compreensão dos processos de sobrevivência tumoral e do interesse terapêutico na modulação da apoptose mitocondrial.

Declaração de Uso de Inteligência Artificial Generativa

Durante a elaboração deste capítulo, foi utilizada a ferramenta ChatGPT, da OpenAI, como apoio à organização e à revisão textual, à elaboração e ao ajuste de scripts para análises bioinformáticas e à preparação de representações gráficas e estruturais com base nos dados e modelos analisados. Os autores revisaram o conteúdo, os códigos, as referências e a interpretação dos resultados, assumindo integral responsabilidade pela precisão, originalidade e integridade científica da versão final.

Referências

- CZABOTAR, Peter E.; GARCIA-SAEZ, Ana J. Mechanisms of BCL-2 family proteins in mitochondrial apoptosis. **Nature Reviews Molecular Cell Biology**, v. 24, p. 732–748, 2023. DOI: <https://doi.org/10.1038/s41580-023-00629-4>.
- GREEN, Douglas R. The mitochondrial pathway of apoptosis part II: the BCL-2 protein family. **Cold Spring Harbor Perspectives in Biology**, v. 14, n. 6, a041046, 2022. DOI: <https://doi.org/10.1101/cshperspect.a041046>.
- HANAHAN, Douglas. Hallmarks of cancer: new dimensions. **Cancer Discovery**, v. 12, n. 1, p. 31–46, 2022. DOI: <https://doi.org/10.1158/2159-8290.CD-21-1059>.
- QIAN, S. et al. The role of BCL-2 family proteins in regulating apoptosis and cancer therapy. **Frontiers in Oncology**, v. 12, 985363, 2022. DOI: <https://doi.org/10.3389/fonc.2022.985363>.
- SINGH, R.; LETAI, A.; SAROSIEK, K. Regulation of apoptosis in health and disease: the balancing act of BCL-2 family proteins. **Nature Reviews Molecular Cell Biology**, v. 20, p. 175–193, 2019. DOI: <https://doi.org/10.1038/s41580-018-0089-8>.

CAPÍTULO 11

LÚPUS ERITEMATOSO SISTÊMICO: BASES MOLECULARES DA IMUNIDADE INATA E VIAS DE SINALIZAÇÃO NA AUTOIMUNIDADE

SYSTEMIC LUPUS ERYTHEMATOSUS: MOLECULAR BASIS OF INNATE IMMUNITY AND SIGNALING PATHWAYS IN AUTOIMMUNITY

DOI: <https://doi.org/10.56001/26.9786501906539.11>

Submetido em: 24/10/2026

Revisado em: 25/10/2026

Publicado em: 26/10/2026

Ruana Carolina Cabral da Silva

Programa de Pós-Graduação em Ciências da Saúde, Universidade Federal da Grande

Dourados, Mato Grosso do Sul, Brasil

<http://lattes.cnpq.br/8634356371128001>

Gleyce Hellen de Almeida de Souza

Programa de Pós-Graduação em Ciências da Saúde, Universidade Federal da Grande

Dourados, Dourados, Mato Grosso do Sul, Brasil

<http://lattes.cnpq.br/8163519978837401>

Jonatas Thiago dos Santos Tenan

Programa de Pós-Graduação em Educação, Trabalho e Inovação em Medicina

Escola Multicampi de Ciências Médicas do Rio Grande do Norte, Caicó, Rio Grande do

Norte, Brasil

<http://lattes.cnpq.br/6848971915302845>

Lais Albuquerque de Oliveira

Programa de Pós-Graduação em Ciências da Saúde, Universidade Federal da Grande

Dourados, Dourados, Mato Grosso do Sul, Brasil

<http://lattes.cnpq.br/2161403193721501>

Mariana Carvalho Sturaro

Programa de Pós-Graduação em Biociências e Biotecnologia, Universidade de São Paulo, Ribeirão Preto, São Paulo, Brasil

<http://lattes.cnpq.br/3041100311056029>

Rousilândia de Araujo Silva

Programa de Pós-Graduação em Ciências Farmacêuticas, Universidade Federal de Pernambuco, Recife, Pernambuco, Brasil

<http://lattes.cnpq.br/8194570266422228>

Maria Cidinaria Silva Alves

Laboratório de Parasitologia Molecular, Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, São Paulo, Brasil

<http://lattes.cnpq.br/1633031262017450>

José Eduardo Souza Echeverria

Programa de Pós-Graduação em Ciências da Saúde, Universidade Federal da Grande Dourados, Dourados, Mato Grosso do Sul, Brasil

<http://lattes.cnpq.br/1188506108074331>

Fabíola Lucini

Programa de Pós-Graduação em Ciências da Saúde, Universidade Federal da Grande Dourados, Dourados, Mato Grosso do Sul, Brasil

<http://lattes.cnpq.br/1608714419112836>

Resumo

O lúpus eritematoso sistêmico envolve alterações na regulação imunológica, com participação de mecanismos de reconhecimento de ácidos nucleicos e sinalização inflamatória. Este estudo teve como objetivo caracterizar bioinformaticamente as proteínas humanas cGAS, STING, TBK1 e IRF3 e integrar suas características moleculares ao mapeamento funcional do eixo cGAS–STING, contextualizando sua relevância para a doença. Foi realizado um estudo descritivo com sequências do UniProtKB, caracterização físico-química com Biopython, anotação pelo InterProScan, análise de modelos do AlphaFold Protein Structure Database e mapeamento de eventos no Reactome. As proteínas apresentaram 522, 379, 729 e 427 aminoácidos, respectivamente. Foram identificadas regiões relacionadas à atividade nucleotidiltransferase em cGAS, domínios transmembrana e de ligação ao ligante em STING, domínios quinase, semelhante à ubiquitina e coiled-coil em TBK1, além de regiões características de IRF3 e de ligação ao DNA. TBK1 apresentou o maior pLDDT médio, de 89,698, e a maior proporção de resíduos com pLDDT ≥ 70 , correspondente a 90,947%. O mapeamento funcional conectou os alvos ao reconhecimento de DNA, à produção de cGAMP, ao recrutamento de componentes da sinalização e à ativação de IRF3. Conclui-se que a integração entre anotações de sequência, modelos estruturais e eventos funcionais oferece uma base para compreender a organização desse eixo e discutir sua participação, dependente do contexto biológico, na imunopatologia do lúpus.

Palavras-Chave: Lúpus eritematoso sistêmico; Imunidade inata; Sinalização celular; Via cGAS–STING; Bioinformática.

Abstract

Systemic lupus erythematosus involves disrupted immune regulation, including mechanisms of nucleic acid recognition and inflammatory signaling. This study aimed to characterize human cGAS, STING, TBK1, and IRF3 proteins using bioinformatics and integrate their molecular features with functional mapping of the cGAS–STING axis, contextualizing its relevance to the disease. A descriptive study was conducted using UniProtKB sequences, physicochemical characterization with Biopython, annotation with InterProScan, analysis of models from the AlphaFold Protein Structure Database, and event mapping in Reactome. The proteins comprised 522, 379, 729, and 427 amino acids, respectively. Identified features included nucleotidyltransferase-related regions in cGAS; transmembrane and ligand-binding domains in STING; kinase, ubiquitin-like, and coiled-coil domains in TBK1; and DNA-binding and IRF3-specific regions in IRF3. TBK1 showed the highest mean pLDDT, at 89.698, and the largest proportion of residues with pLDDT ≥ 70 , at 90.947%. Functional mapping connected the targets to DNA recognition, cGAMP production, recruitment of signaling components, and IRF3 activation. Integrating sequence annotations, structural models, and functional events provides a foundation for understanding the organization of this axis and discussing its context-dependent involvement in lupus immunopathology.

Keywords: Systemic lupus erythematosus; Innate immunity; Cell signaling; cGAS–STING pathway; Bioinformatics.

Introdução

O lúpus eritematoso sistêmico (LES) é uma doença autoimune crônica caracterizada pela heterogeneidade clínica e pelo comprometimento de diferentes órgãos e tecidos. Sua fisiopatologia envolve a interação entre suscetibilidade genética, fatores ambientais e alterações na regulação das respostas imunes inata e adaptativa. A produção de autoanticorpos, a formação de imunocomplexos e a persistência de processos inflamatórios integram esse cenário, no qual diferentes mecanismos moleculares contribuem para a perda da tolerância imunológica e o dano tecidual. Conforme discutem Arnaud, Chasset e Martin (2024), compreender essa diversidade de mecanismos é fundamental para relacionar as manifestações da doença às vias biológicas envolvidas.

Entre os processos associados à imunopatogênese do LES, destaca-se o reconhecimento de ácidos nucleicos por sensores da imunidade inata. Alterações no processamento, na degradação ou na compartimentalização desses materiais podem favorecer sua disponibilidade para sistemas de detecção imunológica. Nesse contexto, moléculas derivadas do próprio organismo podem participar da indução de respostas inflamatórias. Caielli, Wan e Pascual (2023) destacam a importância dos ácidos nucleicos com potencial de induzir interferons e de sua relação com a ativação de componentes da imunidade inata e adaptativa na doença.

Os interferons do tipo I ocupam uma posição relevante nessa rede, conectando o reconhecimento de sinais moleculares à regulação da expressão gênica e das funções de

diferentes células imunes. Sua produção e seus efeitos integram circuitos capazes de sustentar a ativação imunológica. Entretanto, a resposta de interferon no LES resulta da participação de diferentes sensores e mecanismos regulatórios, devendo ser compreendida como parte de uma rede de sinalização. Essa perspectiva permite examinar vias específicas sem reduzir a complexidade da doença a um único processo molecular (Caielli; Wan; Pascual, 2023).

Nesse cenário, o eixo cGAS–STING representa uma interface entre o reconhecimento de DNA e a sinalização inflamatória. A enzima cGAS, denominada sintase de GMP–AMP cíclico, reconhece DNA e promove a produção do segundo mensageiro 2'3'-cGAMP. Esse mensageiro se liga à proteína STING, desencadeando eventos que conectam o reconhecimento inicial à ativação de componentes intracelulares. Zhang e Zhang (2025) ressaltam que a regulação dessa via envolve fatores ativadores e inibidores, organização subcelular e interações com outros processos, produzindo respostas dependentes do contexto biológico.

A transmissão do sinal envolve a participação da quinase TBK1 e do fator regulador de interferon 3, IRF3. STING atua como plataforma para o recrutamento desses componentes, favorecendo a fosforilação de IRF3 por TBK1 e sua participação na regulação transcricional de genes da resposta imune. Os eventos descritos na base Reactome permitem organizar funcionalmente esses alvos, conectando a detecção de DNA à indução de interferon tipo I. Assim, cGAS, STING, TBK1 e IRF3 representam etapas complementares de um circuito de sinalização de interesse para o estudo da imunidade inata.

A investigação desse circuito também exige considerar a organização molecular de seus componentes. Regiões catalíticas, domínios de ligação, segmentos transmembrana e regiões regulatórias contribuem de maneiras distintas para o reconhecimento de moléculas, a transmissão de sinais e o controle da transcrição. Os avanços em biologia estrutural têm ampliado a compreensão dos mecanismos de funcionamento e regulação da via cGAS–STING, aproximando características moleculares de seus diferentes desfechos celulares (Zhang; Zhang, 2025).

Nesse contexto, a bioinformática oferece recursos para integrar informações de sequências, anotações funcionais e modelos estruturais. A caracterização desses quatro componentes permite examinar suas arquiteturas individuais e situá-las em eventos documentados de sinalização. Essa abordagem estabelece uma conexão entre genética

molecular, estrutura proteica e imunologia, oferecendo uma base descritiva para discutir mecanismos associados ao LES.

Diante disso, este estudo teve como objetivo caracterizar bioinformaticamente as proteínas humanas cGAS, STING, TBK1 e IRF3, considerando suas propriedades físico-químicas, domínios e modelos estruturais, e integrar essas informações ao mapeamento funcional do eixo cGAS–STING. Complementarmente, buscou-se discutir a relevância dessa sinalização no contexto da imunidade inata e da fisiopatologia do lúpus eritematoso sistêmico.

Metodologia

Foi realizado um estudo bioinformático descritivo de quatro componentes humanos do eixo de sinalização cGAS–STING, complementado pelo mapeamento funcional da via e pela discussão de sua relação com o lúpus eritematoso sistêmico. Foram selecionadas as proteínas cGAS, STING, TBK1 e IRF3, correspondentes aos acessos UniProtKB Q8N884, Q86WV6, Q9UHD2 e Q14653, respectivamente. As análises foram conduzidas em ambiente Linux, utilizando scripts em Python e as bibliotecas requests, Biopython, pandas e Matplotlib.

As sequências canônicas e suas anotações foram obtidas no UniProtKB, por acesso programático, e armazenadas nos formatos FASTA e JSON. Foram verificadas a identificação da espécie humana e a concordância entre as sequências disponibilizadas nos dois formatos. Para assegurar a rastreabilidade, foram registrados a data de obtenção, os metadados de atualização dos registros e o identificador de integridade SHA-256 de cada sequência. As anotações referentes a domínios, regiões, motivos, segmentos transmembrana, sítios funcionais e modificações de resíduos foram extraídas dos registros JSON e organizadas em tabela, preservando as informações de evidência disponibilizadas pela base.

A caracterização físico-química foi realizada com o módulo ProteinAnalysis, do pacote Bio.SeqUtils.ProtParam. Foram calculados comprimento da sequência, massa molecular teórica, ponto isoelétrico teórico, índice de instabilidade, hidropaticidade média, expressa pelo parâmetro GRAVY, e aromaticidade. Os parâmetros foram comparados de maneira descritiva e interpretados como estimativas derivadas da composição das sequências, sem equivalência direta com medidas experimentais de estabilidade, localização ou atividade proteica.

As sequências foram submetidas ao InterProScan para identificação de assinaturas de famílias proteicas, domínios e regiões conservadas. Os resultados foram obtidos em formato TSV e organizados por proteína, método, identificador da assinatura, descrição e coordenadas inicial e final. Foram examinadas as correspondências das bases Pfam e SMART, além das predições de segmentos transmembrana disponibilizadas pelo Phobius. Registros redundantes e correspondências sobrepostas foram avaliados para evitar sua contabilização como domínios independentes. Na representação da arquitetura proteica, foram utilizadas as coordenadas Pfam para as regiões selecionadas, exceto para o domínio quinase de TBK1, cujo intervalo foi representado conforme a identificação pelo SMART. Os segmentos transmembrana de STING foram apresentados em faixa separada das assinaturas de domínio. Por se tratar de componentes com funções e arquiteturas distintas, não foi realizado alinhamento múltiplo entre as quatro sequências completas.

Os modelos tridimensionais foram obtidos no AlphaFold Protein Structure Database por consulta programática com os identificadores UniProtKB. Foram selecionados modelos com sequência idêntica à referência utilizada no estudo. Após o download dos arquivos PDB, foram verificadas a sequência da cadeia e a numeração dos resíduos. A confiança local foi examinada pelos valores de pLDDT armazenados no campo B dos arquivos, utilizando os átomos de carbono alfa como referência. Foram calculados o pLDDT médio e os percentuais de resíduos com valores iguais ou superiores a 70 e a 90.

A visualização estrutural foi realizada no PyMOL, em representação cartoon. Para a figura comparativa, foram exibidos exclusivamente os resíduos com pLDDT ≥ 70 , diferenciando-se as faixas de confiança alta, de 70 a menos de 90, e muito alta, igual ou superior a 90. Os demais resíduos foram ocultados apenas na visualização, sem exclusão de coordenadas dos modelos originais. As estruturas foram apresentadas com orientações e escalas de exibição individuais, sem sobreposição estrutural. O pLDDT foi interpretado como uma medida de confiança local da predição, sem inferência direta de estabilidade proteica ou de precisão da orientação relativa entre domínios.

O mapeamento funcional foi realizado com base em eventos anotados no Reactome, abrangendo o reconhecimento de DNA citosólico por cGAS, a produção de cGAMP, sua ligação a STING, o recrutamento de TBK1 e IRF3, a fosforilação de IRF3 e sua participação na indução de interferon tipo I. Foram registrados os identificadores e os endereços dos eventos consultados, e as relações foram organizadas em tabela e

representadas em um esquema de sinalização. Essa etapa teve caráter descritivo, sem análise de enriquecimento estatístico, uma vez que os alvos foram previamente selecionados por sua participação na via.

Resultados e Discussão

Organização molecular dos componentes analisados

A caracterização das proteínas humanas cGAS, STING, TBK1 e IRF3 evidenciou arquiteturas distintas, compatíveis com sua participação em etapas complementares da sinalização imune inata. TBK1 apresentou a maior sequência, com 729 aminoácidos, seguida por cGAS, com 522, IRF3, com 427, e STING, com 379. A anotação pelo InterProScan identificou regiões relacionadas à atividade enzimática, à associação com membranas, ao reconhecimento de ligantes e à ligação ao DNA. A Tabela 1 apresenta as principais correspondências selecionadas.

Tabela 1: Comprimento e principais regiões identificadas nos componentes do eixo cGAS–STING.

Proteína	Comprimento (aa)	Região identificada	Intervalo	Método e assinatura
cGAS	522	Domínio nucleotidiltransferase semelhante a Mab-21	214–389	Pfam PF03281
cGAS	522	Domínio semelhante a HhH/H2TH	405–509	Pfam PF20266
STING	379	Domínio transmembrana de STING	45–152	Pfam PF23417
STING	379	Domínio de ligação ao ligante	154–337	Pfam PF15009
TBK1	729	Domínio quinase	9–310	SMART SM00220
TBK1	729	Domínio semelhante à ubiquitina	308–395	Pfam PF18396
TBK1	729	Domínio coiled-coil 1	400–655	Pfam PF18394
IRF3	427	Domínio de ligação ao DNA	8–110	Pfam PF00605
IRF3	427	Região característica de IRF3	202–379	Pfam PF10401

Legenda: aa: aminoácidos. Coordenadas inclusivas. As denominações correspondem às assinaturas identificadas; a tabela reúne regiões selecionadas, sem representar todas as anotações disponíveis.

Em cGAS, as duas correspondências Pfam selecionadas localizaram-se na porção carboxiterminal da sequência. O SMART identificou uma região mais abrangente, entre os resíduos 213 e 513, associada à família MAB21L/cGAS-like. A sobreposição entre

essas anotações demonstra que diferentes métodos podem reconhecer níveis distintos de organização de uma mesma proteína. Assim, a correspondência SMART não foi contabilizada como um domínio adicional às duas regiões Pfam.

Em STING, foram identificadas uma assinatura de domínio transmembrana e outra de ligação ao ligante. Essa organização permite distinguir uma região associada à arquitetura de membrana de outra relacionada ao reconhecimento molecular. Complementarmente, o Phobius identificou quatro segmentos transmembrana, apresentados na Tabela 2. Os intervalos individuais não coincidem integralmente com a assinatura Pfam porque os métodos descrevem características diferentes da sequência.

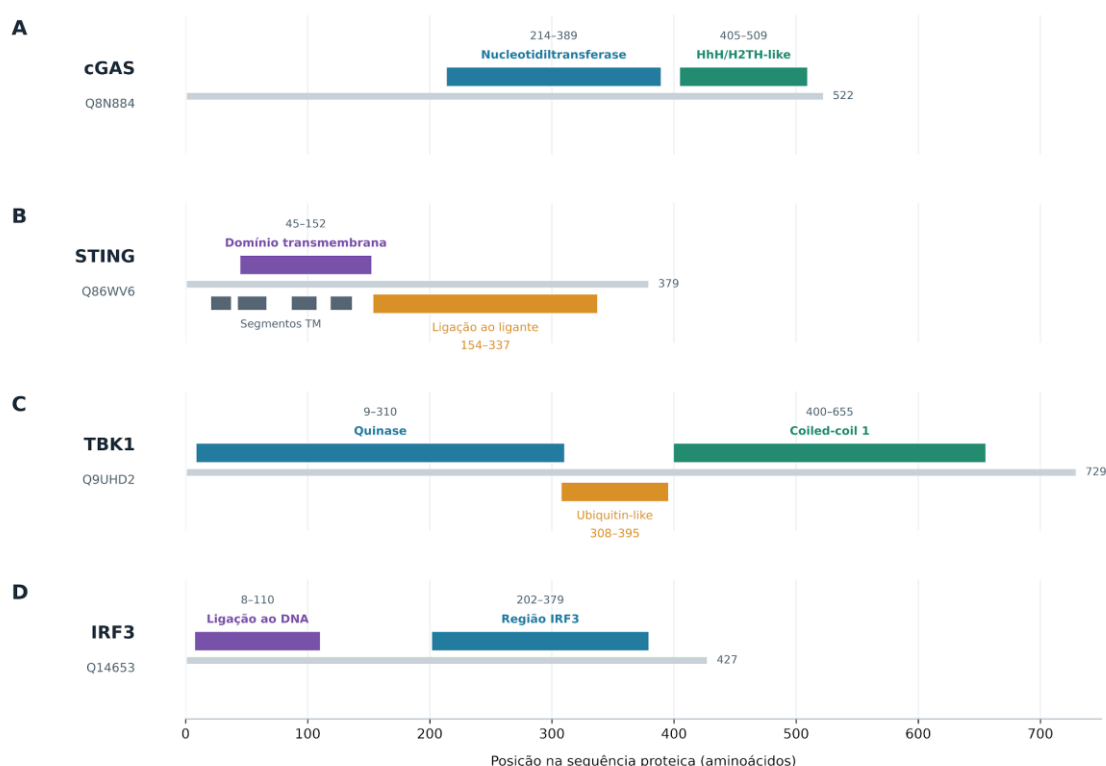
Tabela 2: Segmentos transmembrana preditos em STING pelo Phobius.

Segmento	Intervalo	Comprimento (aa)
TM1	21–37	17
TM2	43–66	24
TM3	87–107	21
TM4	119–136	18

TBK1 apresentou uma organização com domínio quinase, região semelhante à ubiquitina e domínio coiled-coil. O domínio quinase foi delimitado pelo SMART entre os resíduos 9 e 310, enquanto o Pfam identificou uma correspondência mais restrita, entre 13 e 232. Para a representação gráfica, foi empregado o intervalo SMART, mantendo sua pequena sobreposição com a região semelhante à ubiquitina, iniciada no resíduo 308. Essa sobreposição decorre dos limites atribuídos pelos métodos e não comprova uma fronteira estrutural compartilhada.

Em IRF3, a identificação de uma região amino-terminal de ligação ao DNA e de outra assinatura característica de IRF3 permite organizar sua arquitetura em segmentos com anotações distintas. A relação dessa organização com a função transcricional deve ser estabelecida em conjunto com as evidências funcionais disponíveis, pois a identificação de uma assinatura não demonstra, isoladamente, o estado de ativação da proteína. A Figura 1 reúne essas regiões em escala comum de aminoácidos, permitindo comparar sua distribuição sem pressupor homologia entre as quatro proteínas completas.

Figura 1: Arquitetura de domínios dos componentes da via cGAS–STING. Representação de cGAS (A), STING (B), TBK1 (C) e IRF3 (D) em escala comum de aminoácidos. As regiões foram delimitadas pelas assinaturas Pfam, exceto o domínio quinase de TBK1, representado pelo intervalo SMART. Os segmentos transmembrana preditos pelo Phobius são apresentados separadamente em STING.



Confiança local dos modelos estruturais

Os modelos obtidos no AlphaFold Protein Structure Database apresentaram diferentes distribuições de confiança local. TBK1 exibiu o maior pLDDT médio, de 89,698, e a maior proporção de resíduos com pLDDT ≥ 70 , correspondente a 90,947%. cGAS apresentou o menor pLDDT médio, de 76,774, e 68,582% dos resíduos nessa faixa. Os valores estão apresentados na Tabela 3.

Tabela 3: Indicadores de confiança local dos modelos estruturais analisados.

Proteína	Resíduos	pLDDT médio	Resíduos com pLDDT ≥ 70 (%)	Resíduos com pLDDT ≥ 90 (%)
cGAS	522	76,774	68,582	57,471
STING	379	83,738	84,433	51,451
TBK1	729	89,698	90,947	71,605
IRF3	427	80,259	78,454	54,098

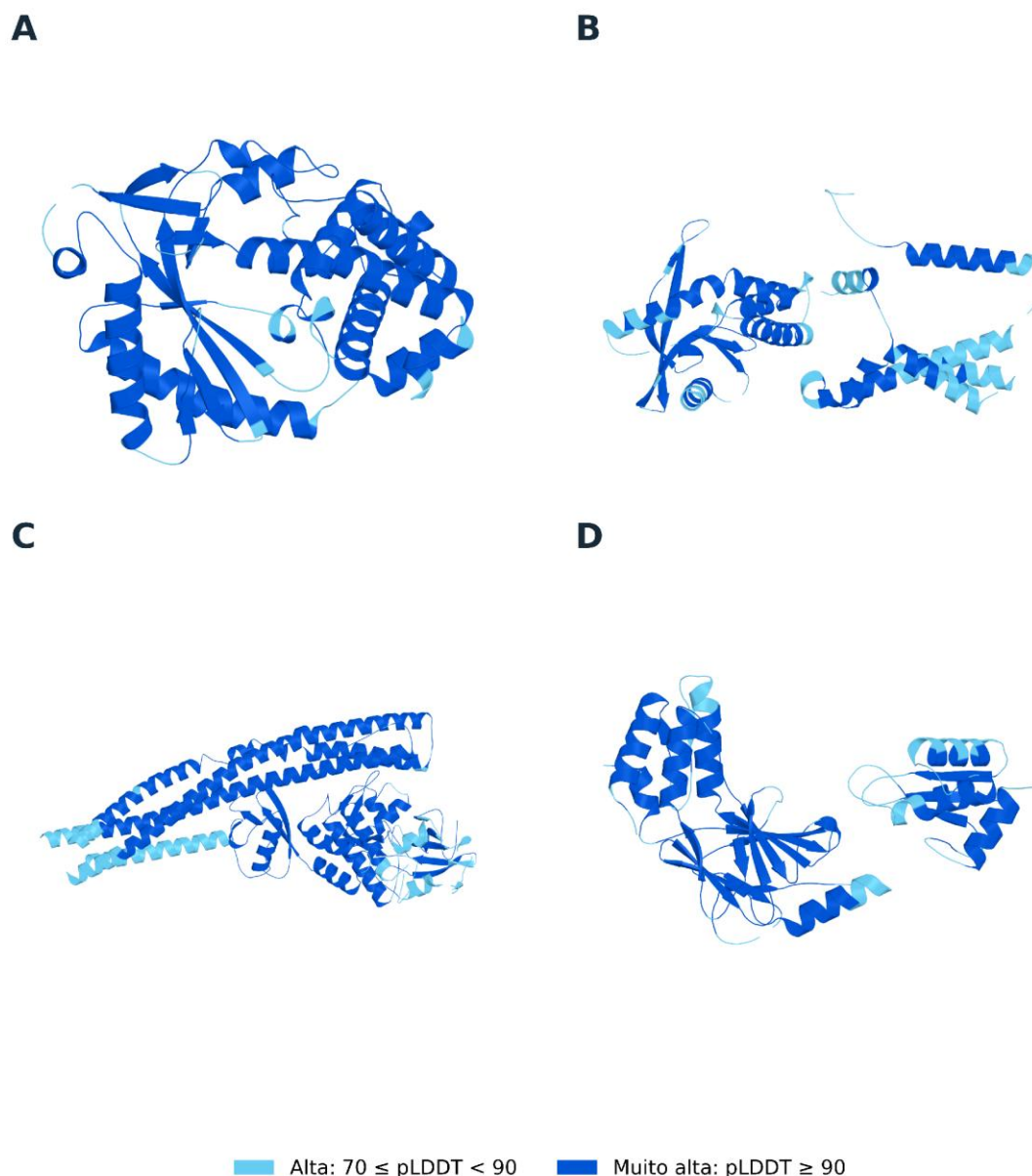
Fonte: elaboração própria, com base nos valores de pLDDT dos modelos obtidos no AlphaFold Protein Structure Database. Nota: o percentual com pLDDT ≥ 90 está incluído no percentual com pLDDT ≥ 70 .

As diferenças entre média e distribuição dos valores também são relevantes. Embora STING tenha apresentado pLDDT médio superior ao de cGAS, sua proporção de resíduos com pLDDT ≥ 90 foi menor. Portanto, uma única média não descreve integralmente a confiança ao longo da sequência.

Para a Figura 2, foram selecionados exclusivamente os resíduos com pLDDT ≥ 70 . O filtro permite destacar regiões com confiança alta ou muito alta, preservando suas coordenadas nos modelos originais. Contudo, a imagem resultante apresenta apenas parte de cada sequência: os segmentos ocultados continuam pertencendo à proteína e não devem ser interpretados como biologicamente ausentes.

A representação também deve distinguir confiança local de organização global. Os valores apresentados não constituem medidas de estabilidade termodinâmica, atividade catalítica ou intensidade de sinalização. Além disso, a exibição individual dos modelos não permite inferir a formação de um complexo entre os quatro componentes.

Figura 2: Regiões de alta e muito alta confiança dos modelos estruturais de cGAS (A), STING (B), TBK1 (C) e IRF3 (D). Representação em cartoon, exibindo exclusivamente resíduos com pLDDT ≥ 70 . Azul claro indica confiança alta e azul escuro, muito alta. Os demais resíduos foram ocultados apenas na visualização. As orientações e escalas de exibição são individuais.



Integração dos alvos no eixo de sinalização

O mapeamento no Reactome conectou os quatro componentes em uma sequência funcional que abrange reconhecimento de DNA, produção de um segundo mensageiro,

organização da sinalização e regulação transcricional. As relações selecionadas estão sintetizadas na Tabela 4.

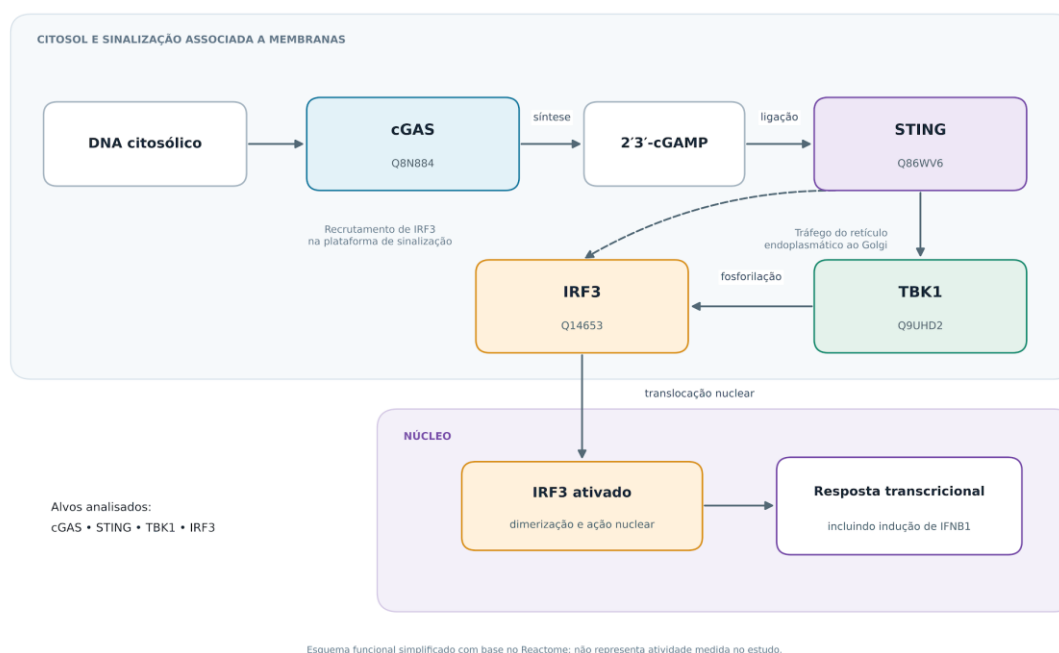
Tabela 4: Eventos selecionados para o mapeamento funcional da via.

Componente ou relação	Evento representado	Identificador Reactome
DNA–cGAS	Reconhecimento de DNA citosólico	R-HSA-3244647
cGAS–cGAMP	Produção de cGAMP	R-HSA-3244614
cGAMP–STING	Ligação do mensageiro a STING	R-HSA-3244643
STING–TBK1–IRF3	Recrutamento dos componentes	R-HSA-1834939
TBK1–IRF3	Fosforilação de IRF3	R-HSA-2396007
IRF3	Participação na indução de interferon tipo I	R-HSA-3270619

Nesse encadeamento, a atividade de cGAS está conectada a STING por intermédio do cGAMP. Portanto, a relação entre esses dois alvos não deve ser representada simplesmente como uma interação física direta. A jusante, STING participa do recrutamento de TBK1 e IRF3, e TBK1 promove a fosforilação de IRF3. Essa organização evidencia a participação de diferentes modalidades de comunicação molecular dentro do mesmo eixo.

O mapeamento complementa a anotação dos domínios ao situar as proteínas em eventos funcionais documentados. Enquanto a arquitetura proteica descreve regiões identificadas nas sequências, o esquema da via organiza relações estabelecidas nas fontes consultadas. A integração dessas abordagens oferece uma interpretação molecular coerente, sem transformar a anotação computacional em demonstração experimental de sinalização.

Figura 3: Organização funcional do eixo cGAS–STING–TBK1–IRF3. O de DNA por cGAS promove a produção de cGAMP, que se liga a STING. A sinalização envolve o recrutamento de TBK1 e IRF3, a fosforilação de IRF3 e sua participação na regulação transcricional, incluindo a indução de *IFNB1*. O esquema reúne eventos selecionados e não representa atividade medida no estudo.



Relação com a autoimunidade e o lúpus eritematoso sistêmico

A conexão entre reconhecimento de DNA e indução de interferon oferece um fundamento para investigar esse eixo na autoimunidade. Chang (2024) discute sua relevância no LES e o interesse em estratégias de modulação da via. Essa perspectiva contextualiza a seleção dos quatro componentes, cuja caracterização permite examinar etapas moleculares potencialmente relevantes para a sinalização inflamatória.

Entretanto, a participação da via deve ser interpretada conforme o mecanismo investigado. Motwani *et al.* (2021), em modelos murinos de lúpus dependentes de receptores Toll-like, observaram que a deficiência de cGAS ou STING não impediu a doença e identificaram efeitos compatíveis com contenção da atividade autoimune nesses contextos. Esses resultados demonstram que o papel desse eixo pode variar entre modelos, reforçando a importância de relacionar cada evidência às condições em que foi produzida.

No presente estudo, os resultados estabelecem uma caracterização dos componentes humanos de referência e de sua integração funcional. Não foram avaliados níveis de expressão, variantes de pacientes ou estados de fosforilação. Assim, a contribuição para a discussão do LES está na organização das bases moleculares da via e na identificação de regiões que podem orientar investigações posteriores.

Em conjunto, a análise evidencia a complementaridade entre sensores, adaptadores, quinases e fatores de transcrição na sinalização imune. A integração de domínios, confiança estrutural e eventos anotados oferece uma base para compreender como proteínas molecularmente distintas participam de um mesmo circuito, aproximando a caracterização bioinformática dos mecanismos discutidos na patologia autoimune.

Considerações Finais

A caracterização bioinformática de cGAS, STING, TBK1 e IRF3 permitiu integrar informações de sequência, organização de domínios, confiança estrutural e participação em eventos de sinalização imune. As arquiteturas identificadas evidenciaram a diversidade molecular de componentes que atuam em etapas complementares, conectando o reconhecimento de DNA à produção de mensageiros, à transmissão do sinal e à regulação transcricional.

O mapeamento funcional contribuiu para situar essas características no eixo cGAS–STING–TBK1–IRF3, oferecendo uma base para discutir sua relevância na imunidade inata e no lúpus eritematoso sistêmico. A representação das regiões de maior confiança dos modelos estruturais acrescentou uma perspectiva espacial à análise, preservando a distinção entre predição computacional e atividade biológica.

Conclui-se que a integração entre bioinformática, biologia estrutural e imunologia favorece a compreensão dos mecanismos moleculares associados à autoimunidade. Os achados fornecem uma referência para investigações sobre variantes, regiões regulatórias e interações desses componentes, contribuindo para o estudo da modulação de vias inflamatórias em contextos específicos do lúpus.

Declaração de Uso de Inteligência Artificial Generativa

Durante a elaboração deste capítulo, foi utilizada a ferramenta ChatGPT, da OpenAI, como apoio à organização e à revisão textual, à elaboração e ao ajuste de scripts para análises bioinformáticas e à preparação de figuras a partir dos dados analisados. Os autores revisaram o conteúdo, os códigos, as referências e a interpretação dos resultados,

assumindo integral responsabilidade pela precisão, originalidade e integridade científica da versão final.

Referências

ARNAUD, Laurent; CHASSET, François; MARTIN, Thierry. Immunopathogenesis of systemic lupus erythematosus: an update. **Autoimmunity Reviews**, v. 23, n. 10, 103648, 2024. DOI: 10.1016/j.autrev.2024.103648.

CAIELLI, Simone; WAN, Zurong; PASCUAL, Virginia. Systemic lupus erythematosus pathogenesis: interferon and beyond. **Annual Review of Immunology**, v. 41, p. 533–560, 2023. DOI: 10.1146/annurev-immunol-101921-042422.

CHANG, Liu. Harnessing cGAS–STING axis for therapeutic benefits in systemic lupus erythematosus. **International Journal of Rheumatic Diseases**, v. 27, n. 7, e15256, 2024. DOI: 10.1111/1756-185X.15256.

MOTWANI, Mona *et al.* cGAS-STING pathway does not promote autoimmunity in murine models of SLE. **Frontiers in Immunology**, v. 12, 605930, 2021. DOI: 10.3389/fimmu.2021.605930.

REACTOME. STING mediated induction of host immune responses. [S. l.]: Reactome, [s. d.]. Identificador: R-HSA-1834941. Disponível em: <https://reactome.org/content/detail/R-HSA-1834941>. Acesso em: 24 out. 2026.

ZHANG, Zhengyin; ZHANG, Conggang. Regulation of cGAS–STING signalling and its diversity of cellular outcomes. **Nature Reviews Immunology**, v. 25, p. 425–444, 2025. DOI: 10.1038/s41577-024-01112-7.

CAPÍTULO 12

IDENTIFICAÇÃO BIOINFORMÁTICA DE ALVOS MOLECULARES PARA TERAPIAS COMBINADAS EM DOENÇAS METABÓLICAS: UMA ABORDAGEM INTEGRATIVA

BIOINFORMATIC IDENTIFICATION OF MOLECULAR TARGETS FOR COMBINATION THERAPIES IN METABOLIC DISEASES: AN INTEGRATIVE APPROACH

DOI: <https://doi.org/10.56001/26.9786501906539.12>

Submetido em: 24/10/2026

Revisado em: 25/10/2026

Publicado em: 26/10/2026

Ruana Carolina Cabral da Silva

Programa de Pós-Graduação em Ciências da Saúde, Universidade Federal da Grande
Dourados, Mato Grosso do Sul, Brasil

<http://lattes.cnpq.br/8634356371128001>

Gleyce Hellen de Almeida de Souza

Programa de Pós-Graduação em Ciências da Saúde, Universidade Federal da Grande
Dourados, Dourados, Mato Grosso do Sul, Brasil

<http://lattes.cnpq.br/8163519978837401>

Fabíola Lucini

Programa de Pós-Graduação em Ciências da Saúde, Universidade Federal da Grande
Dourados, Dourados, Mato Grosso do Sul, Brasil

<http://lattes.cnpq.br/1608714419112836>

Luana Rossato

Laboratório de Pesquisa em Ciências da Saúde, Universidade Federal da Grande
Dourados, Dourados, Mato Grosso do Sul, Brasil

<http://lattes.cnpq.br/6429982229782529>

Resumo

As doenças metabólicas apresentam mecanismos fisiopatológicos complexos, caracterizados pela interação entre alterações no metabolismo lipídico, na sinalização hormonal, na resposta inflamatória e na regulação da homeostase energética. Nesse contexto, abordagens bioinformáticas podem contribuir para a identificação e priorização de alvos moleculares potencialmente relevantes ao desenvolvimento de estratégias terapêuticas combinadas. Este estudo teve como objetivo caracterizar e priorizar alvos moleculares relacionados às doenças metabólicas, utilizando a doença hepática esteatótica associada à disfunção metabólica (MASLD) como modelo demonstrativo. Trata-se de um estudo exploratório, descritivo e computacional, fundamentado na análise de redes de associações proteína-proteína. Foram selecionados 18 genes humanos relacionados ao metabolismo lipídico, à regulação metabólica, à inflamação e à fibrogênese. As associações funcionais foram recuperadas na plataforma STRING, considerando escore mínimo de confiança de 0,700, e analisadas por meio da biblioteca NetworkX, utilizando métricas de grau de conectividade, centralidade de intermediação e PageRank. A análise identificou 33 associações funcionais entre as proteínas investigadas, permitindo construir uma rede molecular e estabelecer uma classificação topológica exploratória dos componentes selecionados. A organização da rede possibilitou representar diferentes processos biológicos associados à fisiopatologia da MASLD, enquanto as métricas de centralidade forneceram informações sobre a importância estrutural relativa das proteínas. Os resultados demonstraram a aplicabilidade da integração entre redes moleculares e análises topológicas na caracterização de componentes relacionados a doenças metabólicas complexas. Conclui-se que a abordagem empregada representa uma estratégia exploratória para a organização de informações moleculares e a formulação de hipóteses sobre potenciais alvos terapêuticos, sendo necessárias análises farmacológicas e validações experimentais para investigar sua aplicabilidade no desenvolvimento de terapias combinadas.

Palavras-Chave: Bioinformática; Alvos moleculares; Doenças metabólicas; MASLD; Redes de interação proteína-proteína.

Abstract

Metabolic diseases involve complex pathophysiological mechanisms characterized by interactions among alterations in lipid metabolism, hormonal signaling, inflammatory responses, and energy homeostasis. In this context, bioinformatics approaches may contribute to the identification and prioritization of molecular targets potentially relevant to the development of combination therapeutic strategies. This study aimed to characterize and prioritize molecular targets associated with metabolic diseases, using metabolic dysfunction-associated steatotic liver disease (MASLD) as a demonstrative model. This exploratory, descriptive, computational study was based on the analysis of protein-protein association networks. Eighteen human genes associated with lipid metabolism, metabolic regulation, inflammation, and fibrogenesis were selected. Functional associations were retrieved from the STRING database using a minimum confidence score of 0.700 and analyzed with the NetworkX library through degree, betweenness centrality, and PageRank metrics. The analysis identified 33 functional associations among the investigated proteins, enabling the construction of a molecular network and the exploratory topological ranking of the selected components. Network organization provided a representation of different biological processes associated with MASLD pathophysiology, while centrality metrics offered insights into the relative structural importance of the proteins. The findings demonstrated the applicability of integrating molecular networks and topological analyses to characterize components involved in complex metabolic diseases. In conclusion, the proposed approach represents an exploratory strategy for organizing molecular information and generating hypotheses regarding potential therapeutic targets. Further pharmacological analyses and experimental validation are required to investigate its applicability to the development of combination therapies.

Keywords: Bioinformatics; Molecular targets; Metabolic diseases; MASLD; Protein-protein association networks.

Introdução

As doenças metabólicas representam um importante desafio para os sistemas de saúde devido à elevada frequência, à evolução crônica e às complicações sistêmicas. Obesidade, diabetes mellitus tipo 2, dislipidemias e doenças hepáticas relacionadas à disfunção metabólica compartilham alterações na sinalização da insulina, no metabolismo energético e na resposta inflamatória. A compreensão desses mecanismos é fundamental para investigar estratégias terapêuticas integradas (Samuel; Shulman, 2012; Hotamisligil, 2017).

Os avanços da genética molecular e da biologia de sistemas demonstram que essas condições envolvem redes de proteínas, receptores e mediadores interconectados, cujas alterações repercutem sobre diferentes processos celulares (Barabási; Gulbahce; Loscalzo, 2011). Essa complexidade tem estimulado a investigação de terapias combinadas direcionadas a mecanismos complementares. Entretanto, a modulação de múltiplos alvos exige avaliação de eficácia, interações farmacológicas e segurança, não garantindo, isoladamente, maior benefício clínico (Hopkins, 2008; Jia *et al.*, 2009).

Nesse contexto, ferramentas bioinformáticas permitem integrar informações moleculares por meio de redes de associações proteína-proteína, métricas de centralidade e análises de enriquecimento funcional. Essas abordagens contribuem para caracterizar processos biológicos e priorizar candidatos à investigação farmacológica, embora seus resultados dependam de interpretação biológica e validação experimental (Szkarczyk *et al.*, 2023; Shannon *et al.*, 2003).

A doença hepática esteatótica associada à disfunção metabólica (MASLD) constitui um modelo relevante para essas investigações por sua associação com obesidade, resistência à insulina, diabetes mellitus tipo 2 e dislipidemias (Rinella *et al.*, 2023). Sua fisiopatologia envolve acúmulo de lipídios, lipotoxicidade, disfunção mitocondrial, estresse oxidativo, inflamação e fibrogênese, evidenciando a participação de mecanismos interdependentes na progressão da doença (Friedman *et al.*, 2018).

A análise de redes pode auxiliar na identificação de proteínas relacionadas a esses mecanismos e na formulação de hipóteses sobre combinações de alvos. Contudo, a centralidade topológica não comprova relevância terapêutica, sendo necessário considerar evidências genéticas e funcionais, expressão tecidual e possibilidades de modulação farmacológica.

Diante disso, este capítulo tem como objetivo explorar ferramentas bioinformáticas para a priorização de potenciais alvos de terapias combinadas em doenças metabólicas, utilizando a MASLD como modelo demonstrativo. Para tanto, analisa uma rede de associações proteína-proteína e suas métricas de centralidade, discutindo a complementaridade entre componentes do metabolismo lipídico, da regulação energética, da inflamação e da fibrose hepática.

Metodologia

Trata-se de um estudo exploratório, descritivo e computacional que utilizou a doença hepática esteatótica associada à disfunção metabólica (MASLD) como modelo para investigar potenciais alvos de terapias combinadas. Foram utilizados exclusivamente dados públicos, sem participação de pacientes ou coleta de material biológico.

Foram selecionados, de maneira dirigida, 18 genes humanos relacionados ao metabolismo lipídico, à regulação metabólica, à sinalização hormonal, à inflamação e à fibrose: THRB, PPARA, PPARG, FGF21, FGFR1, KLB, GLP1R, SREBF1, FASN, ACACA, SCD, PNPLA3, TM6SF2, TNF, IL6, TGFB1, COL1A1 e ACTA2.

As associações funcionais entre as proteínas foram recuperadas pela API da plataforma STRING, restringindo-se à espécie *Homo sapiens* (código taxonômico 9606), com escore mínimo de confiança de 0,700 e sem inclusão de proteínas externas ao painel (Szkarczyk *et al.*, 2023). Os dados foram organizados em formato tabular, contendo os pares de proteínas e seus respectivos escores.

A rede foi construída em Python 3.11, utilizando a biblioteca NetworkX. As proteínas foram representadas como nós e as associações como arestas não direcionadas, mantendo-se os componentes sem conexões. Foram calculados o grau de conectividade, a centralidade de intermediação e o PageRank, utilizado como principal critério de ordenação, considerando os escores de confiança como pesos.

As representações gráficas foram elaboradas com NetworkX e Matplotlib. Na rede, as cores indicaram as categorias funcionais e o tamanho dos nós correspondeu ao grau de conectividade. Utilizou-se o posicionamento spring layout, com semente fixa em 42, e um gráfico de barras para apresentar os valores de PageRank.

A interpretação integrou as métricas topológicas às funções biológicas das proteínas, buscando reconhecer mecanismos complementares para investigações posteriores. A priorização foi considerada exploratória, sem comprovar viabilidade farmacológica, sinergia terapêutica ou eficácia clínica.

Resultados e Discussão

Caracterização da rede molecular associada à MASLD

A análise bioinformática realizada a partir de um painel de 18 genes humanos permitiu identificar 33 associações funcionais entre proteínas relacionadas a diferentes mecanismos fisiopatológicos da doença hepática esteatótica associada à disfunção metabólica (MASLD). As associações foram recuperadas na plataforma STRING, considerando um escore mínimo de confiança de 0,700, e utilizadas para a construção de uma rede molecular destinada à investigação da conectividade e da organização topológica dos componentes selecionados.

O painel analisado contemplou proteínas envolvidas na regulação do metabolismo lipídico, na oxidação de ácidos graxos, na sinalização hormonal, na sensibilidade à insulina, na lipogênese, na inflamação e na fibrogênese. Essa composição permitiu representar diferentes dimensões da fisiopatologia da MASLD, considerando que o desenvolvimento e a progressão da doença envolvem alterações metabólicas e respostas celulares interdependentes.

A representação gráfica da rede, apresentada no **painel A da Figura 1**, possibilitou visualizar as associações recuperadas entre os componentes investigados, bem como distinguir os grupos funcionais previamente estabelecidos. A organização resultante constitui uma representação exploratória das relações moleculares presentes no painel, permitindo observar a coexistência de proteínas associadas a mecanismos metabólicos, inflamatórios e fibróticos.

A identificação dessas associações reforça a pertinência de investigar a MASLD por meio de uma abordagem integrada, em vez de considerar isoladamente os diferentes mecanismos fisiopatológicos. A doença apresenta elevada heterogeneidade biológica, e sua progressão pode envolver alterações na disponibilidade de substratos energéticos, na sinalização metabólica, na resposta imunológica e nos mecanismos de reparação tecidual. Nesse contexto, Tilg, Adolph e Moschen (2021) destacam a importância da interação entre processos metabólicos, inflamatórios e ambientais na evolução da doença hepática gordurosa associada à disfunção metabólica.

Entretanto, a presença de uma associação funcional entre duas proteínas não demonstra necessariamente interação física direta, regulação causal ou participação conjunta em um mesmo mecanismo patológico.

Priorização topológica das proteínas e interpretação biológica

A análise topológica da rede foi conduzida por meio das métricas de grau de conectividade, centralidade de intermediação e PageRank. Esses parâmetros permitiram caracterizar diferentes propriedades estruturais dos componentes investigados e estabelecer uma ordenação relativa das proteínas segundo sua posição na rede.

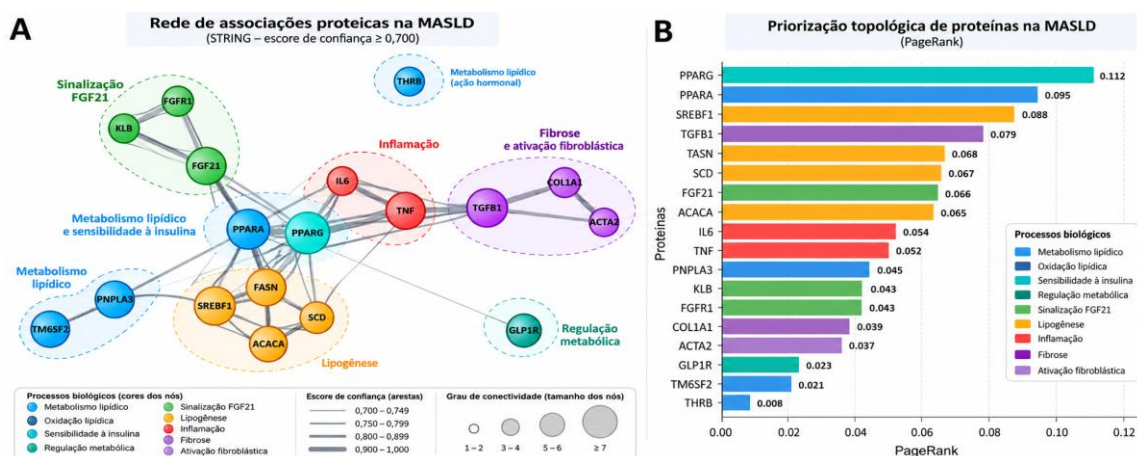
O grau de conectividade foi utilizado para quantificar o número de associações estabelecidas por cada proteína, enquanto a centralidade de intermediação permitiu avaliar sua participação nos caminhos mínimos que conectam diferentes componentes. O PageRank, por sua vez, possibilitou considerar simultaneamente a conectividade de cada nó e a importância estrutural de suas conexões, sendo empregado como critério principal para a elaboração do ranking exploratório.

A aplicação dessas métricas é relevante porque diferentes proteínas podem apresentar padrões distintos de conectividade, mesmo quando pertencem a processos biológicos relacionados. Nesse sentido, uma proteína com elevado número de associações não necessariamente apresenta a mesma importância estrutural de outra que estabelece conexões entre regiões distintas da rede. A utilização de métricas complementares contribui, portanto, para uma caracterização mais abrangente da organização molecular.

O painel B da Figura 1 apresenta a distribuição dos valores de PageRank das proteínas analisadas, permitindo comparar sua importância topológica relativa. A representação gráfica complementa a rede apresentada no painel A, reunindo informações sobre a organização das associações funcionais e a priorização estrutural dos componentes moleculares.

A apresentação conjunta dos dois painéis permite estabelecer uma relação entre a conectividade observada na rede e os valores de centralidade calculados. Enquanto o painel A evidencia a organização das associações entre proteínas, o painel B sintetiza sua classificação segundo o PageRank, proporcionando uma visualização integrada dos resultados.

Figura 1: Caracterização da rede de associações proteicas e priorização topológica de alvos moleculares relacionados à MASLD.



Legenda: (A) Rede de associações funcionais construída a partir de 18 genes humanos relacionados à fisiopatologia da doença hepática esteatótica associada à disfunção metabólica (MASLD). As associações foram recuperadas na plataforma STRING, considerando escore mínimo de confiança de 0,700, totalizando 33 conexões. Os nós representam proteínas, as arestas indicam associações funcionais e as cores diferenciam categorias biológicas previamente estabelecidas. (B) Priorização topológica das proteínas investigadas, apresentada por meio de gráfico de barras horizontais com os valores de PageRank. A classificação representa a importância estrutural relativa dos componentes na rede analisada, não constituindo medida direta de potencial farmacológico ou eficácia terapêutica. Fonte: Elaboração própria, a partir dos resultados da análise bioinformática (2026).

A interpretação conjunta dos painéis evidencia a utilidade da análise de redes para a organização de informações moleculares relacionadas à MASLD. A representação das associações e a classificação topológica permitem reconhecer diferentes padrões de conectividade e identificar proteínas que podem ser consideradas em etapas posteriores de investigação funcional e farmacológica.

Entretanto, a centralidade molecular não constitui, isoladamente, evidência suficiente para caracterizar uma proteína como alvo farmacológico prioritário. Proteínas altamente conectadas podem participar de processos fisiológicos essenciais, de modo que sua modulação indiscriminada pode produzir consequências sistêmicas indesejáveis. Por outro lado, componentes com menor centralidade podem apresentar relevância terapêutica quando associados a mecanismos específicos da doença ou quando sua modulação é sustentada por evidências experimentais e clínicas.

Essa distinção é particularmente importante na investigação de doenças metabólicas, nas quais diferentes tecidos e sistemas participam da manutenção da homeostase energética. A interpretação dos resultados computacionais deve, portanto,

integrar a posição topológica das proteínas com informações sobre suas funções biológicas, distribuição tecidual, mecanismos de regulação e possíveis efeitos decorrentes de sua modulação.

Proteínas relacionadas ao metabolismo lipídico e à regulação energética

Entre os componentes selecionados para a construção da rede, destacam-se proteínas relacionadas à regulação do metabolismo lipídico e à homeostase energética, incluindo *THRB*, *PPARA*, *PPARG*, *FGF21*, *FGFR1*, *KLB* e *GLP1R*. Esses componentes participam de mecanismos distintos, mas potencialmente complementares, relacionados à utilização de substratos energéticos, à sinalização hormonal e à regulação de processos metabólicos.

O receptor beta do hormônio tireoidiano, codificado pelo gene *THRB*, apresenta relevância para o metabolismo hepático de lipídios, uma vez que participa da regulação transcricional de processos associados à utilização e ao processamento de substratos energéticos. A investigação desse receptor como alvo terapêutico tem recebido atenção no contexto das doenças hepáticas metabólicas. Harrison *et al.* (2024), em estudo clínico envolvendo pacientes com esteato-hepatite associada à disfunção metabólica, demonstraram que a modulação seletiva desse receptor por meio do resmetirom apresentou benefícios em desfechos histológicos específicos, incluindo resolução da esteato-hepatite e melhora da fibrose em parte dos participantes.

A inclusão de *THRB* no painel molecular, portanto, apresenta plausibilidade biológica e farmacológica. Entretanto, sua posição na rede não deve ser utilizada isoladamente para inferir magnitude de resposta terapêutica, pois a eficácia de um medicamento depende de características farmacodinâmicas, farmacocinéticas e clínicas que não são representadas diretamente pelas métricas topológicas.

Os receptores ativados por proliferadores de peroxissoma, representados por *PPARA* e *PPARG*, também constituem componentes relevantes da regulação metabólica. Esses receptores nucleares participam de mecanismos associados à oxidação de ácidos graxos, à diferenciação adipocitária e à sensibilidade à insulina. Sua inclusão permite contemplar processos relacionados tanto ao metabolismo hepático quanto à regulação sistêmica da disponibilidade de lipídios.

A sinalização mediada por *FGF21*, *FGFR1* e *KLB* representa outro eixo molecular de interesse. O *FGF21* atua como regulador metabólico, enquanto *FGFR1* e *KLB*

participam de mecanismos envolvidos na resposta celular a esse fator. A investigação desse sistema tem contribuído para o desenvolvimento de estratégias farmacológicas voltadas à modulação do metabolismo energético e à redução de alterações associadas à doença hepática metabólica.

Loomba *et al.* (2023) investigaram o pegozafermin, um análogo de FGF21, em pacientes com esteato-hepatite não cirrótica e demonstraram resultados favoráveis em desfechos histológicos avaliados no estudo. Esses achados fornecem sustentação clínica à investigação do eixo FGF21, embora não permitam concluir que combinações envolvendo esse sistema sejam necessariamente superiores à modulação isolada.

O receptor GLP1R, por sua vez, está relacionado à sinalização das incretinas e à regulação metabólica. Sua relevância decorre da participação em mecanismos associados à secreção de insulina, ao controle glicêmico e à regulação da ingestão alimentar. Newsome *et al.* (2021) demonstraram que o tratamento com semaglutida esteve associado a maior frequência de resolução da esteato-hepatite em comparação ao placebo, embora o desfecho de melhora da fibrose não tenha apresentado diferença estatisticamente significativa no estudo.

Em conjunto, esses achados evidenciam que diferentes componentes do painel molecular apresentam fundamentos biológicos e farmacológicos que justificam sua investigação. A complementaridade entre receptores nucleares, sinalização hormonal e mecanismos de regulação energética constitui uma hipótese relevante para estudos futuros, mas sua tradução em estratégias terapêuticas combinadas depende de evidências específicas sobre eficácia, segurança e interação entre intervenções.

Lipogênese e alterações do metabolismo hepático

Os genes *SREBF1*, *FASN*, *ACACA* e *SCD* foram incluídos no painel por sua participação em mecanismos associados à síntese e ao processamento de lipídios. Esses componentes representam diferentes etapas da regulação lipogênica, contribuindo para a compreensão de processos envolvidos no acúmulo de gordura hepática.

A proteína SREBF1 participa da regulação transcricional de genes relacionados à síntese de ácidos graxos, enquanto FASN e ACACA estão associadas a etapas enzimáticas da lipogênese. A proteína SCD, por sua vez, participa da dessaturação de ácidos graxos, contribuindo para a composição lipídica celular.

A inclusão dessas proteínas é particularmente relevante porque o acúmulo de gordura no fígado não depende exclusivamente da ingestão de lipídios, mas também de

processos endógenos de síntese, utilização e exportação de ácidos graxos. Donnelly *et al.* (2005), ao investigarem as fontes de ácidos graxos hepáticos em indivíduos com doença hepática gordurosa não alcoólica, demonstraram a contribuição de diferentes origens metabólicas para o conteúdo lipídico do fígado, incluindo a lipogênese de novo.

Esses resultados reforçam a importância de considerar a lipogênese como componente da fisiopatologia da MASLD. Entretanto, a identificação de enzimas relacionadas à síntese lipídica não significa que sua inibição farmacológica seja necessariamente benéfica em todos os contextos, uma vez que a regulação do metabolismo de ácidos graxos envolve mecanismos compensatórios e funções fisiológicas essenciais.

Os genes *PNPLA3* e *TM6SF2* foram incorporados ao painel por sua associação com características do metabolismo lipídico hepático. Romeo *et al.* (2008) identificaram associação entre variação genética em *PNPLA3* e diferenças no conteúdo de gordura hepática, contribuindo para o reconhecimento da importância dos determinantes genéticos na susceptibilidade à esteatose.

Posteriormente, Kozlitina *et al.* (2014) demonstraram associação entre uma variante de *TM6SF2* e a susceptibilidade à doença hepática gordurosa não alcoólica. Esses achados evidenciam que alterações genéticas podem influenciar processos relacionados ao acúmulo de lipídios hepáticos, reforçando a natureza multifatorial da doença.

Contudo, a relevância genética de determinado componente não implica necessariamente que sua proteína seja um alvo farmacológico diretamente acessível ou seguro. Essa distinção deve ser considerada na interpretação do painel, especialmente quando se busca transformar informações moleculares em hipóteses de intervenção terapêutica.

Inflamação, ativação fibroblástica e progressão da doença

A participação de mecanismos inflamatórios e fibróticos na progressão da MASLD justificou a inclusão dos genes *TNF*, *IL6*, *TGFB1*, *COL1A1* e *ACTA2*. Esses componentes representam diferentes aspectos da resposta inflamatória, da remodelação tecidual e da ativação de células envolvidas na deposição de matriz extracelular.

TNF e IL6 estão associados à sinalização inflamatória e podem participar de respostas celulares relacionadas à disfunção metabólica. Entretanto, seus efeitos dependem do contexto fisiológico, do tecido envolvido e das características da resposta

imunológica, não sendo adequadamente representados por uma interpretação exclusivamente pró-inflamatória.

A inclusão desses mediadores na rede permite contemplar a dimensão imunometabólica da doença, considerando que alterações metabólicas podem modificar a atividade de células imunológicas e que processos inflamatórios podem, por sua vez, influenciar a homeostase energética e a função hepática.

No contexto da fibrogênese, TGFB1 apresenta relevância por sua participação em mecanismos de sinalização associados à ativação de células produtoras de matriz extracelular. COL1A1 representa um componente estrutural importante da matriz fibrótica, enquanto ACTA2 está relacionado ao fenótipo de células ativadas envolvidas na remodelação tecidual.

Tsuchida e Friedman (2017) destacam que a fibrose hepática envolve a interação entre diferentes populações celulares e mediadores moleculares, com participação de mecanismos inflamatórios, sinalização parácrina e ativação de células estreladas hepáticas. Essa complexidade demonstra que a progressão fibrótica não pode ser compreendida exclusivamente pela atividade de uma única proteína.

A presença desses componentes no painel amplia a abrangência biológica da análise, permitindo considerar não apenas mecanismos associados ao acúmulo inicial de lipídios, mas também processos relacionados à evolução da doença. Entretanto, a rede construída não permite determinar a direção das relações regulatórias nem estabelecer se a modulação de determinada proteína resultaria em redução da fibrose.

Dessa forma, proteínas associadas à inflamação e à fibrogênese devem ser interpretadas como componentes relevantes para a compreensão da progressão da MASLD, mas sua priorização farmacológica exige evidências adicionais sobre seletividade, segurança e efeitos em diferentes tecidos.

Implicações para a investigação de terapias combinadas

A caracterização da rede molecular e a distribuição funcional dos genes selecionados permitem reconhecer diferentes eixos fisiopatológicos potencialmente relevantes para o desenvolvimento de estratégias terapêuticas combinadas. Entre esses eixos, destacam-se a regulação do metabolismo lipídico, a sinalização hormonal, o controle da homeostase energética e os mecanismos relacionados à inflamação e à fibrogênese.

A possibilidade de atuar sobre mecanismos complementares apresenta interesse científico, especialmente em doenças complexas nas quais a modulação isolada de determinado processo pode não contemplar toda a diversidade fisiopatológica. Nesse contexto, a integração entre informações topológicas e evidências farmacológicas pode contribuir para a seleção preliminar de combinações de alvos a serem investigadas.

A análise realizada permite formular hipóteses exploratórias envolvendo proteínas associadas a processos distintos, como receptores relacionados à regulação metabólica e componentes envolvidos na síntese ou utilização de lipídios. Entretanto, essas possibilidades não constituem combinações farmacológicas validadas, uma vez que o estudo não avaliou diretamente medicamentos, afinidade de ligação, interações farmacocinéticas ou efeitos de tratamentos combinados.

A distinção entre associação molecular e potencial terapêutico é particularmente importante na interpretação de redes funcionais. Uma proteína pode apresentar elevada conectividade e, ainda assim, não constituir um alvo adequado para intervenção farmacológica. Da mesma forma, duas proteínas pertencentes a processos diferentes não representam necessariamente uma combinação terapêutica vantajosa.

A transição entre a identificação computacional de alvos e o desenvolvimento de terapias exige uma sequência de avaliações complementares, incluindo validação funcional, investigação da possibilidade de modulação farmacológica, caracterização dos efeitos biológicos e análise de segurança. Assim, a abordagem apresentada deve ser compreendida como uma etapa inicial de organização e priorização de informações moleculares, capaz de orientar perguntas de pesquisa e subsidiar investigações posteriores.

Considerações Finais

A abordagem bioinformática permitiu integrar 18 genes e 33 associações funcionais relacionados ao metabolismo lipídico, à regulação energética, à inflamação e à fibrogênese na MASLD. As métricas de centralidade contribuíram para caracterizar a organização da rede e apoiar a priorização exploratória de potenciais alvos terapêuticos. Entretanto, a centralidade das proteínas não comprova viabilidade farmacológica, eficácia clínica ou sinergia terapêutica. A seleção dirigida dos genes, a dependência de bases públicas e a ausência de validação experimental limitam os resultados à formulação de hipóteses. Conclui-se que a análise de redes constitui uma estratégia útil para orientar

investigações sobre mecanismos complementares na MASLD, devendo ser integrada a evidências genéticas, funcionais e farmacológicas e à validação experimental.

Declaração de Uso de Inteligência Artificial Generativa

Durante a elaboração deste capítulo, foi utilizada a ferramenta ChatGPT, da OpenAI, como apoio à organização e à revisão textual, à elaboração e ao ajuste de scripts para análises bioinformáticas e à preparação de figuras a partir dos dados analisados. Os autores revisaram o conteúdo, os códigos, as referências e a interpretação dos resultados, assumindo integral responsabilidade pela precisão, originalidade e integridade científica da versão final.

Referências

- BARABÁSI, Albert-László; GULBAHCE, Natali; LOSCALZO, Joseph. Network medicine: a network-based approach to human disease. **Nature Reviews Genetics**, v. 12, n. 1, p. 56-68, 2011. DOI: <https://doi.org/10.1038/nrg2918>.
- DONNELLY, Kerry L. *et al.* Sources of fatty acids stored in liver and secreted via lipoproteins in patients with nonalcoholic fatty liver disease. **The Journal of Clinical Investigation**, v. 115, n. 5, p. 1343-1351, 2005. DOI: <https://doi.org/10.1172/JCI23621>.
- FRIEDMAN, Scott L. *et al.* Mechanisms of NAFLD development and therapeutic strategies. **Nature Medicine**, v. 24, p. 908-922, 2018. DOI: <https://doi.org/10.1038/s41591-018-0104-9>.
- HARRISON, Stephen A. *et al.* A phase 3, randomized, controlled trial of resmetirom in NASH with liver fibrosis. **The New England Journal of Medicine**, v. 390, p. 497-509, 2024. DOI: <https://doi.org/10.1056/NEJMoa2309000>.
- HOPKINS, Andrew L. Network pharmacology: the next paradigm in drug discovery. **Nature Chemical Biology**, v. 4, n. 11, p. 682-690, 2008. DOI: <https://doi.org/10.1038/nchembio.118>.
- HOTAMISLIGIL, Gökhan S. Inflammation, metaflammation and immunometabolic disorders. **Nature**, v. 542, n. 7640, p. 177-185, 2017. DOI: <https://doi.org/10.1038/nature21363>.
- JIA, Jia *et al.* Mechanisms of drug combinations: interaction and network perspectives. **Nature Reviews Drug Discovery**, v. 8, n. 2, p. 111-128, 2009. DOI: <https://doi.org/10.1038/nrd2683>.
- KOZLITINA, Julia *et al.* Exome-wide association study identifies a TM6SF2 variant that confers susceptibility to nonalcoholic fatty liver disease. **Nature Genetics**, v. 46, p. 352-356, 2014. DOI: <https://doi.org/10.1038/ng.2901>.

LOOMBA, Rohit *et al.* Randomized, controlled trial of the FGF21 analogue pegozafermin in NASH. **The New England Journal of Medicine**, v. 389, p. 998-1008, 2023. DOI: <https://doi.org/10.1056/NEJMoa2304286>.

NEWSOME, Philip N. *et al.* A placebo-controlled trial of subcutaneous semaglutide in nonalcoholic steatohepatitis. **The New England Journal of Medicine**, v. 384, p. 1113-1124, 2021. DOI: <https://doi.org/10.1056/NEJMoa2028395>.

RINELLA, Mary E. *et al.* A multisociety Delphi consensus statement on new fatty liver disease nomenclature. **Hepatology**, v. 78, n. 6, p. 1966-1986, 2023. DOI: <https://doi.org/10.1097/HEP.0000000000000520>.

ROMEO, Stefano *et al.* Genetic variation in PNPLA3 confers susceptibility to nonalcoholic fatty liver disease. **Nature Genetics**, v. 40, p. 1461-1465, 2008. DOI: <https://doi.org/10.1038/ng.257>.

SAMUEL, Varman T.; SHULMAN, Gerald I. Mechanisms for insulin resistance: common threads and missing links. **Cell**, v. 148, n. 5, p. 852-871, 2012. DOI: <https://doi.org/10.1016/j.cell.2012.02.017>.

SHANNON, Paul *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. **Genome Research**, v. 13, n. 11, p. 2498-2504, 2003. DOI: <https://doi.org/10.1101/gr.1239303>.

SZKLARCZYK, Damian *et al.* The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. **Nucleic Acids Research**, v. 51, n. D1, p. D638-D646, 2023. DOI: <https://doi.org/10.1093/nar/gkac1000>.

TILG, Herbert; ADOLPH, Timon E.; MOSCHEN, Alexander R. Multiple parallel hits hypothesis in nonalcoholic fatty liver disease: revisited after a decade. **Hepatology**, v. 73, n. 2, p. 833-842, 2021. DOI: <https://doi.org/10.1002/hep.31518>.

TSUCHIDA, Takahiro; FRIEDMAN, Scott L. Mechanisms of hepatic stellate cell activation. **Nature Reviews Gastroenterology & Hepatology**, v. 14, p. 397-411, 2017. DOI: <https://doi.org/10.1038/nrgastro.2017.38>.

CAPÍTULO 13

SEQUENCIAMENTO DE NOVA GERAÇÃO E BIOINFORMÁTICA NO DIAGNÓSTICO DE DOENÇAS HUMANAS: APLICAÇÕES, AVANÇOS E PERSPECTIVAS

NEXT-GENERATION SEQUENCING AND BIOINFORMATICS IN THE DIAGNOSIS OF HUMAN DISEASES: APPLICATIONS, ADVANCES, AND PERSPECTIVES

DOI: <https://doi.org/10.56001/26.9786501906539.13>

Submetido em: 24/10/2026

Revisado em: 25/10/2026

Publicado em: 26/10/2026

Ruana Carolina Cabral da Silva

Programa de Pós-Graduação em Ciências da Saúde, Universidade Federal da Grande

Dourados, Mato Grosso do Sul, Brasil

<http://lattes.cnpq.br/8634356371128001>

Gleyce Hellen de Almeida de Souza

Programa de Pós-Graduação em Ciências da Saúde, Universidade Federal da Grande

Dourados, Dourados, Mato Grosso do Sul, Brasil

<http://lattes.cnpq.br/8163519978837401>

Fabíola Lucini

Programa de Pós-Graduação em Ciências da Saúde, Universidade Federal da Grande

Dourados, Dourados, Mato Grosso do Sul, Brasil

<http://lattes.cnpq.br/1608714419112836>

Luana Rossato

Laboratório de Pesquisa em Ciências da Saúde, Universidade Federal da Grande

Dourados, Dourados, Mato Grosso do Sul, Brasil

<http://lattes.cnpq.br/6429982229782529>

Resumo

O sequenciamento de nova geração (NGS), associado às ferramentas de bioinformática, tem ampliado as possibilidades de investigação molecular de doenças humanas, contribuindo para a identificação e interpretação de variantes genéticas relacionadas a diferentes condições patológicas. Este capítulo teve como objetivo analisar as aplicações, os avanços e as perspectivas do NGS e da bioinformática no diagnóstico de doenças humanas, utilizando a fibrose cística e o gene *CFTR* como modelo demonstrativo. Trata-se de um estudo descritivo e exploratório, fundamentado na análise de evidências científicas e na sistematização de etapas computacionais empregadas na investigação genômica. Foram abordados os procedimentos de obtenção e processamento de dados de sequenciamento, controle de qualidade das leituras, identificação de variantes, anotação funcional e interpretação clínica, considerando ferramentas e bases de dados utilizadas na análise molecular. A discussão evidenciou a importância da integração entre procedimentos laboratoriais, recursos computacionais e informações clínicas para a adequada interpretação das alterações genéticas. No estudo de caso, destacou-se a variante p.Phe508del do gene *CFTR*, associada a alterações no dobramento e no processamento intracelular da proteína, demonstrando a relação entre modificações genéticas e mecanismos fisiopatológicos. Entretanto, a identificação de variantes não determina isoladamente sua relevância clínica, sendo necessária a avaliação conjunta de evidências moleculares, funcionais e fenotípicas. Ressalta-se que a abordagem possui caráter demonstrativo, sem processamento original de dados de sequenciamento. Conclui-se que a integração entre NGS e bioinformática apresenta potencial para aprimorar o diagnóstico molecular, ampliar a compreensão das doenças genéticas e favorecer o desenvolvimento da medicina de precisão, desde que acompanhada de rigor metodológico, interpretação especializada e validação adequada dos resultados.

Palavras-Chave: Sequenciamento de nova geração; Bioinformática; Diagnóstico molecular; Variantes genéticas; Fibrose cística.

Abstract

Next-generation sequencing (NGS), combined with bioinformatics tools, has expanded the possibilities for molecular investigation of human diseases, contributing to the identification and interpretation of genetic variants associated with different pathological conditions. This chapter aimed to analyze the applications, advances, and perspectives of NGS and bioinformatics in the diagnosis of human diseases, using cystic fibrosis and the *CFTR* gene as a demonstrative model. This descriptive and exploratory study was based on the analysis of scientific evidence and the systematization of computational procedures employed in genomic investigations. The main steps addressed included sequencing data acquisition and processing, read quality control, variant identification, functional annotation, and clinical interpretation, considering computational tools and databases used in molecular analysis. The discussion highlighted the importance of integrating laboratory procedures, computational resources, and clinical information to ensure the appropriate interpretation of genetic alterations. The case study focused on the p.Phe508del variant in the *CFTR* gene, which is associated with impaired protein folding and intracellular processing, illustrating the relationship between genetic alterations and pathophysiological mechanisms. However, variant identification alone does not establish clinical relevance, requiring the combined assessment of molecular, functional, and phenotypic evidence. This approach was demonstrative in nature and did not involve original sequencing data processing. It is concluded that the integration of NGS and bioinformatics has the potential to improve molecular diagnosis, advance the understanding of genetic diseases, and support the development of precision medicine, provided that methodological rigor, specialized interpretation, and appropriate validation of results are ensured.

Keywords: Next-generation sequencing; Bioinformatics; Molecular diagnosis; Genetic variants; Cystic fibrosis.

Introdução

O avanço das tecnologias de sequenciamento de ácidos nucleicos transformou significativamente a investigação das bases moleculares das doenças humanas, ampliando

a capacidade de identificar alterações genéticas associadas a diferentes condições clínicas. Durante décadas, o sequenciamento pelo método de Sanger representou uma das principais estratégias para a análise de regiões específicas do DNA. Entretanto, a necessidade de investigar simultaneamente múltiplos genes e regiões genômicas impulsionou o desenvolvimento de tecnologias de maior capacidade analítica. Nesse cenário, o sequenciamento de nova geração (*Next-Generation Sequencing* – NGS) consolidou-se como uma ferramenta relevante para a pesquisa biomédica e o diagnóstico molecular, permitindo a obtenção de grandes volumes de dados genéticos em uma única execução. Goodwin, McPherson e McCombie (2016) destacam que o desenvolvimento dessas tecnologias ampliou a escala das análises genômicas, favorecendo investigações anteriormente limitadas por questões técnicas, operacionais e econômicas.

A aplicação do NGS na investigação de doenças humanas envolve diferentes estratégias, incluindo o sequenciamento de painéis gênicos direcionados, o sequenciamento completo do exoma (*Whole-Exome Sequencing* – WES) e o sequenciamento completo do genoma (*Whole-Genome Sequencing* – WGS). Essas abordagens apresentam diferenças quanto à abrangência das regiões analisadas, à capacidade de identificação de variantes e às exigências computacionais. Enquanto os painéis direcionados concentram a investigação em conjuntos previamente definidos de genes, o sequenciamento do exoma permite examinar predominantemente regiões codificantes, e o sequenciamento do genoma amplia a análise para regiões codificantes e não codificantes. A escolha da estratégia depende da hipótese diagnóstica, das características clínicas investigadas e das limitações técnicas de cada método (Goodwin; McPherson; McCombie, 2016).

Entre as principais contribuições dessas tecnologias destaca-se sua aplicação no diagnóstico de doenças raras e hereditárias, frequentemente caracterizadas por manifestações clínicas heterogêneas, sobreposição fenotípica e dificuldades de identificação etiológica. A investigação molecular pode contribuir para reconhecer variantes associadas a condições monogênicas, esclarecer mecanismos de herança e orientar o aconselhamento genético. Em estudo envolvendo famílias com suspeita de doenças raras, a utilização do sequenciamento genômico permitiu estabelecer diagnósticos moleculares definitivos ou prováveis em uma parcela relevante dos casos, demonstrando sua utilidade especialmente em situações nas quais avaliações diagnósticas anteriores não haviam identificado a causa da doença (Bick *et al.*, 2024).

No contexto brasileiro, a incorporação das tecnologias genômicas apresenta particular relevância diante das dificuldades de acesso ao diagnóstico especializado e da diversidade de condições hereditárias presentes na população. A ampliação da investigação molecular pode contribuir para reduzir trajetórias diagnósticas prolongadas, favorecer a identificação de condições geneticamente determinadas e subsidiar decisões assistenciais mais individualizadas. Entretanto, sua implementação exige infraestrutura laboratorial, profissionais qualificados, capacidade computacional e integração entre resultados moleculares e informações clínicas. Manolio et al. (2019) ressaltam que a utilização efetiva da genômica na assistência depende não apenas da disponibilidade tecnológica, mas também da organização dos serviços, da interpretação adequada dos resultados e da incorporação das evidências genéticas à tomada de decisão clínica.

Apesar dos avanços proporcionados pelo NGS, a obtenção das sequências representa apenas uma etapa do processo diagnóstico. A transformação dos dados brutos em informações biologicamente interpretáveis depende de procedimentos bioinformáticos que incluem avaliação da qualidade das leituras, alinhamento a um genoma de referência, identificação de variantes, anotação funcional e priorização de alterações potencialmente relacionadas ao fenótipo investigado. Esses procedimentos são fundamentais para diferenciar variações genéticas comuns de alterações que possam apresentar relevância clínica, considerando aspectos como frequência populacional, consequência molecular, conservação evolutiva e associação previamente descrita com doenças. Dessa forma, a bioinformática constitui um componente indispensável da análise genômica, permitindo organizar e interpretar grandes conjuntos de dados gerados pelas plataformas de sequenciamento.

A identificação de uma variante genética, entretanto, não significa necessariamente que ela seja responsável pelo desenvolvimento de determinada doença. A interpretação clínica exige avaliação criteriosa das evidências disponíveis, considerando informações populacionais, funcionais, familiares e fenotípicas. Richards et al. (2015), por meio das recomendações conjuntas do *American College of Medical Genetics and Genomics* (ACMG) e da *Association for Molecular Pathology* (AMP), estabeleceram critérios para a classificação de variantes em cinco categorias: patogênicas, provavelmente patogênicas, de significado incerto, provavelmente benignas e benignas. Essa sistematização contribuiu para a padronização da interpretação de variantes germinativas associadas a doenças mendelianas, embora sua aplicação deva considerar as

particularidades de cada condição clínica e as atualizações posteriores das recomendações.

Outro aspecto relevante envolve a utilização de bancos de dados genômicos públicos, que disponibilizam sequências de referência, registros de variantes e informações sobre suas possíveis consequências biológicas. Esses recursos favorecem a realização de estudos computacionais reprodutíveis e permitem investigar genes associados a doenças humanas sem a necessidade inicial de produzir novos dados de sequenciamento. Entretanto, a utilização dessas informações exige atenção à procedência dos registros, à qualidade das evidências e à distinção entre variantes identificadas experimentalmente e alterações cuja interpretação permanece incerta. A integração entre dados públicos, ferramentas bioinformáticas e critérios de classificação representa, portanto, uma estratégia relevante para o desenvolvimento de estudos de caso voltados à compreensão dos processos envolvidos no diagnóstico molecular.

Nesse contexto, a elaboração de estudos de caso bioinformáticos apresenta potencial para demonstrar, de maneira aplicada, as etapas que conectam a análise de dados genômicos à identificação e interpretação de variantes genéticas. Essa abordagem permite explorar aspectos técnicos e biológicos do sequenciamento, além de discutir a relevância e as limitações das evidências utilizadas na investigação de doenças humanas. Contudo, os resultados obtidos em análises computacionais exploratórias não substituem a avaliação clínica, a validação laboratorial quando indicada ou a interpretação realizada por profissionais habilitados, devendo ser compreendidos como informações complementares ao processo diagnóstico.

Diante do exposto, este capítulo tem como objetivo investigar, por meio de um estudo de caso bioinformático baseado em dados genômicos públicos, a aplicação de ferramentas computacionais na identificação, anotação e interpretação de variantes genéticas potencialmente associadas a doenças humanas, destacando as contribuições do sequenciamento de nova geração para o diagnóstico molecular e discutindo as possibilidades e limitações da integração entre dados genômicos, evidências científicas e critérios de interpretação clínica.

Metodologia

Delineamento do estudo

Trata-se de um estudo exploratório, descritivo e demonstrativo, fundamentado na análise de evidências científicas e na sistematização de procedimentos bioinformáticos aplicados ao sequenciamento de nova geração (*Next-Generation Sequencing* – NGS). Adotou-se como modelo a fibrose cística, com ênfase no gene *CFTR* e na variante p.Phe508del, buscando discutir as contribuições das ferramentas computacionais para a identificação, caracterização e interpretação de alterações genéticas relacionadas a doenças humanas.

Fontes de dados e caracterização molecular

A investigação foi fundamentada em informações científicas publicadas e em bases genômicas de referência, considerando recursos como o National Center for Biotechnology Information (NCBI), Ensembl, ClinVar e Genome Aggregation Database (gnomAD). Foram abordadas características moleculares do gene *CFTR*, consequências funcionais da variante p.Phe508del e evidências de sua associação com a fibrose cística. Não foram utilizados arquivos individuais de sequenciamento nem executadas análises originais de dados genômicos.

Sistematização do fluxo bioinformático

Foi estruturado um fluxo demonstrativo de análise de NGS, contemplando avaliação da qualidade das leituras, alinhamento ao genoma humano de referência GRCh38, identificação de variantes, anotação funcional e interpretação clínica. Foram consideradas ferramentas como FastQC, MultiQC, BWA-MEM, SAMtools, GATK e Ensembl Variant Effect Predictor (VEP), além de bases como gnomAD e ClinVar. Essas ferramentas foram apresentadas como componentes de um fluxo metodológico de referência, sem execução computacional no presente estudo.

Interpretação das variantes e apresentação dos resultados

A discussão da relevância molecular e clínica das variantes foi orientada por evidências científicas e pelos princípios de interpretação estabelecidos pelo American College of Medical Genetics and Genomics (ACMG) e pela Association for Molecular Pathology (AMP), conforme Richards et al. (2015). A variante p.Phe508del foi utilizada

para exemplificar a relação entre alteração genética, comprometimento funcional da proteína CFTR e mecanismos fisiopatológicos.

Os resultados foram organizados de forma descritiva, acompanhados de figuras esquemáticas referentes ao fluxo de sequenciamento, aos parâmetros de qualidade, à anotação de variantes e à caracterização molecular da CFTR. As representações gráficas possuem finalidade exclusivamente ilustrativa e não correspondem a resultados experimentais ou análises computacionais efetivamente realizadas.

Resultados e Discussão

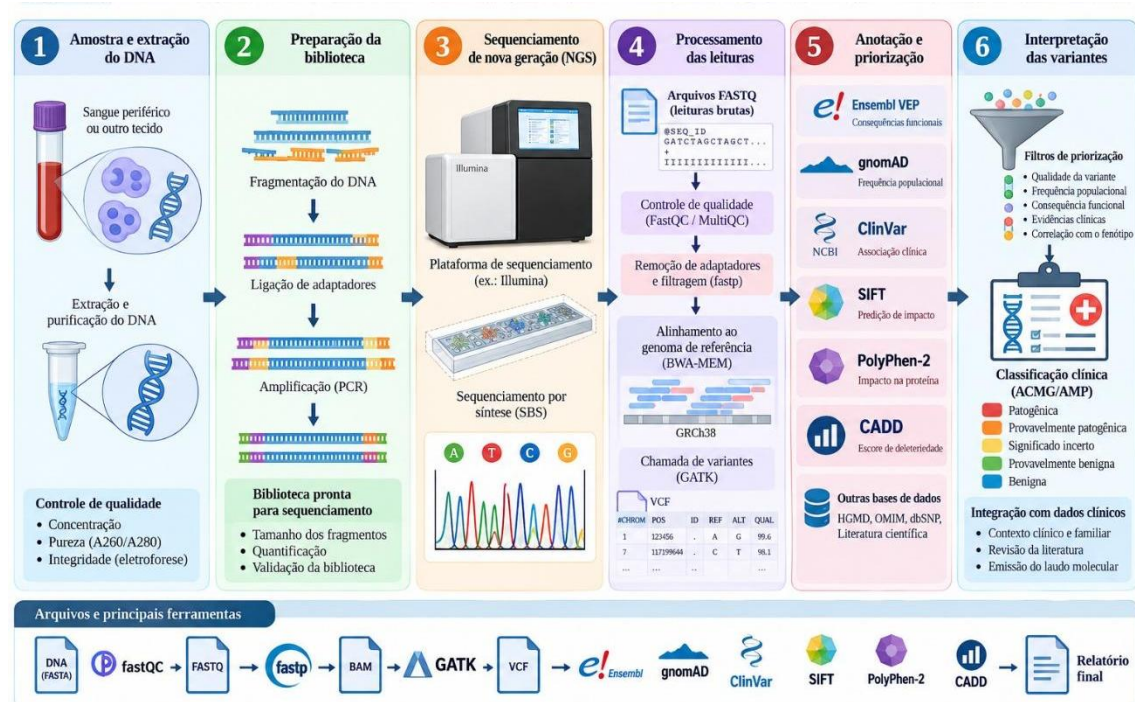
Aplicação do sequenciamento de nova geração à investigação de doenças genéticas

A análise das aplicações do sequenciamento de nova geração (NGS) evidencia sua relevância para a investigação molecular de doenças humanas, especialmente aquelas associadas a alterações em genes envolvidos em processos fisiológicos essenciais. Para demonstrar a integração entre sequenciamento, análise bioinformática e interpretação clínica, adotou-se como modelo a fibrose cística, doença genética de herança autossômica recessiva associada a variantes no gene *CFTR*, responsável pela codificação de uma proteína envolvida no transporte de íons cloreto e bicarbonato através de membranas epiteliais.

A utilização desse modelo permite discutir diferentes etapas da investigação molecular, desde a geração das leituras de sequenciamento até a identificação e interpretação de alterações genéticas. Entretanto, é necessário distinguir os procedimentos que compõem um fluxo bioinformático de resultados efetivamente produzidos a partir de amostras sequenciadas. Neste capítulo, a aplicação ao gene *CFTR* constitui um estudo de caso demonstrativo, fundamentado em evidências moleculares publicadas, não sendo apresentada como análise original de uma amostra clínica.

O fluxo de trabalho representado na Figura 1 contempla seis etapas principais: obtenção e extração do DNA, preparação da biblioteca genômica, sequenciamento, processamento das leituras, anotação e priorização das variantes e interpretação clínica. A integração dessas etapas demonstra que a geração dos dados genômicos representa apenas o início do processo diagnóstico, cuja interpretação depende de procedimentos computacionais e da avaliação das evidências disponíveis.

Figura 1: Fluxo de trabalho do sequenciamento de nova geração e da análise bioinformática para identificação e interpretação de variantes genéticas.



Legenda: Representação esquemática das etapas de investigação molecular por NGS, compreendendo coleta e extração do DNA, preparação da biblioteca, sequenciamento, avaliação da qualidade das leituras, alinhamento ao genoma de referência, identificação de variantes, anotação funcional e interpretação clínica. São apresentados exemplos de ferramentas computacionais e bases de dados utilizadas em diferentes etapas do processamento genômico. O fluxo representa uma abordagem metodológica possível e não corresponde à execução experimental de sequenciamento neste estudo.

Fonte: Elaboração própria (2026).

A integração entre procedimentos laboratoriais e computacionais é fundamental para garantir que alterações identificadas nos dados genômicos sejam interpretadas de maneira adequada. Erros de sequenciamento, regiões com cobertura insuficiente e dificuldades de alinhamento podem comprometer a identificação de variantes, especialmente em regiões repetitivas ou estruturalmente complexas. Dessa forma, a confiabilidade da investigação depende da qualidade dos dados e da aplicação de critérios técnicos apropriados em cada etapa.

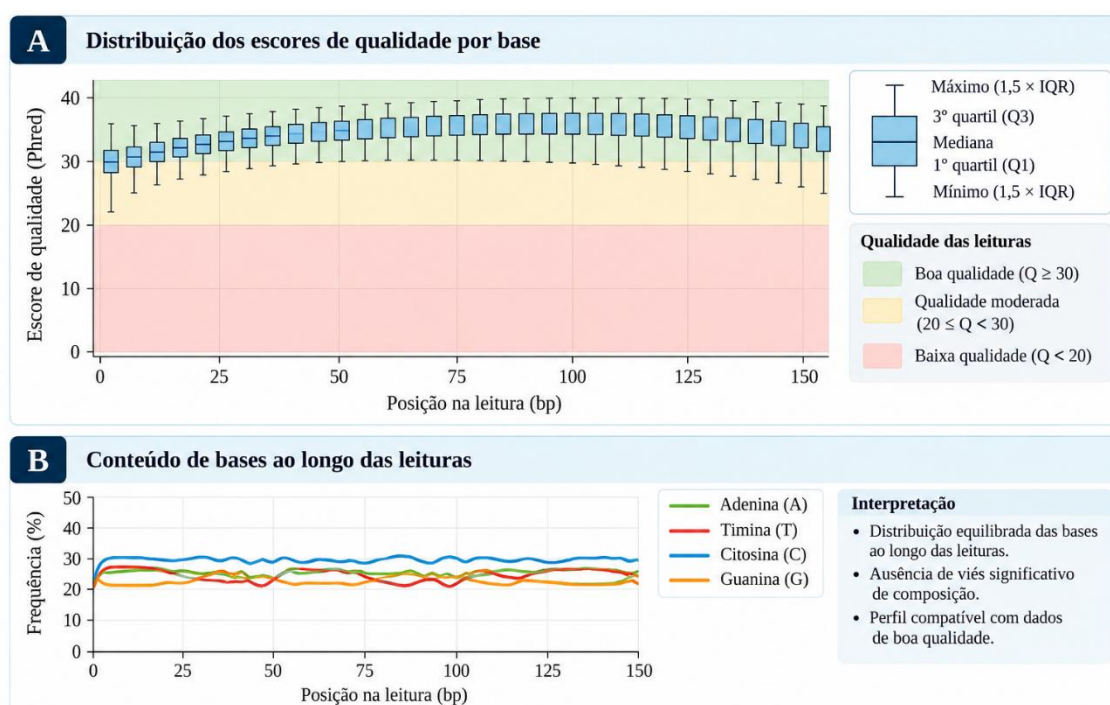
Avaliação da qualidade das leituras e confiabilidade da análise genômica

A qualidade das leituras constitui um dos principais determinantes da confiabilidade das análises realizadas a partir de dados de NGS. Durante o sequenciamento, cada base identificada recebe uma estimativa de qualidade, geralmente expressa por meio do escore Phred, que representa a probabilidade estimada de erro na identificação do nucleotídeo.

A avaliação desses escores permite reconhecer possíveis problemas técnicos, como redução progressiva da qualidade ao longo das leituras, presença de sequências de adaptadores e alterações inesperadas na composição de bases. Essas características podem influenciar o alinhamento ao genoma de referência e a identificação posterior de variantes.

A Figura 2 apresenta uma representação ilustrativa dos principais parâmetros utilizados na avaliação da qualidade das leituras. O painel A demonstra a distribuição dos escores Phred ao longo das posições de sequenciamento, enquanto o painel B representa a frequência relativa das bases nitrogenadas. Os perfis apresentados são esquemáticos e não correspondem a medições obtidas de arquivos FASTQ processados neste estudo.

Figura 2: Representação ilustrativa de parâmetros utilizados na avaliação da qualidade das leituras de sequenciamento.



Legenda: (A) Exemplo esquemático da distribuição dos escores de qualidade Phred ao longo das posições das leituras, destacando diferentes intervalos de qualidade. (B) Exemplo ilustrativo da composição percentual das bases nitrogenadas ao longo das leituras. Os valores apresentados possuem finalidade didática e não representam resultados experimentais de uma amostra analisada.

Fonte: Elaboração própria (2026).

A avaliação da qualidade deve anteceder as etapas de identificação de variantes, pois alterações técnicas podem resultar tanto em chamadas falso-positivas quanto na ausência de detecção de variantes reais. Além disso, a interpretação da composição de bases deve considerar as características da biblioteca e do método de sequenciamento, uma vez que determinados protocolos podem produzir distribuições específicas sem que isso represente necessariamente um problema técnico.

A análise conjunta desses parâmetros demonstra que a qualidade dos dados genômicos não pode ser determinada por um único indicador. A avaliação deve considerar a distribuição dos escores de qualidade, a presença de sequências contaminantes, a proporção de leituras alinhadas e a cobertura das regiões de interesse. Portanto, a obtenção de resultados molecularmente interpretáveis depende de um processo de controle de qualidade compatível com a finalidade diagnóstica da investigação.

Identificação e anotação funcional das variantes genéticas

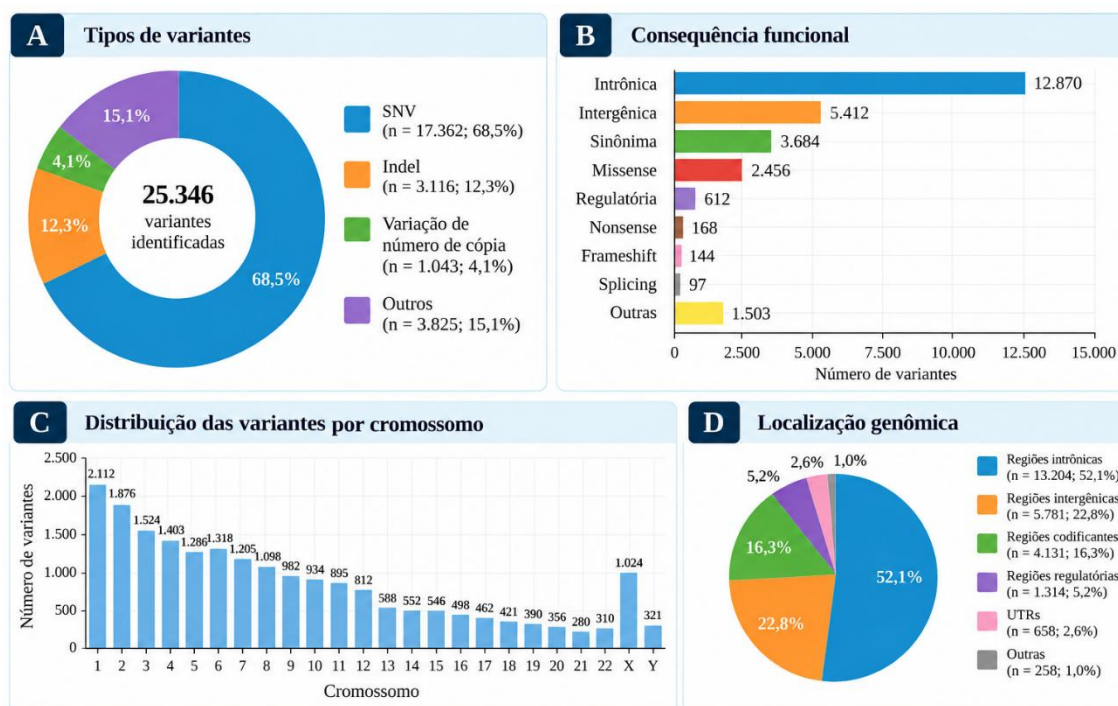
Após o processamento das leituras, a identificação de variantes permite reconhecer diferenças entre as sequências analisadas e o genoma de referência. Entre as alterações frequentemente investigadas estão as variantes de nucleotídeo único, pequenas inserções e deleções, além de determinadas alterações estruturais, cuja detecção depende da tecnologia utilizada e das características do fluxo computacional.

A anotação funcional constitui uma etapa essencial para caracterizar as consequências potenciais dessas alterações. Variantes localizadas em regiões codificantes podem resultar em substituições de aminoácidos, interrupção prematura da tradução ou modificações na matriz de leitura. Alterações em regiões não codificantes também podem apresentar relevância biológica, especialmente quando interferem em mecanismos de regulação gênica ou processamento do RNA.

A Figura 3 apresenta uma representação ilustrativa de diferentes categorias utilizadas na caracterização de variantes genéticas, incluindo tipos de alterações, consequências funcionais, distribuição cromossômica e localização genômica. Essa

representação demonstra possibilidades de organização dos resultados produzidos por ferramentas de anotação, mas os quantitativos apresentados não foram derivados de um conjunto de dados genômicos efetivamente analisado.

Figura 3: Representação ilustrativa da distribuição e da anotação funcional de variantes genéticas em análises de sequenciamento de nova geração.



Legenda: (A) Distribuição exemplificativa de diferentes tipos de variantes. (B) Categorias de consequências moleculares utilizadas na anotação funcional. (C) Representação esquemática da distribuição de variantes entre cromossomos humanos. (D) Categorias de localização genômica das alterações identificadas. Os valores apresentados são exclusivamente ilustrativos e não constituem resultados quantitativos de sequenciamento.

Fonte: Elaboração própria (2026).

A interpretação dessas categorias exige considerar que a frequência de variantes em determinadas regiões genômicas depende da abrangência do sequenciamento, das características do genoma de referência e dos critérios de anotação empregados. Em análises de exoma, por exemplo, espera-se maior representação de regiões codificantes quando comparadas às regiões não codificantes, enquanto análises de genoma completo permitem uma investigação mais abrangente.

Outro aspecto relevante envolve a distinção entre consequência molecular prevista e significado clínico. Uma alteração classificada como *missense* pode modificar um aminoácido sem necessariamente comprometer a função da proteína. Da mesma forma, variantes que introduzem códons de parada ou alteram o processamento do RNA podem

apresentar consequências distintas dependendo da localização, do transcrito envolvido e dos mecanismos celulares de compensação.

Consequentemente, a anotação funcional deve ser compreendida como uma etapa de caracterização e priorização, e não como confirmação automática de patogenicidade. A atribuição de relevância clínica exige integração com evidências adicionais, incluindo dados funcionais, informações populacionais, mecanismos conhecidos de doença e compatibilidade entre genótipo e fenótipo.

Estudo de caso: relevância da variante p.Phe508del no gene CFTR

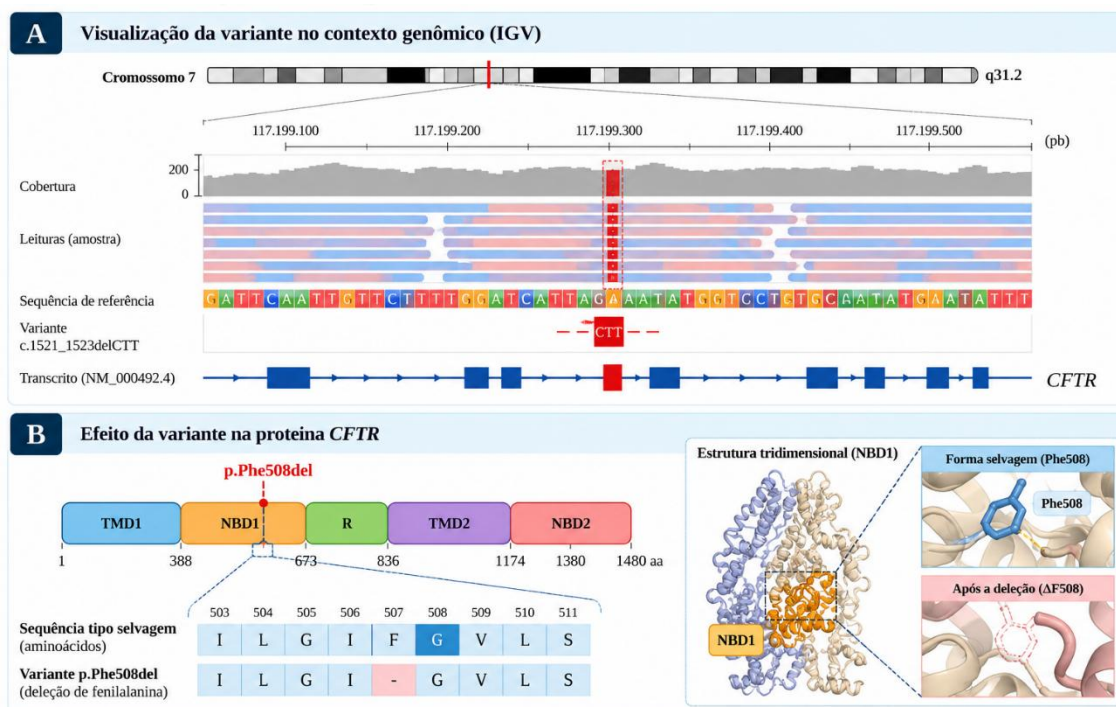
O gene *CFTR*, localizado no cromossomo 7, codifica uma proteína transmembrana responsável pela regulação do transporte de íons em diferentes tecidos epiteliais. Alterações que comprometem sua expressão, processamento ou atividade podem resultar em disfunções no transporte de cloreto e bicarbonato, contribuindo para manifestações clínicas características da fibrose cística.

Entre as variantes associadas à doença, destaca-se a deleção p.Phe508del, tradicionalmente denominada F508del ou $\Delta F508$. Essa alteração corresponde à perda de uma fenilalanina na posição 508 da proteína CFTR e constitui uma das variantes mais estudadas na investigação molecular da fibrose cística.

A relevância dessa alteração está relacionada ao seu impacto sobre o dobramento e o processamento intracelular da proteína. Estudos estruturais e funcionais demonstram que a ausência da fenilalanina compromete a estabilidade conformacional e as interações entre domínios da CFTR, favorecendo sua retenção no retículo endoplasmático e posterior degradação. Shishido *et al.* (2020) demonstraram que diferentes variantes associadas a defeitos de processamento podem interferir em etapas específicas do dobramento cotraducional, contribuindo para compreender a diversidade dos mecanismos moleculares envolvidos na doença.

A Figura 4 apresenta uma representação esquemática da localização da variante p.Phe508del e de suas possíveis consequências estruturais. O painel A ilustra como uma alteração pode ser representada no contexto de leituras genômicas, enquanto o painel B demonstra sua localização no domínio NBD1 da proteína CFTR. A visualização semelhante à interface de um navegador genômico não foi gerada a partir de alinhamentos reais e, portanto, não deve ser interpretada como evidência de detecção da variante em uma amostra.

Figura 4: Representação esquemática da variante p.Phe508del no gene CFTR e de suas consequências moleculares.



Legenda: (A) Ilustração conceitual da localização de uma variante no contexto genômico, utilizando elementos visuais semelhantes aos empregados em navegadores de alinhamento de sequências. A representação não corresponde a dados reais de sequenciamento ou à confirmação experimental de uma variante. (B) Representação da organização dos principais domínios da proteína CFTR, destacando a região NBD1 e a deleção da fenilalanina na posição 508. A alteração está associada a defeitos no dobramento e no processamento intracelular da proteína. Fonte: Elaboração própria (2026), com base na literatura científica sobre a estrutura e a função da CFTR.

A análise da variante p.Phe508del evidencia a importância de integrar informações genéticas e funcionais para compreender a relação entre alterações na sequência do DNA e manifestações clínicas. A identificação de uma deleção específica permite formular hipóteses sobre suas consequências moleculares, mas a interpretação diagnóstica depende de sua associação com o fenótipo e da consideração do padrão de herança.

Sosnay et al. (2013), ao investigarem a relevância clínica de variantes do gene *CFTR*, demonstraram a importância da integração entre informações genotípicas, dados funcionais e características clínicas para determinar sua associação com a fibrose cística. Esse trabalho contribuiu para evidenciar que diferentes variantes de um mesmo gene podem apresentar consequências distintas, dificultando interpretações baseadas exclusivamente na localização ou no tipo de alteração molecular.

No caso da fibrose cística, a identificação de uma única variante patogênica em heterozigose não estabelece, isoladamente, o diagnóstico molecular de uma condição autossômica recessiva. A interpretação deve considerar a identificação de alterações nos dois alelos, sua disposição em *trans*, quando pertinente, e a compatibilidade com os achados clínicos e laboratoriais.

Além disso, a investigação molecular pode apresentar limitações relacionadas à cobertura de determinadas regiões, à detecção de variantes estruturais e à identificação de alterações intrônicas profundas. Portanto, um resultado negativo em determinado ensaio de NGS não exclui necessariamente a presença de uma condição genética, especialmente quando a suspeita clínica permanece elevada.

Contribuições da bioinformática para o diagnóstico molecular e a medicina de precisão

A integração entre sequenciamento de nova geração e ferramentas bioinformáticas ampliou as possibilidades de investigação de doenças genéticas, permitindo examinar simultaneamente múltiplas regiões genômicas e organizar grandes volumes de informações moleculares. Entretanto, a interpretação dessas informações exige uma abordagem que considere tanto a qualidade técnica dos dados quanto a relevância biológica das alterações identificadas.

O estudo de caso envolvendo o gene *CFTR* demonstra como o conhecimento sobre uma variante específica pode contribuir para compreender mecanismos moleculares de doença. A relação entre a alteração p.Phe508del, os defeitos de processamento proteico e a redução da função do canal constitui um exemplo da importância da caracterização funcional das variantes.

Essa compreensão também apresenta implicações para a medicina de precisão, uma vez que determinadas estratégias terapêuticas podem ser direcionadas a mecanismos moleculares específicos. Entretanto, a possibilidade de utilização de tratamentos individualizados depende de evidências clínicas, critérios de elegibilidade e avaliação especializada, não sendo determinada exclusivamente pela identificação de uma alteração genética.

A investigação genômica também apresenta potencial para reduzir incertezas diagnósticas, orientar a investigação familiar e contribuir para o aconselhamento genético. Contudo, a utilização clínica dos resultados exige atenção às limitações das

ferramentas computacionais, à possibilidade de achados incidentais e à necessidade de interpretação contextualizada.

Nesse sentido, a bioinformática deve ser compreendida como uma ferramenta de apoio à investigação molecular, cuja contribuição depende da integração entre dados genômicos, conhecimento científico e avaliação clínica. A utilização responsável dessas tecnologias pode favorecer avanços na identificação de doenças humanas, desde que seus resultados sejam interpretados dentro dos limites das evidências disponíveis.

Considerações Finais

O presente capítulo analisou as contribuições do sequenciamento de nova geração (NGS) e da bioinformática para o diagnóstico molecular de doenças humanas, utilizando a fibrose cística e o gene *CFTR* como modelo demonstrativo. A abordagem evidenciou a importância da integração entre controle de qualidade, identificação de variantes, anotação funcional e interpretação clínica para a compreensão das alterações genéticas associadas a processos patológicos.

A análise da variante p.Phe508del exemplificou a relação entre modificações genéticas, alterações na estrutura e no processamento proteico e mecanismos fisiopatológicos. Entretanto, destacou-se que a identificação de variantes não determina isoladamente sua relevância clínica, sendo necessária a integração de evidências moleculares, funcionais e fenotípicas. Ressalta-se que o estudo apresentou caráter demonstrativo, fundamentado em evidências científicas publicadas, sem a execução de análises originais de sequenciamento.

Conclui-se que a associação entre NGS e bioinformática apresenta potencial para aprimorar a investigação de doenças genéticas e contribuir para o desenvolvimento da medicina de precisão. Contudo, sua aplicação clínica exige rigor metodológico, validação dos resultados e interpretação interdisciplinar, visando diagnósticos mais precisos e estratégias de cuidado individualizadas.

Declaração de Uso de Inteligência Artificial Generativa

Durante a elaboração deste capítulo, utilizou-se o ChatGPT (OpenAI) como ferramenta auxiliar na organização e redação preliminar do texto e na revisão linguística. Os conteúdos gerados foram submetidos à análise crítica e à revisão dos autores, incluindo a verificação das informações e das referências bibliográficas. A interpretação dos estudos, as decisões editoriais e a aprovação da versão final permaneceram sob

responsabilidade dos autores, que assumem integral responsabilidade pela precisão, originalidade e integridade científica do trabalho.

Referências

BICK, David *et al.* Genome sequencing for diagnosing rare diseases. **The New England Journal of Medicine**, v. 390, 2024. DOI: <https://doi.org/10.1056/NEJMoa2314761>.

GOODWIN, Sara; MCPHERSON, John D.; MCCOMBIE, W. Richard. Coming of age: ten years of next-generation sequencing technologies. **Nature Reviews Genetics**, v. 17, n. 6, p. 333-351, 2016. DOI: <https://doi.org/10.1038/nrg.2016.49>.

MANOLIO, Teri A. *et al.* Opportunities, resources, and techniques for implementing genomics in clinical care. **The Lancet**, v. 394, n. 10197, p. 511-520, 2019. DOI: [https://doi.org/10.1016/S0140-6736\(19\)31140-7](https://doi.org/10.1016/S0140-6736(19)31140-7).

RICHARDS, Sue *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. **Genetics in Medicine**, v. 17, n. 5, p. 405-424, 2015. DOI: <https://doi.org/10.1038/gim.2015.30>.

SHISHIDO, Hideki *et al.* CFTR trafficking mutations disrupt cotranslational protein folding by targeting biosynthetic intermediates. **Nature Communications**, v. 11, art. 4258, 2020. DOI: <https://doi.org/10.1038/s41467-020-18101-8>.

SOSNAY, Patrick R. *et al.* Defining the disease liability of variants in the cystic fibrosis transmembrane conductance regulator gene. **Nature Genetics**, v. 45, p. 1160-1167, 2013. DOI: <https://doi.org/10.1038/ng.2745>.

CAPÍTULO 14

PUBLIQUE COM A SCIENCE EM FLUXO CONTÍNUO

PUBLISH WITH SCIENCE IN CONTINUOUS FLOW

DOI: <https://doi.org/10.56001/26.9786501906539.14>

Submetido em: 15/12/2025

Revisado em: 28/12/2025

Publicado em: 29/01/2026

AUTORES

Universidade Federal do Brasil, Faculdade de Ciências, Localidade-PE

<http://lattes.cnpq.br/>

AUTORES

Universidade Estadual do Brasil, Centro de Ciências, Localidade-PB

<https://orcid.org/>

AUTORES

Instituto Federal do Brasil, Departamento de Ciências, Localidade-SE

<http://lattes.cnpq.br/>

Resumo

Texto

Palavras-chave: Words.

Abstract

Texto

Keywords: Words.

Introdução

Aqui começa sua publicação e história de sucesso.

CAPÍTULO 15

PUBLIQUE COM A SCIENCE EM FLUXO CONTÍNUO

PUBLISH WITH SCIENCE IN CONTINUOUS FLOW

DOI: <https://doi.org/10.56001/26.9786501906539.15>

Submetido em: 15/12/2025

Revisado em: 28/12/2025

Publicado em: 29/01/2026

AUTORES

Universidade Federal do Brasil, Faculdade de Ciências, Localidade-PE

<http://lattes.cnpq.br/>

AUTORES

Universidade Estadual do Brasil, Centro de Ciências, Localidade-PB

<https://orcid.org/>

AUTORES

Instituto Federal do Brasil, Departamento de Ciências, Localidade-SE

<http://lattes.cnpq.br/>

Resumo

Texto

Palavras-chave: Words.

Abstract

Texto

Keywords: Words.

Introdução

Aqui começa sua publicação e história de sucesso.

SOBRE OS ORGANIZADORES DO LIVRO DADOS CNPQ:

Prof. Dr. Igor Luiz Vieira de Lima Santos



Possui Graduação em Bacharelado em Ciências Biológicas pela Universidade Federal Rural de Pernambuco (2003) e Mestrado em Genética e Biologia Molecular pela Universidade Federal do Rio Grande do Norte (2006). Doutor em Biotecnologia pela RENORBIO (Rede Nordeste de Biotecnologia (2013), Área de Concentração Biotecnologia em Saúde atuando principalmente com pesquisa relacionada a genética do câncer de mama. Participou como Bolsista de Desenvolvimento Tecnológico Industrial Nível 3 de relevantes projetos tais como: Projeto Genoma *Anopheles darlingi* (de 02/2008 a 02/2009); e Isolamento de genes de interesse biotecnológico para a agricultura (de 08/2009 a 12/2009). Atualmente é Professor Adjunto III da Universidade Federal de Campina Grande-UFCG, do Centro de Educação e Saúde onde é Líder do Grupo de Pesquisa BASE (Biotecnologia Aplicada à Saúde e Educação) e colaborador em ensino e pesquisa da UFRPE, UFRN e EMBRAPA-CNPA. Tem experiência nas diversas áreas da Genética, Fisiologia Molecular, Microbiologia e Bioquímica com ênfase em Genética Molecular e de Microrganismos, Plantas e Animais, Biologia Molecular e Biotecnologia Industrial. Atua em projetos versando principalmente sobre os seguintes temas: Metagenômica, Carcinogênese, Monitoramento Ambiental e Genética Molecular, Marcadores Moleculares Genéticos, Polimorfismos Genéticos, Bioinformática, Biodegradação, Biotecnologia Industrial e Aplicada, Sequenciamento de DNA, Nutrigenômica, Farmacogenômica, Genética na Enfermagem e Educação.

Pós-Dra. Carliane Rebeca Coelho da Silva



Possui Graduação em Bacharelado em Ciências Biológicas pela Universidade Federal Rural de Pernambuco apresentando monografia na área de genética com enfoque em transgenia. Mestrado em Melhoramento Genético de Plantas pela Universidade Federal do Rural de Pernambuco com dissertação na área de melhoramento genético com enfoque em técnicas de imunodeteção. Doutora em Biotecnologia pela RENORBIO (Rede Nordeste de Biotecnologia, Área de Concentração Biotecnologia em Agropecuária) atuando principalmente com tema relacionado a transgenia de plantas. Pós-doutorado em Biotecnologia com concentração na área de Biotecnologia em Agropecuária. Atua com linhas de pesquisa focalizadas nas áreas de defesa de plantas contra estresses bióticos e abióticos, com suporte de ferramentas biotecnológicas e do melhoramento genético. Tem experiência na área de Engenharia Genética, com ênfase em isolamento de genes, expressão em plantas, melhoramento genético de plantas via transgenia, marcadores moleculares e com práticas de transformação de plantas via "ovary drip". Tem experiência na área de genética molecular, com ênfase nos estudos de transcritos, expressão diferencial e expressão gênica. Integra uma equipe com pesquisadores de diferentes instituições como Embrapa Algodão, UFRPE, UEPB e UFPB, participando de diversos projetos com enfoque no melhoramento de plantas.

Prof. Dr. Rafael Trindade Maia



Graduação em Licenciatura Plena Em Ciências Biológicas pela Universidade Federal Rural de Pernambuco (2005), mestrado em Genética, Conservação e Biologia Evolutiva pelo Instituto Nacional de Pesquisas da Amazônia (2008) e doutorado em Biologia Animal pela Universidade Federal de Pernambuco (2013). Foi professor substituto do departamento de biologia da Universidade Federal Rural de Pernambuco (2010). Tem experiência no ensino fundamental II, ensino médio e cursos pré-Enem. Atualmente é professor do Centro de Desenvolvimento Sustentável do Semiárido (CDSA) da Universidade Federal de Campina Grande (UFCG). É docente do mestrado no Programa de Pós-graduação em Ciências Naturais e Biotecnologia de Centro de Educação e Saúde (CES) da Universidade Federal de Campina Grande (UFCG) e do Programa de Pós-graduação em Biologia Celular e Molecular (UFPB). Tem experiência com genética de populações, filogenia, bioinformática, docking molecular, modelagem e dinâmica molecular de proteínas. Atua na área de ensino de ciências e biologia. Lidera os grupos de pesquisa Biologia

SOBRE OS ORGANIZADORES DO LIVRO

Computacional e Teórica (BCT) e Ensino de Ciências e Biologia (ECB). É editor acadêmico e revisor dos periódicos Asian Journal of Biotechnology and Genetic Engineering, da Revista Thema, do Journal of Biomolecular Structure and Dynamics, da Molecular Biotechnology, Molecular Diversity e da Editora internacional de livros científicos IntechOpen.

BIOINFORMÁTICA UMA CIÊNCIA SEM FRONTEIRAS

“Esperamos que tenham aproveitado todos os trabalhos disponíveis na íntegra e gratuitos para seu conhecimento e consulta.

Esta obra objetivou ampliar os seus horizontes sobre a temática proposta além dos muros acadêmicos, proporcionando uma visão mais realista, ampla e multidisciplinar desta área de estudo seus impactos e descobertas.

Os livros da Science compreendem do conhecimento mais simples ao mais complexo, do mais acadêmico ao mais aplicado, procurando sempre a socialização global com conhecimento científico respaldado e de qualidade, para que a sociedade possa se beneficiar em todos os sentidos.

Agradecemos o seu interesse em chegar até o final deste livro na busca por conhecimento. Aguardem novos títulos e eventos da Editora Science sempre comprometida com a qualidade e o sucesso da sua publicação.”

PARA MAIS INFORMAÇÕES E OBRAS DA EDITORA SCIENCE ACESSE:

www.editorascience.com.br

Siga nossas redes sociais e amplie o alcance dos nossos livros:

Facebook: <http://www.facebook.com/editorascience>

Instagram: <https://www.instagram.com/editorascience>



Todos os Direitos Reservados

