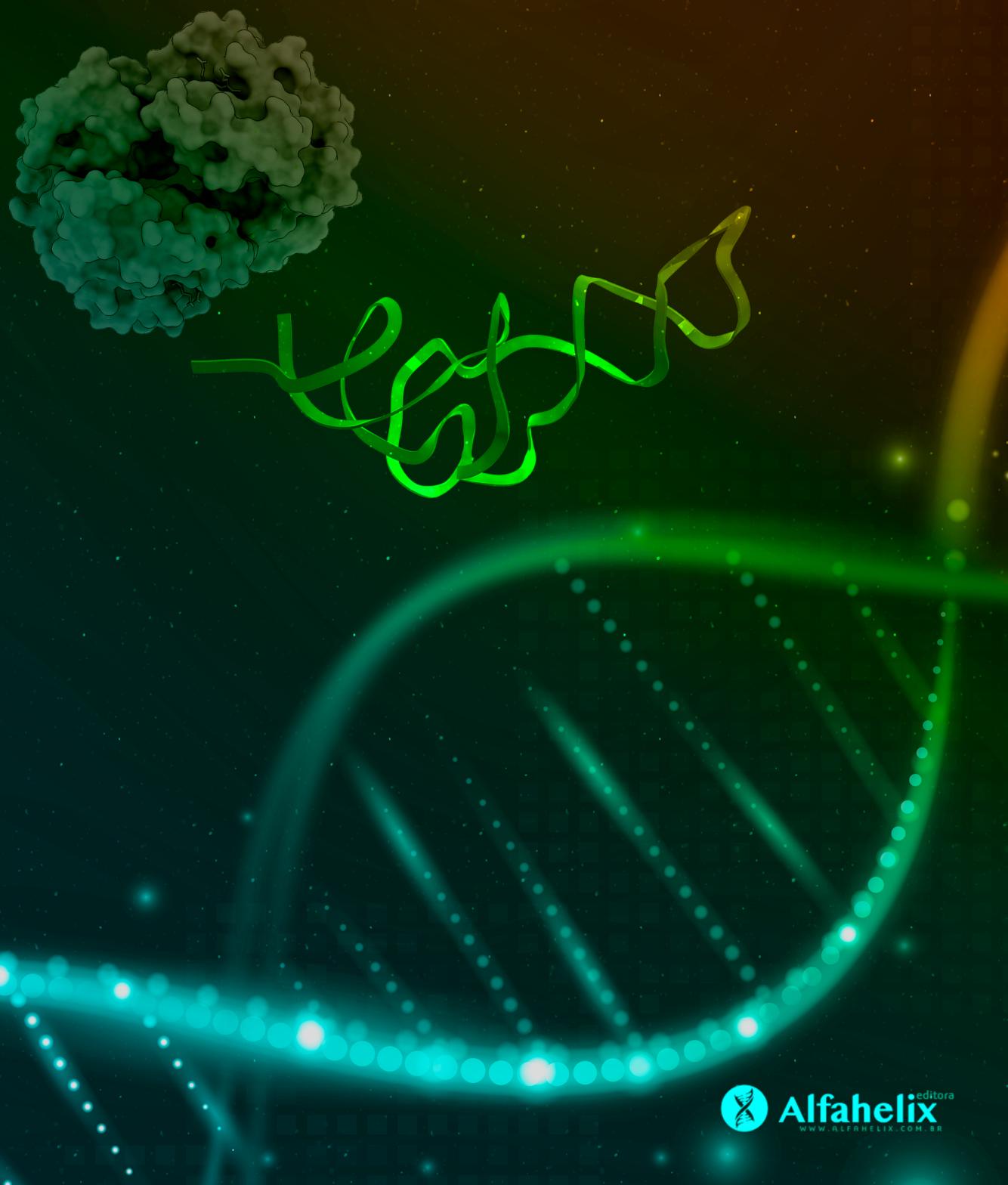




# bioinfo

[www.bioinfo.com.br](http://www.bioinfo.com.br)

Revista Brasileira de Bioinformática  
e Biologia Computacional



**BIOINFO - Revista Brasileira de Bioinformática e Biologia  
Computacional**

ISBN: 978-65-992753-8-8 | doi: 10.51780/978-65-992753-8-8

[www.bioinfo.com.br](http://www.bioinfo.com.br)

Vol. 3 - Set. 2023

**Organização & Revisão**

-  Alessandra Lima
-  Aline de Paula Dias da Silva
-  Aline Sampaio Cremonesi
-  Ana Carolina Silva Bulla
-  Ana Paula Abreu
-  Ariany Rosa Gonçalves
-  Bárbara Rebeca de Macedo Pinheiro
-  Bibiana Fam
-  Bruna Espiño dos Santos
-  Carlos Capelini
-  Diego Lucas Neres Rodrigues
-  Filipe Augusto Teixeira
-  Isaac Farias Cansanção
-  Izabela Mamede
-  Luana Luiza Bastos
-  Lucianna Helene Santos
-  Marcos Antonio Nobrega de Sousa
-  Mira Raya Paula de Lima
-  Rafael Pereira Lemos
-  Savio Costa
-  Thiago M. N. de Camargo
-  Tiago Cabral Borelli
-  Wylerson Nogueira

## Organização & Mídias sociais

-  Angie Atoche Puelles
-  Emília Sousa de Oliveira
-  Glenerson Baptista
-  Joalisson da Costa Moreira
-  Luana Gabrielli Ayres Schekiera
-  Vivian Morais Paixão

## *Editor-in-chief*

Diego Mariano 

Residente pós-doutoral | Departamento de Ciência da Computação (UFMG) | Editor-in-chief Alfahelix

### Ficha catalográfica

Sandro Alex Batista CRB6/2433

Bibliotecário

### Diagramação

Alessandra Lima

Angie Atoche Puelles

Luana Luiza Bastos

Bibiana Fam

Bruna Espiño dos Santos

Diego Mariano

Marcos Antonio Nobrega de Sousa

Thiago de Camargo

### Publicação

Editora Alfahelix, CNPJ: 37.524.984/0001-10

Lagoa Santa, MG, Brasil

[www.alfahelix.com.br](http://www.alfahelix.com.br)

### Capa

Adaptado de rawpixel.com/Freepik. Estruturas da hemoglobina e do tRNA foram obtidas no PDB e renderizadas com ChimeraX.

### Agradecimentos

FAPEMIG - Fundação de Amparo à Pesquisa do Estado de Minas Gerais.

M333m MARIANO, Diego *et al.* (org.).

BIOINFO #03 - Revista Brasileira de bioinformática e Biologia Computacional  
/ Organização de Diego Mariano. Lagoa Santa, MG: Alfahelix, Set. 2023. V.3.  
300 p. il.: imagens P&B e col.

E-book.

ISBN: 978-65-992753-8-8

DOI: 10.51780/97865992753-8-8

1. Bioinformática. 2. Computação. 3. Biologia de sistemas. 4. Evolução.  
5. Biologia estrutural. 6. Aline de Paula dias da Silva *et al.* (orgs.). I. Título.

CDD: 006.5765

CDU: 004.89/576



## BIOINFO: *free & open access*

*Sem custos para autores. Sem custos para leitores.*

Este livro compila os artigos publicados na Revista BIOINFO Vol. 3 (2023). Todos os direitos autorais pertencem aos autores de cada respectivo artigo. Todavia, os autores concordam em compartilhar gratuitamente o conteúdo deste livro e incentivam sua livre distribuição (desde que os autores dos respectivos capítulos sejam corretamente citados e/ou que as leis de uso justo sejam respeitadas). Este livro está compartilhado sob a licença Creative Commons Atribuição 4.0 Internacional - Não comercial (CC BY-NC 4.0). Você pode utilizar qualquer conteúdo aqui apresentado, desde que cite:

Cite este artigo 0.1

MARIANO, DCB (org.) et al. BIOINFO #03 - Revista Brasileira de Bioinformática e Biologia Computacional. 3. Ed. Vol. 3. ISBN: 978-65-992753-8-8 Lagoa Santa: Alfahelix, 2023. doi:  
10.51780/978-65-992753-8-8

Esta é uma obra digital. Note que fazemos uso de links e outras propriedades de hipertexto, o que pode limitar a qualidade de edições impressas deste manuscrito. Entretanto, sinta-se à vontade caso deseje imprimir este conteúdo. A cópia por meio impresso ou digital (PDF) é permitida, sendo exclusivamente vedada a venda visando lucro. Detalhes sobre a licença de uso estão disponíveis em: <https://bioinfo.com.br/licenca-de-uso/>

## Artigos & Autores - BIOINFO #03 (Set. 2023)

#	Categoria	Artigo	Autores	doi
<b>0</b>	Editorial	Editorial – BIOINFO #03	Diego Mariano	<a href="https://doi.org/10.51780/bioinfo-03-00">10.51780/bioinfo-03-00</a>
<b>1</b>	Ensino	Potencialidades do uso da bioinformática como ferramenta de ensino	Francisco de Sousa, Antônia Pedro, Bianca Araújo, Tarcisio Coutinho	<a href="https://doi.org/10.51780/bioinfo-03-01">10.51780/bioinfo-03-01</a>
<b>2</b>	Genômica	Metagenômica e Amplicon: perguntas frequentes e respostas essenciais	Sávio Costa	<a href="https://doi.org/10.51780/bioinfo-03-02">10.51780/bioinfo-03-02</a>
<b>3</b>	Genômica	O problema da nomeação dos genes	Izabela Mamede	<a href="https://doi.org/10.51780/bioinfo-03-03">10.51780/bioinfo-03-03</a>
<b>4</b>	Virologia	O Código COVID: A investigação da pandemia da COVID-19 através da bioinformática	Aline de Paula Dias da Silva	<a href="https://doi.org/10.51780/bioinfo-03-04">10.51780/bioinfo-03-04</a>
<b>5</b>	Genômica	Desafios na padronização da anotação genômica	Diego Lucas Neres Rodrigues	<a href="https://doi.org/10.51780/bioinfo-03-05">10.51780/bioinfo-03-05</a>
<b>6</b>	Virologia	Desafiando a resistência antimicrobiana: o potencial terapêutico da fagoterapia	Bruna Espiño dos Santos	<a href="https://doi.org/10.51780/bioinfo-03-06">10.51780/bioinfo-03-06</a>
<b>7</b>	Genômica	A Bioinformática como aliada da Biotecnologia Agrícola	Ariany Rosa Gonçalves	<a href="https://doi.org/10.51780/bioinfo-03-07">10.51780/bioinfo-03-07</a>
<b>8</b>	Virologia	A bioinformática na era pré e pós-pandemia	Bibiana Fam	<a href="https://doi.org/10.51780/bioinfo-03-08">10.51780/bioinfo-03-08</a>
<b>9</b>	Sequenciamento	Uso do BLASTn na construção de primers	Bárbara Rebeca de M. Pinheiro	<a href="https://doi.org/10.51780/bioinfo-03-09">10.51780/bioinfo-03-09</a>
<b>10</b>	Bancos de dados	Bioinformática na luta contra o câncer: os bancos de dados na pesquisa oncológica	Thayanne Costa, Lara Vitória da Costa Bezerra	<a href="https://doi.org/10.51780/bioinfo-03-10">10.51780/bioinfo-03-10</a>
<b>11</b>	Bioinformática estrutural	Introdução à Biologia Estrutural de Proteínas	Rafael Pereira Lemos, Paulo Santos, Aline Rocha	<a href="https://doi.org/10.51780/bioinfo-03-11">10.51780/bioinfo-03-11</a>
<b>12</b>	Bioinformática estrutural / Tutoriais	Extração de Informações de Sequências e Estruturas de Proteínas	Rafael Pereira Lemos, Paulo Aline Rocha	<a href="https://doi.org/10.51780/bioinfo-03-12">10.51780/bioinfo-03-12</a>
<b>13</b>	Transcriptômica	Um mundo dentro de nós: explorando a microbiota humana através da metatranscriptômica	Aline de Paula Dias da Silva, Monique Cristina dos Santos	<a href="https://doi.org/10.51780/bioinfo-03-13">10.51780/bioinfo-03-13</a>

#	Categoría	Artigo	Autores	doi
14	Ensino	Bioinformática como uma ferramenta didática para o ensino da Genética	Marcos Sousa, Jeniffer Pereira, Ana Luíza Soares, Ana Beatriz Santos, Ricardo Henrique Silva, Arthur Medeiros, Bruna Araujo, Francisca Nobrega	<a href="https://doi.org/10.51780/bioinfo-03-14">10.51780/bioinfo-03-14</a>
15	Evolução	Lá e de volta outra vez: um pouco do passado, presente e futuro da evolução	Tiago Cabral Borelli	<a href="https://doi.org/10.51780/bioinfo-03-15">10.51780/bioinfo-03-15</a>
16	Bioinformática estrutural	De onde vêm as proteínas?	Alisson Silva, Bruno Nunes, Joicymara Xavier	<a href="https://doi.org/10.51780/bioinfo-03-16">10.51780/bioinfo-03-16</a>
17	Virologia	Vírus endógenos humanos: como analisá-los in silico?	Juan Diego Sampaio, João Gonçalves da Costa Neto, Isaac Farias Cansanção	<a href="https://doi.org/10.51780/bioinfo-03-17">10.51780/bioinfo-03-17</a>
18	Evolução	A Bioinformática e a Compreensão da Vida	Thiago Camargo	<a href="https://doi.org/10.51780/bioinfo-03-18">10.51780/bioinfo-03-18</a>
19	Biología de sistemas	Visão Integrativa da Biología de Sistemas	Vitor Lima Coelho	<a href="https://doi.org/10.51780/bioinfo-03-19">10.51780/bioinfo-03-19</a>
20	Bioinformática estrutural	A importância da docagem molecular no combate às bactérias multirresistentes	Aline Sampaio Cremonesi	<a href="https://doi.org/10.51780/bioinfo-03-20">10.51780/bioinfo-03-20</a>
21	Bioinformática estrutural / Tutoriais	Re-docking Molecular Utilizando o PyMOL e AutoDock VINA	Luana Luiza Bastos, Giovana Fiorini	<a href="https://doi.org/10.51780/bioinfo-03-21">10.51780/bioinfo-03-21</a>
22	Bioinformática estrutural / Tutoriais	ColabFold: uma ferramenta web para modelagem de proteínas	Giovana Fiorini, Luana Luiza Bastos, Rafael Pereira Lemos	<a href="https://doi.org/10.51780/bioinfo-03-22">10.51780/bioinfo-03-22</a>
23	Bioinformática estrutural / Tutoriais	AlphaFold 2: revolucionando a modelagem de estruturas 3D de macromoléculas	Vivian Moraes Paixão, Angie Puelles, Eduardo Moreira, Luana Bastos, Raquel Cardoso de Melo-Minardi	<a href="https://doi.org/10.51780/bioinfo-03-23">10.51780/bioinfo-03-23</a>
24	Bioinformática estrutural	Termodinâmica de Proteínas: como as proteínas se enovelam?	Alisson Silva, Bruno Nunes, Joicymara Xavier	<a href="https://doi.org/10.51780/bioinfo-03-24">10.51780/bioinfo-03-24</a>
25	Bioquímica	Avaliação ADMET de substâncias	Artur Gomes Barros	<a href="https://doi.org/10.51780/bioinfo-03-25">10.51780/bioinfo-03-25</a>
26	Computação	Perspectiva histórica de grandes eventos da ciência da computação e na biologia molecular: segunda guerra e guerra fria	Monique Cristina dos Santos, Aline de Paula Dias da Silva	<a href="https://doi.org/10.51780/bioinfo-03-26">10.51780/bioinfo-03-26</a>

# SUMÁRIO

## CAPÍTULO 1

### POTENCIALIDADES DO USO DA BIOINFORMÁTICA COMO

#### FERRAMENTA DE ENSINO

PÁGINA 22

1.1	Introdução .....	23
1.2	Como inserir? .....	24
1.3	A bioinformática não é para toda aula.....	25
1.4	Explore textos científicos e/ou jornalísticos .....	25
1.5	Utilizando banco de dados .....	27
1.6	Alinhamento de sequências .....	29
1.7	Desafios para a inserção .....	29
1.8	Referências.....	31

## CAPÍTULO 2

### METAGENÔMICA E AMPICON: PERGUNTAS FREQUENTES E

#### RESPOSTAS ESSENCIAIS

PÁGINA 35

2.1	O que é Metagenômica?.....	36
2.2	Metagenômica e ampicon? São as mesmas coisas? .....	37
2.3	Como se analisa dados de metagenômica <i>Shotgun</i> ?.....	38
2.4	Como se analisa dados de ampicon? .....	40

2.5	Referências .....	42
-----	-------------------	----

## CAPÍTULO 3

### O PROBLEMA DA NOMEAÇÃO DOS GENES

PÁGINA 46

3.1	Referências .....	49
-----	-------------------	----

## CAPÍTULO 4

### O CÓDIGO COVID: A INVESTIGAÇÃO DA PANDEMIA DA COVID-

19 ATRAVÉS DA BIOINFORMÁTICA

PÁGINA

50

4.1	Referências .....	53
-----	-------------------	----

## CAPÍTULO 5

### DESAFIOS NA PADRONIZAÇÃO DA ANOTAÇÃO GENÔMICA

PÁGINA 54

5.1	Referências .....	58
-----	-------------------	----

## CAPÍTULO 6

### DESAFIANDO A RESISTÊNCIA ANTIMICROBIANA: O POTENCIAL

TERAPÊUTICO DA FAGOTERAPIA

PÁGINA 60

6.1	O que são bacteriófagos? .....	61
6.2	Funcionamento da terapia .....	61
6.3	A fagoterapia é secular .....	63
6.4	Vantagens da fagoterapia .....	64
6.5	O papel da bioinformática .....	64
6.6	Referências .....	65

## CAPÍTULO 7

### A BIOINFORMÁTICA COMO ALIADA DA BIOTECNOLOGIA

AGRÍCOLA

PÁGINA 67

7.1	Referências .....	70
-----	-------------------	----

**CAPÍTULO 8****A BIOINFORMÁTICA NA ERA PRÉ E PÓS-PANDEMIA** \_\_\_\_\_ **PÁGINA 74**

8.1 Referências .....	77
-----------------------	----

**CAPÍTULO 9****USO DO BLASTN NA CONSTRUÇÃO DE PRIMERS** \_\_\_\_\_ **PÁGINA 79**

9.1 Referências .....	82
-----------------------	----

**CAPÍTULO 10****BIOINFORMÁTICA NA LUTA CONTRA O CÂNCER: OS BANCOS DE DADOS NA PESQUISA ONCOLÓGICA** \_\_\_\_\_ **PÁGINA 83**

10.1 Introdução .....	84
10.2 A Bioinformática na Oncologia .....	85
10.2.1 Análise de expressão gênica: .....	86
10.2.2 Análises em proteômica: .....	87
10.2.3 Predição e Mineração de Dados: .....	87
10.3 Por que os bancos de dados são importantes na oncologia? .....	88
10.4 Referências .....	91

**CAPÍTULO 11****INTRODUÇÃO À BIOLOGIA ESTRUTURAL DE PROTEÍNAS****PÁGINA 94**

11.1 Aminoácidos .....	95
11.2 Estrutura proteica .....	97
11.3 Bases de dados usadas em biologia estrutural .....	100
11.4 Referências .....	101

**CAPÍTULO 12****EXTRAÇÃO DE INFORMAÇÕES DE SEQUÊNCIAS E ESTRUTURAS****DE PROTEÍNAS****PÁGINA 103**

12.1 Parte 1 – Identificação da Sequência .....	104
12.2 Parte 2 – Uniprot.....	108
12.3 Parte 3 – Características Físico-Químicas.....	111

12.4	Parte 4 – Predição de Estruturas Secundárias .....	114
12.5	Parte 5 – Identificação de Desordem Estrutural e Domínios Transmembrana .....	118
12.6	Parte 6 – Obtenção de Estruturas 3D .....	122
12.7	Parte 7 – O Formato PDB .....	127
12.8	Parte 8 – Validação Estrutural.....	131
12.9	Parte 9 – Visualização de Estruturas.....	136
12.10	Parte 10 – Introdução ao AlphaFold .....	141
12.11	Parte 11 – Ferramentas Extras .....	145
12.12	Referências.....	147

## CAPÍTULO 13

### **UM MUNDO DENTRO DE NÓS: EXPLORANDO A MICROBIOTA HUMANA ATRAVÉS DA METATRANSCRIPTÔMICA** \_\_\_\_\_ PÁGINA 149

13.1	Por que fazer metatranscriptômica? .....	151
13.2	E como estudar o microbioma através da Bioinformática?.....	153
13.3	Não sou Bioinformata! E agora? .....	153
13.4	O futuro das pesquisas sobre o microbioma humano .....	154
13.5	Referências.....	155

## CAPÍTULO 14

### **BIOINFORMÁTICA COMO UMA FERRAMENTA DIDÁTICA PARA O ENSINO DA GENÉTICA** \_\_\_\_\_ PÁGINA 156

14.1	<i>Abstract</i> .....	157
14.2	Introdução .....	158
	14.2.1 Aspectos metodológicos da atividade didática .....	159
	14.2.2 1) Obtenção das sequências .....	161
	14.2.3 2) Alinhamento das sequências .....	162
14.3	Resultados e discussão .....	164
14.4	Conclusão.....	169
14.5	Referências.....	169

**CAPÍTULO 15****LÁ E DE VOLTA OUTRA VEZ: UM POUCO DO PASSADO, PRESENTE****E FUTURO DA EVOLUÇÃO****PÁGINA 171**

- 15.1 Referências ..... 173

**CAPÍTULO 16****DE ONDE VÊM AS PROTEÍNAS?****PÁGINA 174**

- 16.1 Afinal, o que são proteínas? ..... 175  
16.2 Síntese de proteínas ..... 176  
16.3 Estruturas de proteínas ..... 179  
16.4 Conclusão ..... 182  
16.5 Referências ..... 182

**CAPÍTULO 17****VÍRUS ENDÓGENOS HUMANOS: COMO ANALISÁ-LOS *in silico*?****PÁGINA 184**

- 17.1 Introdução ..... 185  
17.2 Metodologia ..... 186  
17.3 Resultados e Discussão ..... 187  
17.4 Conclusão ..... 191  
17.5 Referências ..... 192

**CAPÍTULO 18****A BIOINFORMÁTICA E A COMPREENSÃO DA VIDA****PÁGINA 194**

- 18.1 Referências ..... 196

**CAPÍTULO 19****VISÃO INTEGRATIVA DA BIOLOGIA DE SISTEMAS****PÁGINA 198**

- 19.1 Referências ..... 201

**CAPÍTULO 20****A IMPORTÂNCIA DA DOCAGEM MOLECULAR NO COMBATE ÀS****BACTÉRIAS MULTIRRESISTENTES****PÁGINA 202**

20.1 Referências .....	206
------------------------	-----

**CAPÍTULO 21****RE-DOCKING MOLECULAR UTILIZANDO O PyMOL E****AUTODOCK VINA****PÁGINA 208**

21.1 Instalando os programas .....	210
21.1.1 AUTODOCKTOOLS – ADT .....	210
21.1.2 AUTODOCK VINA .....	213
21.1.3 PyMOL .....	213
21.2 PLUGIN AUTODOCK VINA PYMOL.....	214
21.3 Re-docking com o PyMOL.....	215
21.4 Conclusões.....	225
21.5 Referências.....	225

**CAPÍTULO 22****COLABFOLD: UMA FERRAMENTA WEB PARA MODELAGEM DE****PROTEÍNAS****PÁGINA 227**

22.1 Modelagem de proteínas e o AlphaFold .....	228
22.2 ColabFold .....	230
22.2.1 Versões do ColabFold .....	233
22.2.2 Conhecendo a interface do ColabFold .....	235
22.3 Conclusão.....	241
22.4 Referências.....	242

**CAPÍTULO 23****ALPHAFOLD 2: REVOLUCIONANDO A MODELAGEM DE****ESTRUTURAS 3D DE MACROMOLÉCULAS****PÁGINA 243**

23.1 Introdução .....	244
23.2 AlphaFold 2 .....	246
23.2.1 Como funciona essa ferramenta? .....	247

23.2.2	Pré-processamento .....	247
23.2.3	Evoformer.....	249
23.2.4	Módulo de estrutura .....	250
23.3	AlphaFold 2 x ColabFold .....	252
23.4	Prática de modelagem .....	253
23.4.1	Etapa 1: Prepare seus dados .....	253
23.4.2	Etapa 2: Acesse e execute o AlphaFold 2 .....	255
23.4.3	Etapa 3: Analisar e interpretar os resultados .....	256
23.4.4	Etapa 4: Validação das estruturas previstas (Bônus) .....	259
23.5	Prática de modelagem de complexos .....	259
23.5.1	Etapa 1: Adquirindo as sequências .....	260
23.5.2	Etapa 2: Modelando o complexo.....	261
23.5.3	Etapa 3: Avaliando os resultados .....	262
23.6	Conclusão.....	264
23.7	Referências.....	265

## CAPÍTULO 24

### TERMODINÂMICA DE PROTEÍNAS: COMO AS PROTEÍNAS SE ENOVELAM? \_\_\_\_\_ PÁGINA 268

24.1	Técnicas de determinação de estruturas 3D .....	272
24.1.1	Técnicas experimentais .....	272
24.1.2	Difração de raios-X .....	272
24.1.3	Calorimetria de Varredura Diferencial .....	273
24.1.4	Ressonância Magnética Nuclear .....	273
24.1.5	Dicroísmo Circular .....	273
24.2	Métodos computacionais .....	274
24.2.1	Métodos de modelagem comparativa .....	274
24.3	Métodos de reconhecimento de padrões de enovelamento .....	275
24.4	Métodos ab initio ou de novo .....	275
24.4.1	AlphaFold.....	275
24.5	Conclusão.....	277
24.6	Referências.....	277

**CAPÍTULO 25****AVALIAÇÃO ADMET DE SUBSTÂNCIAS****PÁGINA 280**

25.1	Introdução .....	281
25.2	Como são realizados os testes ADMET <i>in silico</i> .....	282
25.3	Software para avaliações ADMET .....	283
25.3.1	SwissADME .....	283
25.3.2	pkCSM.....	286
25.4	Conclusão.....	288
25.5	Referências.....	289

**CAPÍTULO 26****PERSPECTIVA HISTÓRICA DE GRANDES EVENTOS DA CIÊNCIA****DA COMPUTAÇÃO E NA BIOLOGIA MOLECULAR: SEGUNDA****GUERRA E GUERRA FRIA****PÁGINA 291**

26.1	Quando tudo ainda era mato! .....	292
26.2	Chegou a Guerra Fria.....	294
26.3	Nova Era.....	298
26.4	Conclusão.....	298
26.5	Referências.....	299

# Editorial - BIOINFO #03

**A** Revista BIOINFO é a primeira revista brasileira voltada a publicação de trabalhos de divulgação científica em bioinformática e biologia computacional, publicada em língua portuguesa e voltada ao público universitário. De setembro de 2022 a setembro de 2023, a revista recebeu mais de 24 mil visitantes únicos. No mesmo período entre os anos de 2021 e 2022, o número de visitantes foi um pouco superior a 17 mil, o que indica um aumento de mais de 40% no total de acessos. Todos esses fatores destacam a importância que a revista adquiriu recentemente.

Nesta **terceira edição**, 26 artigos foram aceitos para a publicação. Os artigos foram publicados em 12 principais categorias. A categoria com maior número de submissões foi Bioinformática Estrutural (8), seguido por Virologia (4), Tutoriais (4), Genômica (4), Evolução (2), Ensino (2), Sequenciamento (1), Transcriptômica (1), Biologia de sistemas (1), Bioquímica (1), Bancos de dados (1) e Computação (1).

A principal novidade foi a subcategoria **Opiniões & Perspectivas** (OP), que recebeu 11 artigos de até 1000 palavras. Essa subcategoria foi criada para atender autores que desejavam submeter manuscritos de único autor descrevendo visões pessoais de seus campos de pesquisa.

A tabela a seguir apresenta os 26 artigos publicados na terceira edição da revista BIOINFO (ordenados por data de publicação):

#	Categoría	OP	Artigo	Autores
01	Ensino		Potencialidades do uso da bioinformática como ferramenta de ensino	Francisco de Sousa, Antônia Pedro, Bianca Araújo, Tarcisio Coutinho
02	Genômica		Metagenômica e Amplicon: perguntas frequentes e respostas essenciais	Sávio Costa
03	Genômica	x	O problema da nomeação dos genes	Izabela Mamede
04	Virologia	x	O Código COVID: A investigação da pandemia da COVID-19 através da bioinformática	Aline da Silva
05	Genômica	x	Desafios na padronização da anotação genômica	Diego Rodrigues
06	Virologia	x	Desafiando a resistência antimicrobiana: o potencial terapêutico da fagoterapia	Bruna Santos
07	Genômica	x	A Bioinformática como aliada da Biotecnologia Agrícola	Ariany Gonçalves
08	Virologia	x	A bioinformática na era pré e pós-pandemia	Bibiana Fam
09	Sequenciamento	x	Uso do BLASTn na construção de primers	Bárbara Pinheiro
10	Bancos de dados		Bioinformática na luta contra o câncer: o bancos de dados na pesquisa oncológica	Thayanne Costa, Lara Bezerra
11	Bioinformática estrutural		Introdução à Biologia Estrutural de Proteínas	Rafael Pereira Lemos, Paulo Santos, Aline Rocha
12	Bioinformática estrutural Tutoriais		Extração de Informações de Sequências e Estruturas de Proteínas	Rafael Pereira Lemos, Paulo Santos, Aline Rocha
13	Transcriptômica		Um mundo dentro de nós: explorando a microbiota humana através da metatranscriptômica	Aline da Silva, Monique dos Santos
14	Ensino		Bioinformática como uma ferramenta didática para o ensino da Genética	Marcos de Sousa, Jeniffer Pereira, Ana Soares, Ana dos Santos, Ricardo da Silva, Arthur de Medeiros, Bruna de Araujo, Francisca Nobrega
15	Evolução	x	Lá e de volta outra vez: um pouco do passado, presente e futuro da evolução	Tiago Borelli
16	Bioinformática estrutural		De onde vêm as proteínas?	Alisson da Silva, Bruno Nunes, Joicymara Xavier
17	Virologia		Vírus endógenos humanos: como analisá-los in silico?	Juan Sampaio, João Neto, Isaac Cansanção
18	Evolução	x	A Bioinformática e a Compreensão da Vida	Thiago de Camargo
19	Biologia de sistemas	x	Visão Integrativa da Biologia de sistemas	Vitor Coelho

#	Categoria	OP	Artigo	Autores
20	Bioinformática estrutural	x	A importância da docagem molecular no combate às bactérias multirresistentes	Aline Cremonesi
21	Bioinformática estrutural Tutoriais		Re-docking Molecular Utilizando o PyMOL e AutoDock VINA	Luana Bastos, Giovana Fiorini
22	Bioinformática estrutural Tutoriais		ColabFold: uma ferramenta web para modelagem de proteínas	Giovana Fiorini, Luana Bastos, Rafael Pereira Lemos
23	Bioinformática estrutural Tutoriais		AlphaFold 2: revolucionando a modelagem de estruturas 3D de macromoléculas	Vivian Moraes Paixão, Angie Puelles, Eduardo Moreira, Luana Bastos, Raquel Melo-Minardi
24	Bioinformática estrutural		Termodinâmica de Proteínas: como as proteínas se enovelam?	Alisson da Silva, Bruno Nunes, Joicymara Xavier
25	Bioquímica		Avaliação ADMET de substâncias	Artur Barros
26	Computação		Perspectiva histórica de grandes eventos da ciência da computação e na biologia molecular: segunda guerra e guerra fria	Monique dos Santos, Aline da Silva

### Comitê editorial

Nesta edição, o comitê editorial atingiu o maior número de participantes entre todas as edições. O comitê editorial é responsável pelo processo de revisão de artigos e pela decisão de publicação. Em média, nesta edição, cada artigo foi avaliado por dois ou mais revisores independentes.

Membros atuais do comitê editorial:

-  Alessandra Lima
-  Aline de Paula Dias da Silva
-  Aline Sampaio Cremonesi
-  Ana Carolina Silva Bulla
-  Ana Paula Abreu
-  Ariany Rosa Gonçalves
-  Bárbara Rebeca de Macedo Pinheiro
-  Bibiana Fam
-  Bruna Espiño dos Santos
-  Carlos Capelini

-  Diego Lucas Neres Rodrigues
-  Diego Mariano
-  Filipe Augusto Teixeira
-  Isaac Farias Cansanção
-  Izabela Mamede
-  Luana Luiza Bastos
-  Lucianna Helene Santos
-  Marcos Antonio Nobrega de Sousa
-  Mira Raya Paula de Lima
-  Rafael Pereira Lemos
-  Savio Costa
-  Thiago Camargo
-  Tiago Cabral Borelli
-  Wylerson Nogueira

Cabe ressaltar que o processo de decisão feito pelo editor atendeu ao parecer dos revisores. A taxa de aprovação para publicação da terceira edição foi de 81,25%. Um total de 18,75% dos artigos submetidos não atenderam a critérios técnicos e não puderam ser aprovados para a publicação. Aos autores que não tiveram seus artigos aceitos, recomendamos que confirmam a [lista de artigos publicados em edições anteriores da revista](#). Confira também os artigos que descrevem nossa [política editorial](#), [política de ética e de conduta](#) e [licença de uso](#). Além disso, recomendamos como material externo de consulta o livro "[Bioinformática: da Biologia à Flexibilidade Molecular](#)", de Verli e colaboradores (2014). Para autores com artigos aceitos, leiam o artigo "[Meu artigo foi aceito. E agora?](#)".

### **Comitê de mídias sociais**

Por fim, é importante destacar o papel do comitê de mídias sociais para o impacto da revista. O comitê de mídias sociais é responsável pela divulgação da revista em redes sociais, como o Instagram, Linkedin e Facebook. A função deste comitê é atrair leitores e potenciais autores para a revista. Nesta edição, o comitê

atuou ativamente na divulgação de trabalhos anteriores e na chamada para novos autores. Alguns membros do comitê de mídias sociais participaram também do processo de revisão.

Membros atuais:

-  Angie Atoche Puelles
-  Emilia Sousa de Oliveira
-  Glenerson Baptista
-  Joalisson da Costa Moreira
-  Luana Gabrielli Ayres Schekiera
-  Vivian Moraes Paixao

### Considerações finais

Até a segunda edição, o modelo de publicação da Revista BIOINFO era baseado no registro como e-book, sendo cada artigo publicado como um capítulo de livro. No fim do ano de 2022, a Revista BIOINFO recebeu do CBISSN (Centro Brasileiro do ISSN) autorização para atuar como periódico acadêmico de divulgação científica com tiragem anual, recebendo o código ISSN 2764-8273. Portanto, a partir da terceira edição, as publicações serão indexadas como artigos. Por fim, estas serão compiladas em um e-book, organizado pelos membros do comitê editorial.

Em nome do comitê editorial, agradeço a todos os leitores e autores pela confiança no trabalho deste comitê e desejo a todos uma boa leitura.

Autores 0.1

Diego Mariano 

Revisão: Alessandra Lima 

Saiba mais 0.1

O editorial está disponível em <https://bioinfo.com.br/editorial-bioinfo-03/>

Cite este artigo 0.2

Mariano, DCB. **Editorial - BIOINFO #03.** BIOINFO. ISSN: 2764-8273. Vol. 3. p.00 (2023). doi:  
10.51780/bioinfo-03-00

# 1

## POTENCIALIDADES DO USO DA BIOINFORMÁTICA COMO FERRAMENTA DE ENSINO

### Autores 1.1

Francisco Bruno de Sousa , Antônia Natália Pedro , Bianca Nascimento Araújo , Tarcisio José Domingos Coutinho 

Revisão: Angie Atoche Puelles 

### Cite este artigo 1.1

Sousa, FB *et al.* Potencialidades do uso da bioinformática como ferramenta de ensino.  
BIOINFO. ISSN: 2764-8273. Vol. 3. p.01 (2023). doi: 10.51780/bioinfo-03-01

### Resumo 1.1

Ao longo dos últimos anos, a bioinformática vem ganhando cada vez mais espaço em pesquisas relacionadas a diversas áreas das ciências biológicas e exatas. No entanto, ainda é pouco abordada em salas de aula. O distanciamento entre o ambiente de pesquisa e a docência deve-se ao fato de a Bioinformática ser relativamente recente e tal motivo pode-se ser explicado por não ter sido abordada durante o processo de formação da maioria dos professores. Porém, mesmo com esta questão relacionada à formação docente, a bioinformática não pode mais estar ausente das aulas, quer seja na educação básica ou no ensino superior. Por isso, é importante que os docentes sejam estimulados a introduzir em suas práticas pedagógicas elementos de bioinformática. Introduzir temas de bioinformática ou até mesmo utilizá-la como uma ferramenta para ensinar conteúdos moleculares pode ser uma importante e eficiente estratégia pedagógica, de forma a contribuir significativamente ao processo de ensino e aprendizagem.

## 1.1 Introdução

**A** bioinformática é, na sua essência, uma área interdisciplinar, pois conhecimentos sobre ciência da computação, biologia, química, física, estatística, dentre outras áreas das ciências biológicas e exatas, são requisitados para que análises, especialmente moleculares, sejam realizadas [1].

Quando analisamos o ambiente acadêmico/escolar percebemos que a bioinformática está presente em diversas linhas de pesquisa de programas de pós graduação assim como nas pesquisas desenvolvidas pelos mais diferentes laboratórios tanto no Brasil quanto em outros países, no entanto quando olhamos para a graduação e a educação básica percebemos que esta área está muito distante das salas de aula destas modalidades de ensino [2-4].

Vários aspectos podem ser considerados para tal distanciamento, desde questões relativas à formação docente, passando pela baixa publicação de artigos que tratem de experiências na área, até mesmo a ausência de ambientes minimamente estruturados para que as aulas sejam ministradas. Vale ressaltar que a ideia abordada aqui não é o ensino de bioinformática em si e de como deverá ser aplicada na sua mais diversas análises; mas sim em como abordar temas relacionados a ela dentro das mais diversas áreas da biologia molecular, com a finalidade de melhorar o processo de ensino e aprendizagem [5-6].

Compreender como as moléculas biológicas funcionam ou até mesmo como evoluem não é uma das atividades mais simples da vida de um discente, seja ele do ensino médio ou da graduação. Esta dificuldade está relacionada, em vários casos, com a capacidade de abstrair, ou seja, de imaginar a atuação das moléculas nas diversas vias metabólicas, bem como com as metodologias utilizadas para a explicação do conteúdo, que normalmente são pouco dinâmicas e priorizam apenas a memorização de informações sem estimular o senso crítico dos discentes [7].

Nesse contexto, explorar temas relacionados com a bioinformática, assim como dados biológicos moleculares reais em sala de aula, permite ao docente uma nova forma de ministrar os conteúdos, bem como ao discente uma maneira diferente de visualizar o tema abordado, ou seja, a bioinformática, antes restrita aos ambientes de pesquisa, agora pode ser inserida na sala de aula atuando como uma importante ferramenta pedagógica [8].

## 1.2 Como inserir?

Inserir a bioinformática na sala de aula pode ser um processo demorado e, para isso requer tempo, determinação e planejamento pela parte do docente, uma vez que o objetivo não é simplesmente substituir uma metodologia de ensino por outra, mas sim, criar um novo ambiente que facilite tanto a ministração quanto a aprendizagem do conteúdo que será abordado. Sendo assim, é importante que o docente que se propõe a inserir a bioinformática em sua sala de aula comece da

forma mais simples possível e à medida que for ganhando confiança explore cada vez mais as possibilidades [9].

O que será abordado a seguir são sugestões simples e objetivas de como docentes, independente da experiência nos temas e do nível de ensino, utilizem a bioinformática em suas salas de aula.

### **1.3 A bioinformática não é para toda aula**

O ideal quando o docente planeje seu plano de aula, se faz necessário levar em consideração a abordagem os diversos temas de sua componente curricular, diversificando as metodologias empregadas. Esta não é uma ação fácil de ser executada, pois muitas vezes os docentes utilizam uma única metodologia, ora por questão de ser mais confortável ou menos desafiador, ora por questão de disponibilidade de tempo, já que muitas vezes o mesmo está envolvido em várias atividades além de suas aulas [10].

Por mais interessante que a metodologia seja e por mais que os discentes sejam envolvidos por ela, a repetição demasiada a tornará um elemento comum que com o passar do tempo, provavelmente, perderá o impacto alcançado anteriormente. Portanto, é importante que durante o planejamento haja uma organização no sentido de qual metodologia seja mais adequada para a ministração de determinado conteúdo [11].

Desse modo, antes de colocar qualquer uma das sugestões a seguir em prática, é importante que cada docente conhecendo bem sua componente curricular escolha o(s) melhor(es) momento(s) onde o mesmo possa explorar o conteúdo utilizando elementos de bioinformática em sua aula.

### **1.4 Explore textos científicos e/ou jornalísticos**

Uma das formas mais simples de inserir a bioinformática nas aulas é utilizando trechos ou textos completos de artigos científicos, ou jornalísticos, pois permitem

ao docente mostrar tanto o impacto do ponto de vista científico quanto a atualidade do tema que está sendo tratado [12].

Um aspecto importante que deve ser considerado nesta abordagem é esforçar-se na contextualização dos recortes aos discentes sobre a importância da bioinformática e das análises computacionais realizadas através da abordagem aplicada pela parte dos pesquisadores e/ou jornalistas que escreveram o texto utilizado na aula. Sem a contextualização, muitos discentes não entenderão sobre os avanços aplicados na compreensão dos mecanismos moleculares, que só foram possíveis devido à análise de sequências biológicas através de ferramentas computacionais.

A forma como o texto será explorado depende muito do docente e do tema da aula, mas imaginando, por exemplo, que a aula seja sobre a genes e tamanho de genomas, o docente pode iniciar a aula com a seguinte pergunta: como sabemos quantos genes cada organismo possui? A partir daí os discentes formularão suas hipóteses e logo após isso uma notícia ou um artigo sobre projeto genoma humano, ou de outros organismos podem ser compartilhado para a leitura, em seguida o docente estabelece um debate sobre as hipóteses levantadas e o texto compartilhado, realçando a importância das ferramentas computacionais utilizadas, bem como dos profissionais bioinformáticos responsáveis pelas análises.

Outro exemplo seria em uma aula sobre heterocromatina e eucromatina relacionado com a diferenciação celular, o docente solicita previamente que os discentes tragam imagens impressas de células humanas (quantidade a critério do docente) e informações básicas como órgão ao qual elas pertencem e suas respectivas funções. Ao chegar na sala com esse material os discentes serão convidados a construírem um quadro comparativo (modelo a critério do docente) mostrando as principais semelhanças e diferenças entre os tipos celulares e sugerir, do ponto de vista molecular, uma explicação para as igualdades e diferenças encontradas. Uma vez o quadro pronto e as hipóteses levantadas, o docente compartilha um artigo ou texto jornalístico que trate da expressão diferencial de genes e coordena uma discussão sobre como o quadro, as hipóteses e o artigo/texto

interagem e ainda aborda sobre RNAs e as análises computacionais relacionadas com essas moléculas.

## 1.5 Utilizando banco de dados

As sequências, estruturas e até mesmo informações relacionadas com as funções das moléculas e suas participações nos mecanismos moleculares dos mais diversos organismos normalmente são organizadas em ambientes virtuais de acesso livre e gratuito que conhecemos como banco de dados. Com certeza qualquer pessoa que realize uma busca simples, minimamente orientada por um tema, na internet por esses ambientes virtuais ficará surpreso com a quantidade e diversidade dos mesmos [13].

Alguns dos bancos de dados mais utilizados e conhecidos por quem usa a bioinformática ou algo relacionado a ela, estão contidos no site do Centro Nacional para Informação Biotecnológica (do inglês, *National Center for Biotechnology Information – NCBI*) (<https://www.ncbi.nlm.nih.gov/>) que é organizado e financiado pelo governo dos Estados Unidos da América. No NCBI encontramos bancos de dados que permitem ao usuário (que é quem está buscando a informação) encontrar artigos científicos como o PubMed e o PubMed Central, genomas de organismos como o *Genome* e sequências de ácidos nucleicos como o *Nucleotide* e proteínas como o *Protein* [14].

Além dos bancos de dados contidos no NCBI há diversos outros que podem ser utilizados em aulas que tratem de temas relacionados com a biologia molecular, como, por exemplo: o *ViralZone* (<https://viralzone.expasy.org/>) que consiste em um ambiente que contém informações diversas sobre estrutura, funcionamento e organização do material genético, dentre outros aspectos, de centenas de vírus, permitindo que o docente utilize amplamente os recursos disponíveis para explorar diversas questões moleculares relacionadas a estes organismos [15].

Outro banco de dados interessante e que pode ser utilizado é o *VFDB – virulence factor database* (<http://www.mgc.ac.cn/Vfs/main.htm>), que consiste em uma ferramenta que

permite que várias análises comparativas em relação à diversidade e quantidade de fatores de virulência em bactérias. O docente pode comparar diversos fatores de virulência de forma inter e intra gêneros bacterianos de forma a permitir aos discentes um aprofundamento sobre os mecanismos adaptativos moleculares relacionados com a patogenicidade [16].

Mais um banco de dados interessante é o *PDB – Protein Data Bank* (<https://www.rcsb.org/>) que é um banco de dados de estruturas tridimensionais de proteínas que é excelente para mostrar aos discentes aspectos relacionados com a forma e função de diversas proteínas. Outro aspecto que chama a atenção nesse banco é na existência de uma seção voltada exclusivamente para o ensino, onde o docente pode encontrar recursos educacionais e materiais didáticos prontos para o uso em sala [17-18].

Sobre esses e outros bancos de dados que podem ser utilizados é importante ressaltar que a maioria deles não foi planejado e/ou projetado para ser uma ferramenta pedagógica, uma vez que os mesmos foram pensados para serem ambientes de auxílio para pesquisas científicas e, portanto, não exploram os aspectos pedagógicos. Então, por que usá-los? Há várias respostas para isso, mas seremos breves.

Os bancos de dados na sua maioria apresentam curadoria, ou seja, as informações que estão contidas nos mesmos são validadas por especialistas, o que permite que o docente trabalhe com dados seguros [19]. Além disso, possuem interface gráfica amigável (visualmente são atrativos e partem do princípio da simplicidade para manuseio dos dados) o que é uma excelente oportunidade para o docente utilizar em suas aulas, pois permitem que os discentes acessem informações de uma forma mais confortável e amistosa, o que pode estimular a curiosidade dos mesmos [20-21]. Por fim, o docente pode utilizá-los não apenas para demonstrar um aspecto de sua aula, mas também pode utilizar esses ambientes para que seus discentes realizem pesquisas científicas através, por exemplo, do que chamamos de aprendizagem baseada em problemas, ou seja, usando dados reais para estimular seus discentes a resolverem situações reais e assim desenvolverem ainda mais o raciocínio e o senso crítico [22-23].

## 1.6 Alinhamento de sequências

A análise mais simples, do ponto de vista técnico, que é realizada por um bioinformata é o alinhamento entre duas ou mais sequências de moléculas biológicas, especialmente DNA e proteínas [24].

De forma simplificada, um alinhamento entre sequências consiste em colocar uma embaixo da outra de forma a comparar as mesmas, posição por posição de cada nucleotídeo ou aminoácido em regiões específicas, ou ao longo de toda a molécula. Essa análise de similaridade (o quanto que elas são semelhantes e entenda semelhança pelo número de caracteres iguais nas mesmas posições nas sequências) é um dos passos mais importantes para várias análises em bioinformática [25] (Figura 1.1).

Seq 1:	A	T	G	C	A	A	T	G	G	T	C	C	T	A	G
Seq 2:	A	T	G	C	A	A	T	G	G	T	C	C	T	A	A
Seq 3:	A	T	G	C	A	A	T	G	G	T	C	C	T	A	A
Seq 4:	A	T	G	C	A	A	T	G	G	T	C	C	C	A	A

Figura 1.1: Exemplo de alinhamento entre 4 sequências de DNA (presença de timina – T) mostrando algumas regiões conservadas (os nucleotídeos não mudam de uma sequência para a outra) e variáveis (nucleotídeos mudam de uma sequência para a outra). Fonte: próprio autor.

## 1.7 Desafios para a inserção

Inserir a bioinformática na sala de aula com certeza é uma tarefa desafiadora por vários motivos para os docentes, inclusive alguns já foram tratados ao longo do texto, no entanto, é uma temática que não pode mais ficar de fora das salas de aula tanto graduação quanto do ensino médio, pois além de importante em relação aos avanços científicos que temos presenciado ao longo dos últimos anos, pode ser

uma forma eficiente de explicar conteúdos e atrair cada vez mais a atenção dos discentes para as áreas das ciências biológicas e exatas [31-32].

Um dos principais desafios é com relação à formação dos docentes que desejam inserir a bioinformática em suas aulas, pois no Brasil essa área é relativamente nova e não faz parte da maioria dos cursos de graduação, especialmente licenciaturas, das ciências biológicas e exatas, o que faz com que a maioria dos docentes que estão espalhados pelo país não tenham conhecimento necessário para inserir essa temática em suas práticas pedagógicas [6, 33].

Uma alternativa para esta questão da formação docente está em produzir cursos tanto presenciais quanto online, que tratem a bioinformática como ela está sendo proposta aqui, ou seja, como um meio que facilite o aprendizado, dessa forma os docentes sentiriam mais confiança em inserir esta área em suas salas de aula e os discentes além de melhorarem a compreensão em alguns temas teriam conhecimento sobre uma área em expansão, o que seria importante para a divulgação da bioinformática no país [34-36].

Ainda atrelado à formação docente, há poucos artigos científicos mostrando experiências relacionadas com o uso da bioinformática em sala de aula no país, da mesma forma como é escassa a produção de material didático com esta finalidade, o que difere muito quando buscamos por artigos produzidos por pesquisadores de outros países. É importante que pesquisadores brasileiros que conhecem a realidade do país produzam conteúdo que colabore com colegas docentes que atuam em nessas condições, ou seja, precisamos de exemplos e alternativas viáveis de acordo com nossa situação e não apenas nos espelharmos em experiências que foram executadas em condições totalmente diferentes [37-39].

Por fim, há um desafio do ponto de vista estrutural, especialmente quando tratamos de escolas públicas, que é a dificuldade de existir um ambiente com computadores e internet para a execução de algumas atividades propostas aqui. A dificuldade estrutural em nosso país não está limitada a uma sala de computadores, sabemos que muitas escolas não possuem laboratório para aulas práticas, às vezes até a própria condição da sala de aula não é a mais adequada [40-42].

Por isso, o uso de textos, como citado anteriormente, pode ser uma alternativa assim como o estabelecimento de parcerias com instituições de ensino superior que disponibilizem seus laboratórios de informática, da mesma forma que caso o docente tenha à sua disposição um computador e um projetor, pode baixando as sequências e softwares específicos como o MEGA, realizar algumas atividades.

Por fim, outro fator que pode complicar as atividades é que todos os bancos de dados e softwares que foram tratados aqui, na maioria dos casos, não foram projetados com a finalidade pedagógica, como já citado, nos quais se encontram em língua inglesa. É preciso estimular também a criação de bancos de dados e softwares brasileiros e em português com estas finalidades, uma vez que as dificuldades já são enormes para o professor brasileiro conseguir executar minimamente seu trabalho, no entanto, enquanto isto não for uma realidade, sugerimos ações interdisciplinares, especialmente nas escolas de ensino médio, entre os docentes de linguagens, biológicas e exatas.

Apesar das dificuldades levantadas e sabendo que os docentes já possuem tantas outras no seu dia a dia e sem lhes impor mais uma responsabilidade ou tratá-los como super-heróis capazes de superar todas as dificuldades, acreditamos que inserir a bioinformática na prática pedagógica pode render bons resultados para todos os envolvidos. Portanto, caso tenha interesse e disposição, convidamos a todos a inserir a bioinformática em sua sala de aula.

Saiba mais 1.1

Este artigo está disponível em <https://bioinfo.com.br/potencialidades-do-uso-da-bioinformatica-como-ferramenta-de-ensino/>

## 1.8 Referências

[1] BAIN, S. A. et al. Bringing bioinformatics to schools with the 4273pi project. *PLoS Computational Biology*, v. 18, n. 1, p. 1-12, 2022.

[2] MACHLUF, Y. et al. Making authentic science accessible—the benefits and challenges of integrating bioinformatics into a high-school science curriculum. *Briefings in Bioinformatics*, v. 18, n. 1, p. 145-159, 2017.

- [3] DALL'ALBA, G. et al. Estudo bibliométrico sobre bioinformática: um levantamento na biblioteca brasileira de teses e dissertações. *Revista NBC*, Belo Horizonte, v. 9, n. 18, p. 17-27, 2019.
- [4] DE LAS RIVAS, J.; BONAVIDES-MARTÍNEZ, C.; CAMPOS-LABORIE, F. J. Bioinformatics in Latin America and SoIBio impact, a tale of spin-off and expansion around genomes and protein structures. *Briefings in bioinformatics*, v. 20, n. 2, p. 390-397, 2019.
- [5] MACHLUF, Y.; YARDEN, A. Integrating bioinformatics into senior high school: design principles and implications. *Briefings in Bioinformatics*, v. 14, n. 5, p. 648-660, 2013.
- [6] MORAES, I. O.; CEZAR-DE-MELLO, P. F. T. O que pensam os docentes sobre o uso da bioinformática no ensino de biologia. *Revista Brasileira de Ensino de Ciência e Tecnologia*, Ponta Grossa. v. 14, n. 2, p. 75-94, 2021.
- [7] JENKINSON, J. Molecular Biology Meets the Learning Sciences: Visualizations in Education and Outreach. *Journal of Molecular Biology*, v. 430, n. 21, p. 4013-4027, 2018.
- [8] MARTINS, A. et al. Bioinformatics-Based Activities in High School: Fostering Students' Literacy, Interest, and Attitudes on Gene Regulation, Genomics, and Evolution. *Frontiers Microbiology*, v. 11, n. 1, p.1-15, 2020.
- [9] FORM, D.; LEWITTER, F. Ten Simple Rules for Teaching Bioinformatics at the High School Level. *PLoS Computational Biology*. v. 7, n. 10: p. 1-2, 2011.
- [10] NICOLA, J. A.; PANIZ, C. M. A importância da utilização de diferentes recursos didáticos no Ensino de Ciências e Biologia. *InFor*, v. 2, n. 1, p. 355-381, 2017.
- [11] SANTOS, A. N. B.; LIMA, F. G. C. Ensino de ciências e biologia: avanços e perspectivas a partir de reflexões e contextos da atualidade. *Revista Ibero-Americana de Humanidades, Ciências e Educação*, São Paulo, v. 7, n. 2, p. 370-384, 2021.
- [12] MONERAT, C. A.; ROCHA, M. B. Biologia celular em revista: análise de textos de divulgação científica. *Ensino, Saúde e Ambiente*, v. 10, n. 3, p. 16-33, 2017.
- [13] DUCK, G.; NENADIC, G.; FILANNINO, M. A Survey of Bioinformatics Database and Software Usage through Mining the Literature. *PLoS ONE*, v. 11, n. 6: p. 1-25, 2016.
- [14] NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, v. 44, n. 1, p. 7-19, 2016.
- [15] HULO, C. et al. ViralZone: a knowledge resource to understand virus diversity. *Nucleic Acids Research*, v. 39, n. 1, p. 576-582, 2011.
- [16] LIU, B. et al. VFDB 2022: a general classification scheme for bacterial virulence factors. *Nucleic Acids Research*, 2022; v. 50, n. 1: p. 912-917, 2022.

- [17] BERMAN, H. M. et al. The Protein Data Bank. *Nucleic Acids Research*, v. 28, n. 1, p. 235-242, 2000.
- [18] ZARDECKI, C. et al. PDB-101: Educational resources supporting molecular explorations through biology and medicine. *Protein science: a publication of the Protein Society*, v. 31, n. 1, p. 129-140, 2022.
- [19] SIB Swiss Institute of Bioinformatics Members. The SIB Swiss Institute of Bioinformatics' resources: focus on curated databases. *Nucleic Acids Research*, v. 44, n. 1, p. 27-37, 2016.
- [20] HERNANDEZ-DE-DIEGO, R. et al. The eBioKit, a stand-alone educational platform for bioinformatics. *PLoS Computational Biology*, v. 13, n. 9, p. 1-14, 2017.
- [21] RAMOS, P. I. P. et al. Leveraging User-Friendly Network Approaches to Extract Knowledge From High-Throughput Omics Datasets. *Frontiers in Genetics*, v. 10, n. 1, p. 1-19, 2019.
- [22] PROCHAZKOVA, K. et al. Teaching a difficult topic using a problem-based concept resembling a computer game: development and evaluation of an e-learning application for medical molecular genetics. *BMC Medical Education*, v. 19, n. 1, p. 1-8, 2019.
- [23] SEIDLEIN, A. H., et al. Gamified E-learning in medical terminology: the terminator tool. *BMC Medical Education*, v. 20, n. 1, p. 1-10, 2020.
- [24] ZIELEZINSKI, A. et al. Alignment-free sequence comparison: benefits, applications, and tools. *Genome Biology*, v. 18, n. 1, p. 1-17, 2017.
- [25] WANG, Y.; WU, H.; CAI, Y. A benchmark study of sequence alignment methods for protein clustering. *BMC Bioinformatics*, v. 19, n. 1, p. 95-104, 2018.
- [26] KLEINSCHMIT, A. et al. Sequence Similarity: An inquiry based and “under the hood” approach for incorporating molecular sequence alignment in introductory undergraduate biology courses. CourseSource, 2019.
- [27] TAMURA, K.; STECHER, G.; KUMAR, S. MEGA11: Molecular Evolutionary Genetics Analysis Version 11. *Molecular Biology and Evolution*, v. 38, n. 7, p. 3022–3027, 2021.
- [28] SIEVERS, F. et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, v. 7, n. 1, p. 1-6, 2011.
- [29] EDGAR, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, v. 32, n. 5: p. 1792-1797, 2004.
- [30] NOTREDAME, C.; HIGGINS, D. G.; HERINGA, J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology*, v. 302, n. 1, p. 205-217, 2000.

- [31] SAYRES, M. A. W. et al. Bioinformatics core competencies for undergraduate life sciences education. *PLoS ONE*, v. 13, n. 6, p. 1-20, 2018.
- [32] WHITLEY, K. V.; TUELLER, J. A.; WEBER, K. S. Genomics Education in the Era of Personal Genomics: Academic, Professional, and Public Considerations. *International Journal of Molecular Sciences*, v. 21, n. 3, p. 1-19, 2020.
- [33] LOUISA, W.; GEBHARDT, P. Bioinformatics goes to school-new avenues for teaching contemporary biology. *PLoS Computational Biology*, v. 9, n. 6: p. 1-6, 2013.
- [34] MARQUES, I. et al. Bioinformatics projects supporting life-sciences learning in high schools. *PLoS Computational Biology*, v. 10, n. 1, p. 1-6, 2014.
- [35] ZUVANOV, L. et al. The experience of teaching introductory programming skills to bioscientists in Brazil. *PLoS Computational Biology*, v. 17, n. 11, p. 1-16, 2021.
- [36] SILVA, A. L. et al. From In-Person to the Online World: Insights Into Organizing Events in Bioinformatics. *Frontiers in Bioinformatics*, v. 1, n. 1, p. 1-10, 2021.
- [37] RIBEIRO JUNIOR, H. L.; OLIVEIRA, R. T. G.; CECCATTO, V. M. Bioinformática como recurso pedagógico para o curso de ciências biológicas na Universidade Estadual do Ceará – UECE – Fortaleza, Estado do Ceará. *Acta Scientiarum. Education*, v. 34, n. 1, p. 129-140, 30 abr. 2012.
- [38] FREIRE, C. M. A. S. et al. Proposta pedagógica em prática no ensino de bioquímica: aproveitamento de softwares livres como facilitador do processo de ensino e de aprendizagem. *Revista Thema*, v. 15, n. 4, p. 1442–1455, 2018.
- [40] NOBRE, R. H.; SOUSA, J. A.; NOBRE, C. S. P. Uso dos laboratórios de informática em escolas do ensino médio e fundamental no interior nordestino. *Revista Brasileira de Informática na Educação*, v. 23, n. 3, p. 68-80, 2015.
- [41] VASCONCELOS, J. C. et al. Infraestrutura escolar e investimentos públicos em Educação no Brasil: a importância para o desempenho educacional. *Ensaio: Avaliação e Políticas Públicas em Educação*, v. 29, n. 113, p. 874-898, 2021.
- [42] SABIA, C. P. P.; SORDI, M. R. L. Um olhar para a dimensão infraestrutura como uma das condições objetivas possibilidadoras da qualidade em escolas públicas. *Revista Ibero-Americana de Estudos em Educação*, Araraquara, v. 16, n. 1, p. 127-152, 2021.

# 2

## METAGENÔMICA E AMPLICON: PERGUNTAS FREQUENTES E RESPOSTAS ESSENCIAIS

### Autores 2.1

Sávio de Souza Costa 

Revisão: Diego Mariano 

### Cite este artigo 2.1

Costa, SS. Metagenômica e Amplicon: perguntas frequentes e respostas essenciais.

BIOINFO. ISSN: 2764-8273. Vol. 3. p.02 (2023). doi: 10.51780/bioinfo-03-02

## Resumo 2.1

Neste artigo, você irá aprender sobre Metagenômica e Amplicon.

### 2.1 O que é Metagenômica?

OBTER informações genéticas sobre comunidades microbianas presentes nos mais variados ambientes com as técnicas clássicas de biologia molecular possui um grande obstáculo, que é como acessar o genoma desses organismos se apenas 1% das bactérias conhecidas podem ser cultivadas utilizando metodologias atuais? [1]. Dessa forma, uma alternativa foi a clonagem de genes específicos ao invés da clonagem do genoma completo [2].

Na década de 1980, trabalhos moleculares utilizaram a técnica de PCR para explorar a diversidade de sequências de RNA ribossômico, levando a ideia de clonar DNA diretamente de amostras ambientais já em 1985. [3]. Após este ponto de partida, foi publicado o primeiro trabalho a utilizar o termo metagenômica. Este trabalho analisou as bases para clonagem para a análise funcional de microrganismos do solo [2] e, a partir desta técnica, também surgiu a chamada análise por Amplicon. Atualmente a metagenômica e a Amplicon são técnicas que garantem acesso a diversidade e caracterização microbiana nos mais variados ambientes sem a necessidade de cultura dos microrganismos [4].

*"Dessa forma, a **metagenômica** é uma valiosa ferramenta para a descoberta de novos genes, vias metabólicas e enzimas que são de extrema importância biotecnológica e para saúde [5]."*

Antes de se encontrar as respostas pra pergunta fundamental sobre metagenômica, ecologia microbiana e tudo mais. Primeiro deve-se saber a pergunta, e as principais perguntas que são respondidas em um estudo de metagenômica são “Quem são os microrganismos presentes em um determinado ambiente”, “O que eles estão fazendo” e “Como estão fazendo isto?”.

## 2.2 Metagenômica e amplicon? São as mesmas coisas?

**A resposta rápida é: não!** Contudo precisamos entender que no início ambos eram sinônimos para análises de genes de microrganismos amplificados/clonados direto de um determinado ambiente. Atualmente essas análises metagenômicas são diferenciadas e chamadas de metagenômica e amplicon.

*"A terminologia **amplicon** vem da amplificação de um determinado gene marcador presente no ambiente que será analisado."*

Tal gene marcador é geralmente caracterizado como *house-keeping*, ou seja, essencial para certo grupo de microrganismos. Além disso, ele precisa ter outras características, como ser altamente conservado em uma espécie, mas ter certas diferenças em outras espécies. Isso permite que tais genes possam ser usados como uma ferramenta para distinguir diferentes tipos de microrganismos [6].

A utilização do gene **rRNA 16S** como marcador filogenético está consolidada principalmente por apresentar características como: baixa taxa de transferência horizontal, baixa taxa de recombinação gênica, sequência extremamente conservada, mas, que possui regiões hipervariadas que são espécie-específicas [6]. Portanto, a partir da análise dessa região hipervariada, é possível se identificar a nível taxonômico e distinguir diferentes espécies bacterianas [7]. Outros exemplos de genes marcadores de microrganismos incluem o gene **recA** em bactérias, o gene **ITS** em fungos e o gene **SSU rRNA** em eucariotos unicelulares. A importância de utilizar genes bem descritos na literatura permite a realização de comparações entre diferentes comunidades, ressalvadas as diferenças metodológicas empregadas [7][8]. A metodologia padrão desse tipo de análise busca formar clusters com as sequências dos genes marcadores semelhantes, esses clusters são denominados de “Unidade taxonômica operacional” (OTU’s) [6].

Já o termo metagenômica é normalmente utilizado para tratar da metagenômica *shotgun* que é uma técnica que permite analisar o material genético de uma amostra complexa contendo todos os genes de muitas espécies diferentes. Ou seja, além de amplificar e obter genes, como o 16s bacteriano, no método *shotgun*, o DNA da amostra é fragmentado aleatoriamente em pequenos pedaços sendo, em seguida, sequenciado em massa para assim se obter uma grande variedade de genes presentes neste ambiente [9][10]. A análise metagenômica *shotgun* tem a uma ampla variedade de aplicações, incluindo a investigação de ecossistemas microbianos em ambientes naturais, estudos de microbiomas humanos para entender a relação entre a microbiota intestinal e a saúde humana, e até mesmo para descobrir novas enzimas e produtos naturais produzidos por microrganismos [11].

## 2.3 Como se analisa dados de metagenômica

### *Shotgun?*

Esta pergunta é bastante ampla, uma vez que a metagenômica gera uma grande quantidade de dados a serem analisados. Dentro do “tiroteio de *shotgun*”, existem todos os genes presentes na amostra, como genes marcadores como 16S, genes relacionados às funções metabólicas dos microrganismos, dentre outros. Assim, os objetivos fundamentais da metagenômica podem ser a análise do perfil das comunidades bacterianas, do perfil funcional dos genes e até mesmo a montagem metagenômica para tentativa de obter genomas a partir do metagenoma [10][12]. As análises destes dados ocorre através de ferramentas de **Bioinformática** onde após o tratamento das leituras brutas oriundas do sequenciamento metagenômico, as metodologias computacionais podem prosseguir para diversos caminhos dependendo da pergunta fundamental que será respondida.

Com os dados do metagenoma tratados, é possível responder à pergunta “quem são os microrganismos presentes”. Isso pode ser feito comparando as sequências com bancos de dados de sequências conhecidas, como o banco de dados NCBI, SILVA [13] e RDP [14], usando ferramentas como BLAST [15]

ou DIAMOND [16]. Contudo, essa metodologia pode ser considerada bastante trabalhosa. Dessa forma, surgiram softwares que fazem essas análises utilizando seus próprios métodos, como Kraken2 [17], MetaPhlAn2 [18], MEGAN [19] e o MGRAST [20].

A utilização destas ferramentas propicia a obtenção da diversidade e abundância dos microrganismos presentes, sendo que cada uma utiliza sua própria metodologia computacional. Por exemplo, uma maior diversidade bacterianas foi encontrada pela metodologia shotgun do que usando o método 16S rRNA, o que permitiu ainda prever a classificação de táxons de maneira mais eficaz em nível de filo e, em menor grau, em nível de gênero [21].

Outra abordagem é a obtenção de *contigs* por meio de softwares conhecidos como montadores. Os montadores convencionais utilizam diversas abordagens, uma delas é a de grafo *De Bruijn*, o qual divide as leituras em k-mers e reduz a demanda de memória do computador. Alguns exemplos de montadores baseados em grafo *De Bruijn* incluem o MetaVelvet [22], IDBA-UD [23], MEGAHIT [24] e o metaSPAdes [25]. A escolha do montador irá depender do dado e variar para cada amostra, por isso sempre importante comparar as montagens utilizando o **METAQUAST**.

Após a obtenção de *contigs*, é possível aplicar diversas metodologias para análises dos dados dos metagenomas. Uma dessas técnicas consiste em construir genomas a partir de metagenomas (MAGs – *Metagenome-assembled genomes*). Nesse caso, o primeiro passo consiste no *binning* das leituras, que irá agrupar as *contigs* com base em suas características, como base nas frequências de tetranucleotídeos (TNFs), abundância de genes marcadores e uso de códons [26]. Os softwares mais utilizados com intuito de obter os genomas são o MetaBAT [27], CONCOCT [28] e MaxBin2 [29]. O grau de contaminação de um MAG pode ser analisado através do software CheckM [30]. Assim, a taxa de contaminação depende do método de obtenção do MAG e da diversidade dos microrganismos na amostra do metagenoma. Os MAGs que apresentam alta completude e baixos níveis de contaminação são então selecionados para posterior anotação taxonômica e predição de genes.

As análises apresentadas a seguir podem ser utilizadas tanto para contigs montados quanto para os MAGs preditos. Além disso, eles respondem a duas perguntas funcionais que são: “o que estão fazendo” e “como estão fazendo?”. Dessa forma, para fazer a análise das funções metabólicas dos genes encontrados, pode-se utilizar ferramentas baseadas em homologia, como BLAST [15], para comparar as sequências de genes previstos com as de genes conhecidos. No entanto, métodos modernos, como eggNOG-mapper [31], GhostKOALA [32], MG-RAST [33] e PANNZER2 [34], empregam estratégias de alinhamento otimizadas que permitem alinhamentos rápidos de sequências de genes com bancos de dados. O MG-RAST [33] fornece uma interface de análise metagenômica *online* que inclui *upload* de dados, controle de qualidade e alinhamento com bancos de dados de referência. Ele permite a análise tanto funcional dos genes quanto a predição das comunidades microbianas ali presentes. Dessa forma, percebe-se que a abordagem de quem está presente, o que estão fazendo e como estão fazendo possui vários caminhos e, tudo isso, depende da metodologia escolhida para obter a resposta mais completa. Entretanto, estas são só algumas das ferramentas mais utilizadas para obter estas respostas.

## 2.4 Como se analisa dados de amplicon?

A tecnologia de amplicon é uma abordagem mais comum para a análise de dados ambientais, pois utiliza a amplificação de regiões específicas do DNA para estudar a diversidade e função de comunidades microbianas em diferentes ambientes. Essa abordagem pode ser usada para estudar comunidades microbianas em diversos tipos de amostras, incluindo solo, água, fezes e mucosas. Uma das principais vantagens da tecnologia de amplicon é a sua alta sensibilidade e especificidade, que permite a detecção de baixas abundâncias de microrganismos e a identificação de espécies específicas dentro de uma comunidade complexa. Além disso, a tecnologia de amplicon é relativamente simples e acessível, o que torna essa abordagem uma das mais populares na análise de dados de metagenômica [35] [36].

A técnica de amplicon é baseada em homologia e predição. As espécies podem ser identificadas com base na sequência lida da região variável dos genes

marcadores. O método requer o alinhamento de uma sequência do gene escolhido com todas as sequências de um banco de dados de referência. Dentre os bancos de dados utilizados, encontram-se alguns similares a metagenômica *shotgun*, como SILVA [13] e RDP [14]. Algumas ferramentas e *pipelines* estão disponíveis para a análise dessas sequências de forma automatizada, como QIIME (*Quantitative Insights Into Microbial Ecology*) [37], MOTHUR [38] e USEARCH [39], bem como opções mais recentes DADA2 [40] e Qiime2-Deblur [41].

Os softwares QIIME, MOTHUR e USEARCH-UPARSE agrupam sequências com 97% de identidade em Unidades Taxonômicas Operacionais (OTUs). Já os Qiime2-Deblur, DADA2 e USEARCH-UNOISE3 tentam reconstruir as sequências biológicas exatas presentes na amostra, chamadas de *Amplicon Sequence Variants* (ASVs). Assim, uma OTU representa um grupo de sequências muito próximas (>97% de identidade), que se separa das demais OTUs pela aplicação de técnicas de agrupamento hierárquico utilizando limites de identidade de sequência independentemente de inferências filogenéticas [7].

ASVs são referidos por outros autores como “*zero noise OTUs*” ou “*sub-OTUs*”. Assim, elas buscam identificar e distinguir sequências de amplicons individuais com base em diferenças nucleotídicas. Em vez de agrupar as sequências em OTUs, o método ASV atribui um número de identificação único para cada sequência de amplicon presente nos dados, representando assim variantes únicas. O uso de ASVs permite uma resolução mais alta na análise da diversidade microbiana. Ao considerar cada sequência individualmente, é possível identificar diferenças sutis entre as variantes, como mutações ou polimorfismos, que poderiam ser agrupados em uma única OTU utilizando uma abordagem baseada em similaridade [42][43].

O sequenciamento de metagenoma também é particularmente útil no estudo de comunidades virais. Como os vírus carecem de um marcador filogenético universal compartilhado, a única maneira de acessar a diversidade genética da comunidade viral de uma amostra ambiental é por meio da metagenômica. A metagenômica tem o potencial de avançar o conhecimento em uma ampla variedade de campos. Também pode ser aplicado para resolver desafios práticos em medicina, engenharia, agricultura, sustentabilidade e ecologia [11][21]. Por

fim, cabe ressaltar que a escolha entre as abordagens de metagenômica por amplicon ou *shotgun* depende principalmente dos objetivos da pesquisa, dos recursos disponíveis e do tipo de amostra a ser analisada.

Saiba mais 2.1

Este artigo está disponível em <https://bioinfo.com.br/a-bioinformatica-na-metagenomica-e-amplicon-perguntas-frequentes-e-respostas-essenciais/>

## 2.5 Referências

- [1] Hawksworth, D. L. The magnitude of fungal diversity: the 1.5 million species estimate revisited. *Mycological Research*, 105(12), 1422–1432. (2001). doi:10.1017/s0953756201004725.
- [2] Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem Biol*. Oct;5(10):R245-9. (1998). doi: 10.1016/s1074-5521(98)90108-9. PMID: 9818143.
- [3] Pace NR, Stahl DA, Lane DJ, Olsen GJ. “The Analysis of Natural Microbial Populations by Ribosomal RNA Sequences”. In Marshall KC (ed.). *Advances in Microbial Ecology*. Vol. 9. Springer US. pp. 1–55. (1986). doi:10.1007/978-1-47570611-6.
- [4] Langille, M. G. I., Zaneveld, J., Caporaso, J. G., McDonald, D., Knights, D., Reyes, J. A., Huttenhower, C. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nature Biotechnology*, 31(9), 814–821. (2013). doi:10.1038/nbt.2676.
- [5] Culligan, E. P., Marchesi, J. R., Hill, C., Sleator, R. D. Combined metagenomic and phenomic approaches identify a novel salt tolerance gene from the human gut microbiome. *Frontiers in Microbiology*, 5. (2014). doi:10.3389/fmicb.2014.00189
- [6] Tikhonov, Mikhail, Robert W. Leach, and Ned S. Wingreen. “Interpreting 16S metagenomic data without clustering to achieve sub-OTU resolution.” *The ISME journal* 9.1 (2015): 68-80.
- [7] Yarza, P., Yilmaz, P., Pruesse, E. et al. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nat Rev Microbiol* 12, 635–645 (2014). <https://doi.org/10.1038/nrmicro3330>
- [8] Harris, J. K., Kelley, S. T. Pace, N. R. New perspective on uncultured bacterial phylogenetic division OP11. *Appl. Environ. Microbiol.* 70, 845–849 (2004).

- [9] Wang, T. et al. Structural segregation of gut microbiota between colorectal cancer patients and healthy volunteers. *ISME J.* 6, 320–329 (2012).
- [10] Quince, C., Walker, A., Simpson, J. et al. Shotgun metagenomics, from sampling to analysis. *Nat Biotechnol* 35, 833–844 (2017). <https://doi.org/10.1038/nbt.3935>
- [11] Madhavan, A., Sindhu, R., Parameswaran, B. et al. Metagenome Analysis: a Powerful Tool for Enzyme Bioprospecting. *Appl Biochem Biotechnol* 183, 636–651 (2017). <https://doi.org/10.1007/s12010-017-2568-3>
- [12] Scholz, M. et al. Strain-level microbial epidemiology and population genomics from shotgun metagenomics. *Nat. Methods* 13, 435–438 (2016).
- [13] Quast C, Pruesse E, Yilmaz P, Gerken J, Schneemann T, Yarza P, Peplies J, Glöckner FO. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. [Opens external link in new window](#)*Nucl. Acids Res.* 41 (D1): D590-D596. (2013).
- [14] Cole, J. R. The Ribosomal Database Project (RDP-II): previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy. *Nucleic Acids Research*, 31(1), 442–443. (2003). doi:10.1093/nar/gkg039
- [15] Altschul S, Gish W, Miller W, Myers E, Lipman D: Basic local alignment search tool. *J Mol Biol* 1990, 215(3):403–410.
- [16] Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods*. 2015 Jan;12(1):59-60. doi: 10.1038/nmeth.3176. Epub 2014 Nov 17. PMID: 25402007.
- [17] Wood, D.E., Lu, J. Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol* 20, 257 (2019). <https://doi.org/10.1186/s13059-019-1891-0>
- [18] Truong, D., Franzosa, E., Tickle, T. et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat Methods* 12, 902–903 (2015). <https://doi.org/10.1038/nmeth.3589>
- [19] Huson DH, Auch AF, Qi J, Schuster SC. MEGAN analysis of metagenomic data. *Genome Res.* 2007 Mar;17(3):377-86. doi: 10.1101/gr.5969107. Epub 2007 Jan 25. PMID: 17255551; PMCID: PMC1800929.
- [20] Keegan KP, Glass EM, Meyer F. MG-RAST, a Metagenomics Service for Analysis of Microbial Community Structure and Function. *Methods Mol Biol.* 2016;1399:207-33. doi: 10.1007/978-1-4939-3369-3\_3. PMID : 26791506.
- [21] Ranjan R, Rani A, Metwally A, McGee HS, Perkins DL. Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. *Biochem Biophys Res Commun.* 2016 Jan 22;469(4):967-77. doi: 10.1016/j.bbrc.2015.12.083. Epub 2015 Dec 22. PMID: 26718401; PMCID: PMC4830092.

- [22] Namiki, T., Hachiya, T., Tanaka, H., Sakakibara, Y. (2012). MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic acids research*, 40(20), e155. DOI: 10.1093/nar/gks678
- [23] Peng, Y., Leung, H. C., Yiu, S. M., Chin, F. Y. (2012). IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, 28(11), 1420-1428. DOI: 10.1093/bioinformatics/bts174
- [24] Li, D., Liu, C. M., Luo, R., Sadakane, K., Lam, T. W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*, 31(10), 1674-1676. DOI: 10.1093/bioinformatics/btv033
- [25] Nurk, S., Meleshko, D., Korobeynikov, A., Pevzner, P. A. (2017). metaSPAdes: a new versatile metagenomic assembler. *Genome research*, 27(5), 824-834 DOI: 10.1101/gr.213959.116
- [26] Setubal JC. Metagenome-assembled genomes: concepts, analogies, and challenges. *Biophys Rev*. 2021 Nov 4;13(6):905-909. doi: 10.1007/s12551-021-00865-y. PMID: 35059016; PMCID: PMC8724365.
- [27] Kang, D. D., Froula, J., Egan, R., Wang, Z. (2015). MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ*, 3, e1165. DOI: 10.7717/peerj.1165
- [28] Alneberg, J., Bjarnason, B. S., de Bruijn, I., Schirmer, M., Quick, J., Ijaz, U. Z., ... Quince, C. (2014). Binning metagenomic contigs by coverage and composition. *Nature methods*, 11(11), 1144-1146. DOI: 10.1038/nmeth.3103
- [29] Wu, Y. W., Simmons, B. A., Singer, S. W. (2016). MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics*, 32(4), 605-607. DOI: 10.1093/bioinformatics/btv638
- [30] Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res*. 2015 Jul;25(7):1043-55. doi: 10.1101/gr.186072.114. Epub 2015 May 14. PMID: 25977477; PMCID: PMC4484387.
- [31] Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M. C., ... Bork, P. (2019). eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic acids research*, 47(D1), D309-D314 DOI: 10.1093/nar/gky1085
- [32] Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic acids research*, 44(D1), D457-D462. DOI: 10.1093/nar/gkv1070
- [33] Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E. M., Kubal, M., ... Edwards, R. A. (2008). The metagenomics RAST server—a public resource for the automatic

phylogenetic and functional analysis of metagenomes. *BMC bioinformatics*, 9(1), 386. DOI: 10.1186/1471-2105-9-386

[34] Tonini, M., Ureta-Vidal, A., Bateman, A. (2021). PANNZER2: a rapid functional annotation web server. *Nucleic acids research*, 49(W1), W542-W546. DOI: 10.1093/nar/gkab408

[35] Liu YX, Qin Y, Chen T, Lu M, Qian X, Guo X, Bai Y. A practical guide to amplicon and metagenomic analysis of microbiome data. *Protein Cell.* (2021) May;12(5):315-330. doi: 10.1007/s13238-020-00724-8. Epub 2020 May 11. PMID: 32394199; PMCID: PMC8106563.

[36] Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JI, et al. QIIME allows analysis of high-throughput community sequencing data. *Nat Methods.* (2010) ;7:335–336

[37] Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., ... Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature methods*, 7(5), 335-336. DOI: 10.1038/nmeth.f.303

[38] Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., ... Weber, C. F. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and environmental microbiology*, 75(23), 7537-7541. DOI: 10.1128/AEM.01541-09

[39] Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26(19), 2460-2461. DOI: 10.1093/bioinformatics/btq461

[40] Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J., Holmes, S. P. (2016). DADA2: high-resolution sample inference from Illumina amplicon data. *Nature methods*, 13(7), 581-583. DOI: 10.1038/nmeth.3869

[41] Amir, A., McDonald, D., Navas-Molina, J. A., Kopylova, E., Morton, J. T., Xu, Z. Z. Knight, R. (2017). Deblur rapidly resolves single-nucleotide community sequence patterns. *mSystems*, 2(2), e00191-16. DOI: 10.1128/mSystems.00191-16.

[42] Callahan, B. J., McMurdie, P. J., Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME journal*, 11(12), 2639-2643. DOI: 10.1038/ismej.2017.119

[45] Edgar, R. C. (2018). Updating the 97% identity threshold for 16S ribosomal RNA OTUs. *Bioinformatics*, 34(14), 2371-2375. DOI: 10.1093/bioinformatics/bty113

# 3 O PROBLEMA DA NOMEAÇÃO DOS GENES

## Autores 3.1

Izabela Mamede 

Revisão: Luana Bastos , Bárbara Rebeca de Macedo Pinheiro 

## Cite este artigo 3.1

Conceição, IMCA. **O problema da nomeação dos genes.** BIOINFO. ISSN: 2764-8273. Vol. 3. p.03 (2023). doi: 10.51780/bioinfo-03-03

### Resumo 3.1

#### Opiniões & Perspectivas

O PROJETO **Genoma Humano** começou em 1990 e alcançou seu primeiro grande resultado em 2001, quando foi publicada a primeira montagem completa de todos os cromossomos humanos simultaneamente nas revistas *Nature* e *Science* [1]. Essa primeira montagem nos trouxe informações que ainda são discutidas na comunidade científica, como o fato de que apenas 3% do genoma humano representa sequências associadas a genes e que o ser humano possui cerca de 30 a 40 mil genes, valor muito similar ao de vários parasitas, seres que eram considerados “menos complexos”. Em 2022, com a possibilidade de sequenciarmos fragmentos de tamanhos maiores, alcançamos o objetivo final criado em 1990: um sequenciamento e montagem completa de um genoma humano do início ao fim [2].

Esse novo sequenciamento foi chamado “**de telômero a telômero**”, os telômeros sendo uma região altamente repetitiva no início e no fim de todos os cromossomos. Essa nova montagem do genoma humano de forma impressionante reduziu ainda mais **o número de genes para entre 20 e 25 mil**. De 2001 até hoje, foi descoberto que a grande parte da variabilidade humana estava na produção de RNA e não do DNA, isso ocorre, pois durante processamento do RNA recém transcrito do “molde” do gene, cada gene pode ser processado em vários RNAs maduros diferentes e estes, inclusive podem codificar diferentes proteínas e vários RNAs que não codificam proteínas. Genes que são, a princípio, não codificadores de proteínas, como os RNAs longos não codificadores, possuem ainda mais formas de RNA maduro diferentes se comparados com genes codificadores [3] (Figura 3.1). Essa grande variabilidade no nível do RNA é o que difere os vertebrados dos demais seres vivos e vários RNAs produzidos a partir do mesmo gene podem ser processados e ter funções completamente diferentes da proteína que é associada à tradução daquele locus.

Sendo assim, como nomear cada gene? Normalmente o nome dado a um gene está associado a sua função, por exemplo, o **gene humano TP53** vem do inglês

*Tumor Protein 53*, ou proteína associada a tumores 53. No entanto, esse gene possui 27 RNAs que são transcritos a partir de seu locus, a grande maioria deles produtores de proteínas e que variam de 175 a 2580 pares de base no tamanho do RNA. Por essa grande variabilidade de tamanho é de se esperar que essas proteínas possuem funções distintas umas das outras, então o nome TP53 acaba não representando a função do gene, somente a primeira função que foi associada a ele.

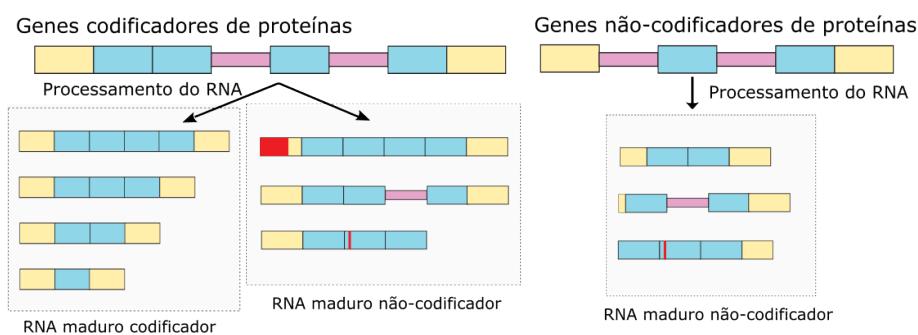


Figura 3.1: Multiplos RNAs maduros são transcritos a partir do mesmo molde gênico. Exons em azul, introns em rosa, regiões perdidas em vermelho. Fonte: Autor

Um artigo recentemente publicado [4] discute exatamente isso, que as nomeações de genes que temos hoje prejudicam análises subsequentes e mesmo a interpretação de características moleculares associadas a certo gene. Por exemplo, se o gene GENE está associado à doença Y, nada impede um clínico que estuda a doença Z e encontra esse gene presente em seus pacientes, pense que ambas as doenças poderiam compartilhar mecanismos parecidos. Ou mesmo que a presença da expressão um gene em certa condição seja erroneamente associada a presença de sua proteína, quando na maioria das vezes não é o caso. Os autores chegaram à conclusão de que a melhor forma seria de que cada gene fosse anotado e depositado em bancos públicos associados a todas as suas possíveis funções e a todos os RNAs associados a este que também têm funções biológicas descritas. Isso permitiria uma melhor descrição do gene com todas as suas funções e auxiliaria bioinformáticos e outros profissionais na interpretação de resultados de sequenciamentos.

### Saiba mais 3.1

Este artigo está disponível em <https://bioinfo.com.br/o-problema-da-nomeacao-dos-genes/>

## 3.1 Referências

- [1] LANDER, E. S. et al. Initial sequencing and analysis of the human genome. *Nature*, v. 409, n. 6822, p. 860–921, 2001.
- [2] NURK, S. et al. The complete sequence of a human genome. *Science (New York, N.Y.)*, v. 376, n. 6588, p. 44–53, 2022.
- [3] CONCEIÇÃO, I. M. C. A. et al. Metformin treatment modulates long non-coding RNA isoforms expression in human cells. *Non-coding RNA*, v. 8, n. 5, p. 68, 2022.
- [4] AMARAL, P. et al. The status of the human gene catalogue. 2023. Disponível em: <<http://arxiv.org/abs/2303.13996>>. Acesso em: 13 jul. 2023.

# 4 O CÓDIGO COVID: A INVESTIGAÇÃO DA PANDEMIA DA COVID-19 ATRAVÉS DA BIOINFORMÁTICA

## Autores 4.1

Aline de Paula Dias da Silva 

Revisão: Isaac Farias Cansanção , Tiago Cabral Borelli 

## Cite este artigo 4.1

Silva, APD. **O Código COVID: A investigação da pandemia da COVID-19 através da bioinformática.** BIOINFO. ISSN: 2764-8273. Vol. 3. p.04 (2023). doi: 10.51780/bioinfo-03-04

## Resumo 4.1

### Opiniões & Perspectivas

No final do ano de 2019, um vírus desconhecido se espalhou repentinamente no mundo, causando uma síndrome respiratória aguda grave [1]. A severidade da doença e a forma rápida de transmissão do vírus ligou o alerta em todo o mundo como uma emergência de saúde pública, como resultado, milhões de pessoas estavam sendo acometidas rapidamente e milhares de mortes ocorrendo diariamente [2]. Dessa forma, as principais questões precisavam ser respondidas, como: o agente etiológico causador da doença, as formas de transmissão, características clínicas, os mecanismos de patogênese e principalmente, os métodos de tratamento e prevenção (Figura 4.1).

## INVESTIGAÇÃO DE UMA NOVA DOENÇA

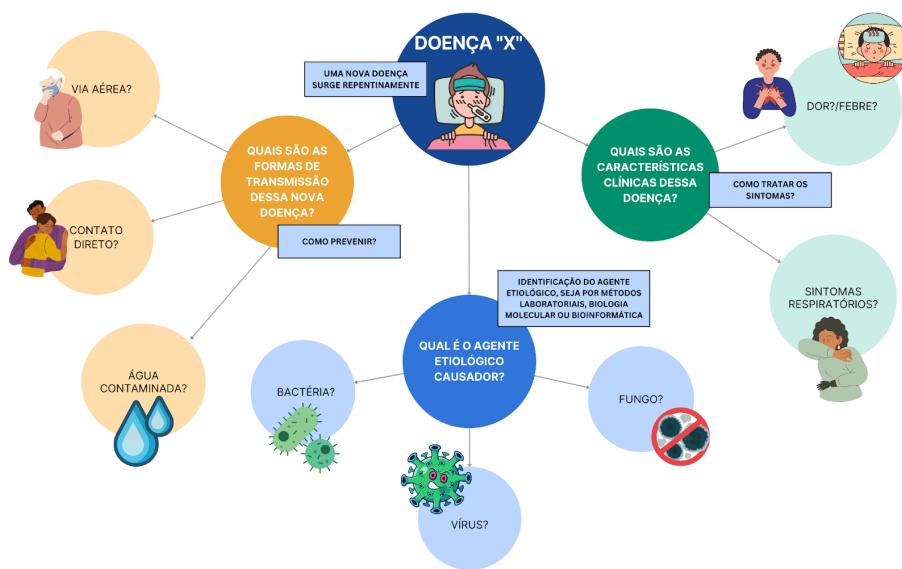


Figura 4.1: Representação das principais questões a serem respondidas com o surgimento de uma nova doença. Fonte: autoria própria.

A bioinformática, presente desde meados da década de 1970 e em grande evolução desde então, **se tornou uma das grandes ferramentas na investigação da pandemia de COVID-19**. Como um dos principais métodos empregados na

biologia molecular e na bioinformática, as técnicas de sequenciamento de nova geração (NGS) foram os protagonistas durante o período pandêmico. Através dessa metodologia, o sequenciamento desse vírus pôde ser executado de forma rápida e eficiente. Após a obtenção da sequência genômica do vírus e as análises de bioinformática serem realizadas, foi demonstrado que a doença era causada por um novo Coronavírus pertencente a linhagem dos betacoronavírus, possuindo um genoma de mais de 29 kB, permitindo o conhecimento do agente etiológico causador da doença [3]. Porém, algumas perguntas ainda precisavam ser respondidas, como: “Qual é a origem dessa doença em humanos?”. Para isto, através da análise dos resultados do sequenciamento, alinhamento de sequências e análises de filogenética por ferramentas de bioinformática, foi revelado que o novo betacoronavírus, agora denominado SARS-CoV-2, possuía uma grande similaridade com outros betacoronavírus originários de morcegos [4], que já haviam sido detectados muito anteriormente na China e uma similaridade alta com o MERS-CoV, um coronavírus causador da Síndrome Respiratória Médio Oriente (MERS) em 2012 [5].

O volume de dados gerados por bioinformática desde o início da pandemia é expressivo, atualmente ultrapassando mais de 15 milhões de genomas de SARS-CoV-2 sequenciados no mundo inteiro e hospedadas em um único banco de dados [6-7]. Isso tem aberto novas portas para o desenvolvimento e triagem de potenciais novos antivirais e o desenvolvimento de novas vacinas. Através desses dados, houve o surgimento de outros bancos de dados, desta vez de potenciais antivirais para SARS-CoV-2, auxiliando na previsão de proteínas-alvo do vírus e triagem de novos compostos contra alvos que podem ter efeitos inibitórios na polimerase viral do SARS-CoV-2 por meio de *docking* molecular [8-9].

Muitas perguntas foram respondidas neste período graças ao avanço da tecnologia e bioinformática, e ainda há muitas outras a serem resolvidas futuramente. Ainda teremos que lidar futuramente com surgimento de novos desafios, como novos agentes causadores de doenças, e para isto, novas técnicas, novos protocolos de biologia molecular e bioinformática deverão ser desenvolvidos. Como, por exemplo, o surgimento de novos modelos tecnológicos, como o desenvolvimento atual de modelos de inteligência artificial que vêm sendo

aplicados e estão avançando de forma tão abrupta quanto o espalhamento da COVID-19 no mundo. Dessa forma, é esperado que futuramente consigamos utilizar desta ferramenta também para responder novos desafios e epidemias.

Saiba mais 4.1

Este artigo está disponível em <https://bioinfo.com.br/o-codigo-covid-a-investigacao-da-pandemia-da-covid-19-atraves-da-bioinformatica/>

## 4.1 Referências

- [1] ZHU, N. et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N Engl J Med*, v. 382, n. 8, p. 727-733, 20 fev. 2020. DOI: 10.1056/NEJMoa2001017. PMID: 31978945; PMCID: PMC7092803
- [2] WU, J.T.; LEUNG, K.; LEUNG, G.M. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. *Lancet*, [S.l.], v. 395, n. 10225, p. 689-697, fev. 2020. DOI: 10.1016/S0140-6736(20)30260-9.
- [3] Chan JF, et al. Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan. *Emerg Microbes Infect*. 2020 Jan 28;9(1):221-236. doi: 10.1080/22221751.2020.1719902.
- [4] Zhou P et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*. 2020 Mar;579(7798):270-273. doi: 10.1038/s41586-020-2012-7.
- [5] Lu R et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet* 2020. 395(10224):565-574. doi: 10.1016/S0140-6736(20)30251-8.
- [6] Banco de dados de genomas Gisaid, EpiCov. Acessado em 18 de abril de 2023. Disponível em: <https://www.epicov.org>
- [7] Libório, L; Resende, V. Introdução aos bancos de dados biológicos. In: BIOINFO – Revista Brasileira de Bioinformática. Ed. 1. Julho, 2021. doi: 10.51780/978-6-599-275326-16
- [8] Martin et al. CORDITE: The Curated CORona Drug InTERactions Database for SARS-CoV-2. *iScience*. 2020 Jul 24;23(7):101297. doi: 10.1016/j.isci.2020.101297
- [9] Santos, L. Docagem molecular: em busca do encaixe perfeito e acessível. In: BIOINFO – Revista Brasileira de Bioinformática. Edição 01. Julho, 2021. DOI: 10.51780/978-6-599-275326-09

# 5 DESAFIOS NA PADRONIZAÇÃO DA ANOTAÇÃO GENÔMICA

## Autores 5.1

Diego Lucas Neres Rodrigues 

Revisão: Ana Carolina Silva Bulla , Bibiana Fam 

## Cite este artigo 5.1

Rodrigues, DLN. Desafios na padronização da anotação genômica. BIOINFO. ISSN: 2764-8273. Vol. 3. p.05 (2023). doi: 10.51780/bioinfo-03-05

## Resumo 5.1

### Opiniões & Perspectivas

**A**BIOINFORMÁTICA é uma área relativamente nova na intersecção da biologia e ciência da computação. Ainda assim, trouxe consigo diversos avanços tecnológicos, atualmente considerados imprescindíveis para o meio científico mundial [1]. Podemos citar a exemplo as diversas atualizações das tecnologias de sequenciamento que ocorreram somente nas últimas duas décadas [2]. Todavia, sendo uma área do saber recente, a bioinformática quanto à ciência ainda possui pontos de melhoria. Nesse contexto, um tópico que sempre está em voga é a pouca padronização da anotação genômica [3].

O **processo de anotação** é essencial para o desenvolvimento de metodologias cuja base parte da análise do material genético, tais como pan-genômica e taxogenômica. Em suma, essa etapa da genômica comparativa consiste em dar sentido às sequências biológicas, indicando pontos relacionados aos padrões de presença gênica, possíveis produtos e possíveis funções [4]. Contudo, esse é um passo que, acima de qualquer outro ponto, é dependente de comparações contra bancos de dados. Afinal, para se conhecer o produto que melhor deriva de uma sequência, é preciso comparar a sequência em questão contra outras possíveis candidatas já estudadas [5]. Porém, em bancos de dados imensos e não-curados, sequências idênticas podem possuir diferentes nomes de produto e estarem relacionadas a diferentes genes, ou seja, duas sequências iguais recebem por vezes nomes não-sinônimos que confundem as ferramentas de predição de contaminam os bancos de dados com informações equivocadas. Outra questão é a presença de genes ou proteínas fragmentadas em bancos de dados utilizados para anotação. Utilizar dados dessa qualidade diminui a acurácia da anotação, já que por vezes pode alterar a fase de leitura ou levar a **predição de pseudogenes** [6-7].

Pela lógica, então, o problema da padronização da anotação genômica deriva diretamente da ausência de padronização dos bancos de dados utilizados para essa função [5-8].

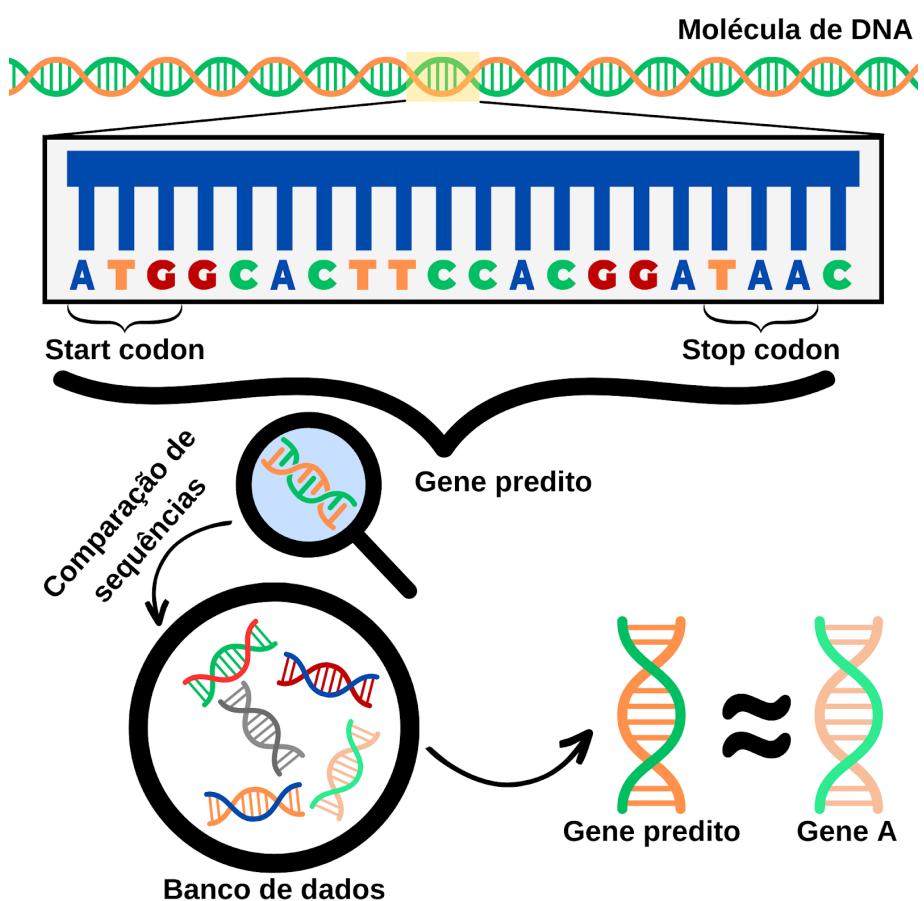


Figura 5.1: Simplificação gráfica do processo de anotação gênica em um genoma bacteriano. Fonte: próprio autor.

Não obstante, existem alternativas que contornam o problema. Caso apenas uma anotação simples seja necessária, é possível realizar uma anotação automática de todo um *dataset* de genomas por meio de uma mesma ferramenta ou *pipeline*. Esse passo visa padronizar os meios de comparação entre as sequências incluídas na análise, garantindo que todas foram comparadas pelo mesmo método e contra as mesmas referências. Porém, essa opção culmina em um passo adicional no *pipeline* de genômica comparativa, que por vezes é por si só extenso, e pode ser custoso temporalmente a depender do *dataset* inicial. Ao utilizar uma mesma ferramenta para a comparação e o mesmo banco de dados biológicos, espera-se que sequências similares sejam computadas pela mesma métrica e possuam a mesma terminologia e nomenclatura [7]. Tal fato implica em uma melhor acurácia do resultado final, e permite que o analista trabalhe com um dado limpo e não redundante.

Outra alternativa é utilizar anotações públicas de um mesmo banco de dados genômico. Com a criação do banco RefSeq, houve um aumento na qualidade da anotação de genomas disponíveis na plataforma *National Center for Biotechnology Information* (NCBI) – um dos maiores repositórios de genes e genomas do mundo [9-10]. Essa melhora nas características gerais de anotações públicas do NCBI se deve a uma curadoria fina dos dados depositados globalmente pelos desenvolvedores e envolvidos com a plataforma. A utilização de dados curados aumenta o número de proteínas hipotéticas preditas ao final do processo de anotação, porém eleva a qualidade dos produtos anotados, pois garante sua veracidade. Esse foi um passo importante para melhorar a reproduzibilidade de trabalhos genômicos, todavia, essa padronização ainda carece de curadoria por parte daqueles que obtêm esses dados [11].

Portanto, cabe ao bioinformata ou simpatizante da bioinformática a tarefa de selecionar o melhor banco de dados para realizar o processo de anotação, além de desenvolver métodos que solucionem o problema a longo prazo. Vale ressaltar que a bioinformática é uma potência científica, tendo como proposta auxiliar na busca pela solução de problemas biológicos complexos, sendo, portanto, a área mais adequada para lidar com essa problemática relacionada a ninguém menos que ela mesma.

### Saiba mais 5.1

Este artigo está disponível em <https://bioinfo.com.br/desafios-na-padronizacao-da-anotacao-genomica/>

## 5.1 Referências

- [1] Bayat A 2002 Science, medicine, and the future: Bioinformatics BMJ: British Medical Journal 324 1018. DOI: 10.1136/bmj.324.7344.1018
- [2] Meera Krishna B, Khan M A and Khan S T (2019). Next-Generation Sequencing (NGS) Platforms: An Exciting Era of Genome Sequence Analysis Microbial Genomics in Sustainable Agroecosystems: Volume 2 ed V Tripathi, P Kumar, P Tripathi, A Kishore and M Kamle (Singapore: Springer) pp 89–109. DOI: 10.1007/978-981-32-9860-6\_6
- [3] Salzberg S L 2019 Next-generation genome annotation: we still struggle to get it right Genome Biology 20 92. DOI: 10.1186/s13059-019-1715-2
- [4] Dominguez Del Angel V, Hjerde E, Sterck L, Capella-Gutierrez S, Notredame C, Vinnere Pettersson O, Amselem J, Bouri L, Bocs S, Klopp C, Gibrat J-F, Vlasova A, Leskosek B L, Soler L, Binzer-Panchal M and Lantz H 2018 Ten steps to get started in Genome Assembly and Annotation F1000Res 7 ELIXIR-148. DOI: 10.12688/f1000research.13598.1
- [5] Médigue C and Moszer I 2007 Annotation, comparison and databases for hundreds of bacterial genomes Research in Microbiology 158 724–36. DOI: 10.1016/j.resmic.2007.09.009
- [6] Costa M A, Guterres A. A Bioinformática na busca implacável: Onde estão os pseudogenes? In: BIOINFO – Revista Brasileira de Bioinformática. Edição #01. Julho, 2022. Acesso: <https://bioinfo.com.br/a-bioinformatica-na-busca-implacavel-onde-estao-os-pseudogenes>. doi: 10.51780/978-65-992753-5-7
- [7] Klimke W, O'Donovan C, White O, Brister J R, Clark K, Fedorov B, Mizrahi I, Pruitt K D and Tatusova T 2011 Solving the Problem: Genome Annotation Standards before the Data Deluge Stand Genomic Sci 5 168–93. DOI: 10.4056/sigs.2084864
- [8] Maia G A, Filho V B, Kawagoe E K, Teixeira Soratto T A, Moreira R S, Grisard E C and Wagner G 2022 AnnotaPipeline: An integrated tool to annotate eukaryotic proteins using multi-omics data Frontiers in Genetics 13. DOI: <https://doi.org/10.3389/fgene.2022.1020100>
- [9] Nasko D J, Koren S, Phillippy A M and Treangen T J 2018 RefSeq database growth influences the accuracy of k-mer-based lowest common ancestor species identification Genome Biol 19 165. DOI: 10.1186/s13059-018-1554-6

[10] O'Leary N A, et al. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation Nucleic Acids Res 44 D733-745. DOI: 10.1093/nar/gkv1189

[11] McDonnell E, Strasser K and Tsang A (2018). Manual Gene Curation and Functional Annotation Methods Mol Biol 1775 185–208. DOI: 10.1007/978-1-4939-7804-5\_16.

# 6

## DESAFIANDO A RESISTÊNCIA

### ANTIMICROBIANA: O POTENCIAL TERAPÊUTICO DA FAGOTERAPIA

Autores 6.1

Bruna Espiño dos Santos 

Revisão: Diego Mariano 

Cite este artigo 6.1

Santos, BES. **Desafiando a resistência antimicrobiana: o potencial terapêutico da fagoterapia.** BIOINFO. ISSN: 2764-8273. Vol. 3. p.06 (2023). doi: 10.51780/bioinfo-03-06

## Resumo 6.1

### Opiniões & Perspectivas

**A** UTILIZAÇÃO de vírus como nossos aliados pode causar certa apreensão, especialmente após a experiência trágica vivida durante a pandemia de COVID-19. No entanto, um diferente problema de saúde coletiva tem permanecido invisível aos olhos da população: a resistência antimicrobiana (RAM). Atualmente, a RAM tem se tornado uma crescente ameaça à saúde pública global, afetando seres humanos, animais e vegetais. O uso indiscriminado de antibióticos tem intensificado esse problema, levando ao crescimento do número de mortes por infecções bacterianas – uma cifra que se aproxima dos óbitos por HIV e malária [1]. Desse modo, a comunidade científica tem dado grande ênfase ao desenvolvimento de novos tratamentos eficazes, como é o caso da utilização de bacteriófagos.

## 6.1 O que são bacteriófagos?

Os bacteriófagos são vírus que possuem a habilidade de predar bactérias (Figura 6.1). Tratam-se da entidade mais amplamente distribuída no planeta, podendo ser encontrada em qualquer ambiente que contenha uma bactéria. Além disso, segundo Keen (2015) [2], estima-se que haja cerca de 1 trilhão de fagos para cada grão de areia no mundo.

## 6.2 Funcionamento da terapia

A terapia baseada em bacteriófagos, conhecida como “fagoterapia”, é uma estratégia terapêutica promissora, que pode ser administrada por meio de coqueteis – intravenosos, por via oral ou nasal – contendo diferentes fagos. O mecanismo de ação dos bacteriófagos ocorre de maneira altamente específica, por meio do reconhecimento de receptores presentes na superfície das bactérias-alvo – ao contrário dos antibióticos convencionais, que acabam por degradar a microbiota essencial ao ser humano. Além disso, os bacteriófagos isolados nos

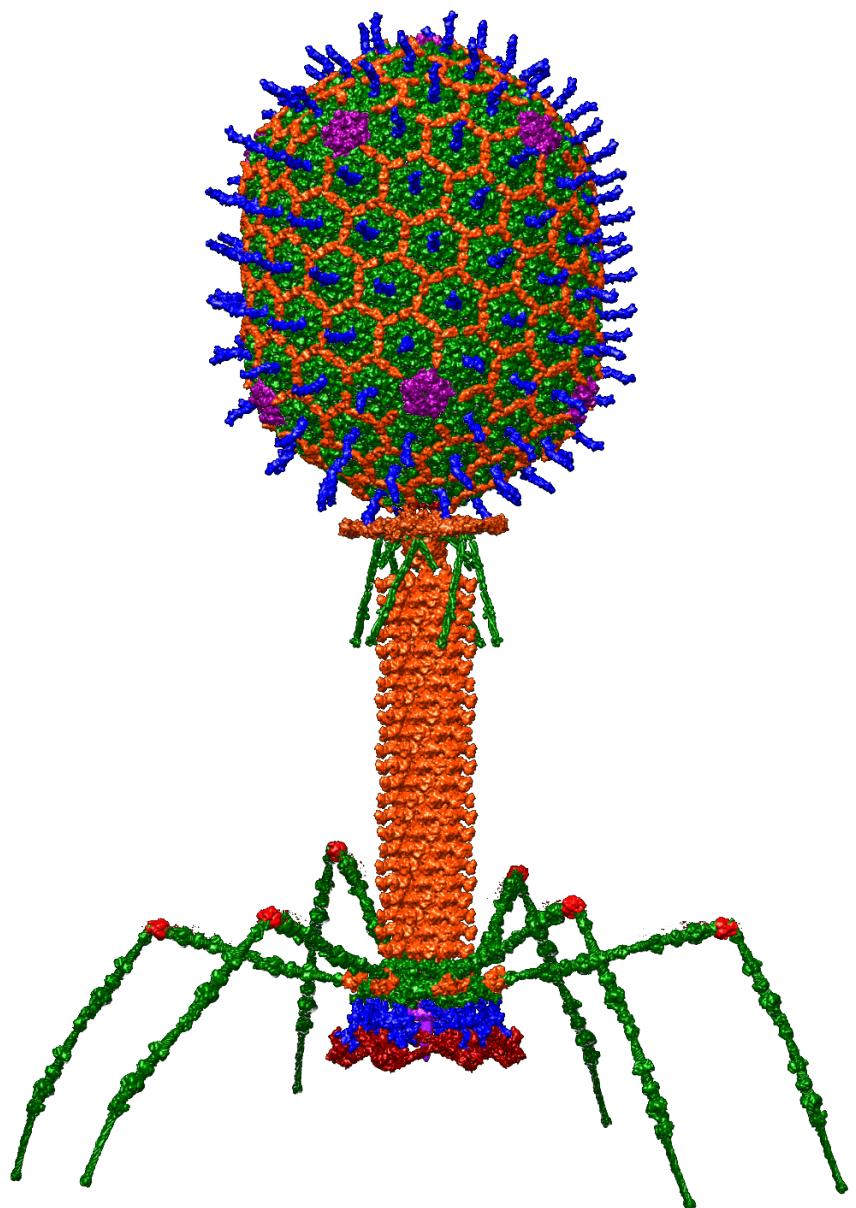


Figura 6.1: Estrutura 3D de bacteriófago T4. Imagem gerada através do Software ChimeraX. Fonte: Victor Padilla-Sánchez (CC-BY 4.0) [3].

coquetéis podem ser naturais ou geneticamente modificados e são selecionados com base em seu potencial de interação com a bactéria-alvo. Visto isso, a ligação do fago à célula hospedeira desencadeia a liberação do material genético viral, culminando na lise da bactéria (Figura 6.2). Portanto, diante da crescente resistência antimicrobiana, a fagoterapia surge como uma alternativa terapêutica de grande relevância.

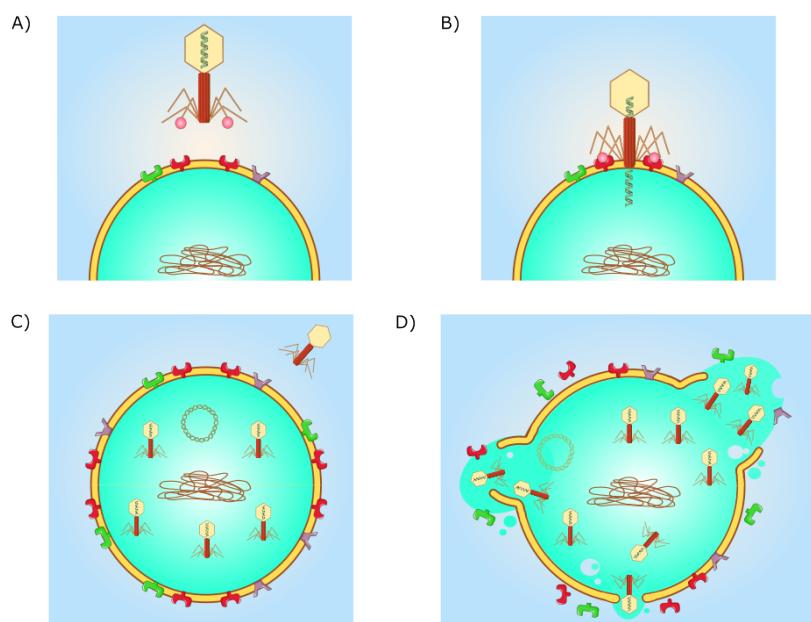


Figura 6.2: Representação do ciclo lítico do bacteriófago. A) O fago possui proteínas ligantes (representadas em esferas rosas) que reconhecem receptores de membrana específicos (receptores vermelhos) da bactéria-alvo. B) A partir do reconhecimento receptor-ligante, o bacteriófago adere-se à membrana da bactéria e, em seguida, libera o seu material genético. C) O material genético introduzido na célula é reconhecido por proteínas bacterianas, que transcrevem e traduzem os genes virais, levando à síntese de novos bacteriófagos. D) Além dos próprios bacteriófagos, o material genético viral também contém genes que codificam enzimas destrutivas, capazes de lisar a célula hospedeira. Dessa forma, os bacteriófagos recém-sintetizados são liberados e inicia-se um novo ciclo de infecção. Fonte: Próprio autor.

### 6.3 A fagoterapia é secular

O bacteriófago foi descoberto por William Twort (1915) e Felix d'Herelle (1917) na Europa e seus primeiros ensaios clínicos como tratamento antimicrobiano

datam desse mesmo período. Desse modo, há mais de 1 século a fagoterapia começou a ser aplicada em seres humanos, principalmente em países do leste europeu, para combater infecções como febre tifoide, cólera e disenteria. No entanto, com a Guerra Fria, a fagoterapia foi deixada de lado pelos países ocidentais e substituída por novos antibióticos, como a penicilina (EUA) e sulfonamidas (Alemanha). Entretanto, com o aumento dos casos de RAM, durante as décadas de 1990 e 2000, os estudos envolvendo a fagoterapia foram resgatados mundialmente. Países do leste europeu nunca interromperam sua pesquisa e aplicação terapêutica e possuem destaque na fundação de centros especializados, como a *Phage Therapy Unit* (PTU) – associada ao *Ludwik Hirschfeld Institute of Immunology and Experimental Therapy* – na Polônia.

## 6.4 Vantagens da fagoterapia

As bactérias podem se defender dos vírus por meio de mutações que afetam seus receptores. Isso pode levar à seguinte questão: por que as bactérias não simplesmente modificam ou excluem o receptor e se tornam imunes ao vírus? No entanto, a resposta é mais complexa do que parece, uma vez que esses receptores possuem diferentes funções, como a liberação de substâncias tóxicas e a movimentação celular – como em casos de receptores incrustados no flagelo bacteriano –, e não apenas a adesão do vírus. Ademais, tanto o DNA bacteriano quanto o DNA viral são sujeitos a mutações durante o processo de replicação no citoplasma hospedeiro, o que permite que a mutação viral naturalmente acompanhe a mutação bacteriana.

## 6.5 O papel da bioinformática

A bioinformática constitui uma etapa imprescindível para o delineamento de um coquetel fagoterápico. Para que um bacteriófago seja eficaz na luta contra as bactérias, é crucial que ele tenha ação lítica, ou seja, que tenha capacidade de inserir seu DNA na célula-alvo, resultando na autodestruição desta última. Assim, técnicas de sequenciamento e mapeamento genético possibilitam a identificação de vírus que apresentem genes codificantes de proteínas relacionadas a esse

processo, como holinas e endolisinas. Além disso, uma vez que os fagos são específicos em relação aos receptores de superfície bacterianos, a identificação e seleção da linhagem que ataca a bactéria em questão torna-se fundamental.

Saiba mais 6.1

Este artigo está disponível em <https://bioinfo.com.br/desafiando-a-resistencia-antimicrobiana-o-potencial-terapeutico-da-fagoterapia/>

## 6.6 Referências

- [1] MURRAY, C. J. L. et al. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *The Lancet*, v. 399, n. 10325, p. 629–655, 2022.  
[https://doi.org/10.1016/S0140-6736\(21\)02724-0](https://doi.org/10.1016/S0140-6736(21)02724-0).
- [2] KEEN, E. C. A century of phage research: Bacteriophages and the shaping of modern biology. *BioEssays: news and reviews in molecular, cellular and developmental biology*, v. 37, n. 1, p. 6–9, 2015. <https://doi.org/10.1002/bies.201400152>.
- [3] Victor Padilla-Sánchez (CC-BY 4.0). [https://commons.wikimedia.org/wiki/File:Bacteriophage\\_t4\\_and\\_pack\\_machine\\_wiki.png](https://commons.wikimedia.org/wiki/File:Bacteriophage_t4_and_pack_machine_wiki.png).
- [4] BARRON, M. Phage Therapy: Past, Present and Future.  
<https://asm.org/443/Articles/2022/August/Phage-Therapy-Past,-Present-and-Future>.
- [5] BRIVES, C.; POURRAZ, J. Phage therapy as a potential solution in the fight against AMR: obstacles and possible futures. *Palgrave Communications*, v. 6, n. 1, p. 1–11, 2020.  
<https://doi.org/10.1057/s41599-020-0478-4>.
- [6] GONZALEZ, F; SCHARE, B. E. Natural Bacteria Killers: How Bacteriophages Find and Eliminate Their Hosts. <https://kids.frontiersin.org/articles/10.3389/frym.2021.574664>.
- [7] GUZINA, J.; DJORDJEVIC, M. Bioinformatics as a first-line approach for understanding bacteriophage transcription. *Bacteriophage*, v. 5, n. 3, p. e1062588, 2015.  
<https://doi.org/10.1080/21597081.2015.1062588>.
- [8] LAXMINARAYAN, R. The overlooked pandemic of antimicrobial resistance. *The Lancet*, v. 399, n. 10325, p. 606–607, 2022. [https://doi.org/10.1016/S0140-6736\(22\)00087-3](https://doi.org/10.1016/S0140-6736(22)00087-3).
- [9] SANCHEZ, B. C. et al. Development of Phage Cocktails to Treat *E. coli* Catheter-Associated Urinary Tract Infection and Associated Biofilms. *Frontiers in Microbiology*, v. 13, 2022. <https://doi.org/10.3389/fmicb.2022.796132>.

[10] SVIRCEV, A.; ROACH, D.; CASTLE, A. Framing the Future with Bacteriophages in Agriculture. *Viruses*, v. 10, n. 5, 2018. <https://doi.org/10.3390/v10050218>.

[11] ŹACZEK, M. et al. Phage Therapy in Poland – a Centennial Journey to the First Ethically Approved Treatment Facility in Europe. *Frontiers in Microbiology*, v. 11, 2020. <https://doi.org/10.3389/fmicb.2020.01056>

# 7

## A BIOINFORMÁTICA COMO ALIADA DA BIOTECNOLOGIA AGRÍCOLA

### Autores 7.1

Ariany Rosa Gonçalves 

Revisão: Wylerson Guimarães Nogueira 

### Cite este artigo 7.1

Gonçalves, AR. **A Bioinformática como aliada da Biotecnologia Agrícola.** BIOINFO. ISSN: 2764-8273. Vol. 3. p.07 (2023). doi: 10.51780/bioinfo-03-07

## Resumo 7.1

### Opiniões & Perspectivas

**A**BIOINFORMÁTICA é um campo interdisciplinar que combina ciência da computação, estatística, matemática e biologia, e tem desempenhado um papel crucial na pesquisa genômica [1, 2]. Com o rápido avanço das tecnologias de sequenciamento de alto rendimento (HTS, do inglês *High Throughput Sequencing*) e o grande volume de dados gerados, a bioinformática se tornou uma ferramenta essencial para gerenciar, processar e analisar esses dados [3]. Na área da biotecnologia agrícola, a bioinformática também tem ganhado um papel importante. Nos últimos anos, estudos genômicos de microrganismos, especialmente de bactérias promotoras de crescimento vegetal (PGPB, do inglês *Plant Growth-Promoting Bacteria*) [4–7], têm sido cada vez mais relevantes, uma vez que essas bactérias, além de promoverem o crescimento das plantas, podem aumentar a produtividade de diversas culturas agrícolas, reduzindo o uso de fertilizantes e pesticidas sintéticos [8–12].

Estudos genômicos de PGPB revelam informações importantes sobre sua composição genética e capacidades funcionais [13, 14]. Abordagens comparativas podem identificar, por exemplo, genes envolvidos na síntese de hormônios de crescimento vegetal, como auxinas e citocininas, e genes envolvidos na biossíntese de metabólitos secundários com atividade antimicrobiana [15–17]. Esses estudos também podem indicar a presença de genes envolvidos na síntese de sideróforos, que são compostos quelantes de ferro que podem acelerar o crescimento vegetal a partir do aumento da disponibilidade deste micronutriente [18].

A metagenômica [19, 20] é outra abordagem de bioinformática que tem sido amplamente utilizada na pesquisa agrícola [21]. Ao analisarem dados metagenômicos de amostras da rizosfera de plantas, pesquisadores podem identificar novas cepas de PGPB com funções de interesse biotecnológico [22, 23]. Os resultados dos estudos genômicos e metagenômicos de PGPB têm levado a importantes avanços na agricultura. Por exemplo, foram identificadas cepas da bactéria *Bacillus amyloliquefaciens* que produzem compostos antifúngicos,

sendo essa relação de antagonismo aos fungos um fator de promoção à resistência de doenças vegetais [24, 25]. Além disso, cepas de PGPB têm sido utilizadas para promover o crescimento e o desenvolvimento de plantas em ambientes estressantes, como solos salinos e secos [26–28].

A bioinformática desempenha um papel fundamental na identificação de novas rotas metabólicas em microrganismos, impulsionando o desenvolvimento de produtos inovadores com aplicações biotecnológicas [29]. Um exemplo notável é a bactéria *Azospirillum brasilense*, que realiza a fixação biológica de nitrogênio e sintetiza o ácido indolacético (AIA), um hormônio vegetal essencial [30, 31]. Nesse contexto, a bioinformática torna-se uma ferramenta bastante útil, permitindo a identificação dos genes envolvidos na síntese desse fitormônio e otimizando sua produção em sistemas de fermentação [32].

A aplicação da bioinformática na agricultura vai além do estudo de microrganismos. A análise de dados genômicos de plantas permite identificar genes envolvidos em processos de interesse agrícola, como a resistência a doenças e a tolerância a estresses bióticos e abióticos [33, 34]. Essas informações são valiosas para o desenvolvimento de variedades de plantas mais resistentes e adaptadas a condições adversas, tornando-se especialmente relevantes diante das mudanças climáticas [35, 36]. Em resumo, a tecnologia HTS revolucionou o campo da genômica, permitindo o sequenciamento rápido de genomas, de tal modo que as ferramentas de bioinformática se tornaram cruciais para o processamento e análise desses dados, possibilitando a identificação de genes envolvidos em vários processos biológicos, como a absorção de nutrientes, interações planta-microrganismos e resposta ao estresse. Essas informações são fundamentais em diversos campos do conhecimento e, em especial, na biotecnologia agrícola.

#### Saiba mais 7.1

Este artigo está disponível em <https://bioinfo.com.br/a-bioinformatica-como-aliada-da-biotecnologia-agricola/>

## 7.1 Referências

- [1] Diniz, W. J. S., & Canduri, F. (2017). Bioinformatics: an overview and its applications. *Genetics and Molecular Research*, 16(1). DOI: <https://doi.org/10.4238/gmr16019645>
- [2] Lage, F. S. D., & Santos, F. B. (2021). Biologia e Computação: Um Casamento Perfeito. In: BIOINFO – Revista Brasileira de Bioinformática e Biologia Computacional (Edição 1.). DOI: <https://doi.org/10.51780/978-6-599-275326-02>
- [3] Freitas, A. S., & Barboza Pinto, H. (2021). Sequenciamento NGS: Status e Perspectivas. In: BIOINFO – Revista Brasileira de Bioinformática e Biologia Computacional. Ed.1. Vol. 1. DOI: 10.51780/978-6-599-275326-04
- [4] Jiang, L., Seo, J., Peng, Y., Jeon, D., Park, S. J., Kim, C. Y., ... Lee, J. (2023). Genome insights into the plant growth-promoting bacterium *Saccharibacillus brassicae* ATSA2T. *AMB Express*, 13(1), 9. DOI: <https://doi.org/10.1186/s13568-023-01514-1>
- [5] Schwab, S., Terra, L. A., & Baldani, J. I. (2018). Genomic characterization of *Nitrospirillum amazonense* strain CBAmC, a nitrogen-fixing bacterium isolated from surface-sterilized sugarcane stems. *Molecular Genetics and Genomics*, 293(4), 997–1016. DOI: <https://doi.org/10.1007/s00438-018-1439-0>
- [6] Matteoli, F. P., Passarelli-Araujo, H., Reis, R. J. A., da Rocha, L. O., de Souza, E. M., Aravind, L., ... Venancio, T. M. (2018). Genome sequencing and assessment of plant growth-promoting properties of a *Serratia marcescens* strain isolated from vermicompost. *BMC Genomics*, 19(1), 750. DOI: <https://doi.org/10.1186/s12864-018-5130-y>
- [7] Glick, B. R. (2012). Plant Growth-Promoting Bacteria: Mechanisms and Applications. *Scientifica*, 2012, 1–15. DOI: <https://doi.org/10.6064/2012/963401>
- [8] Andrade, L. A., Santos, C. H. B., Frezarin, E. T., Sales, L. R., & Rigobelo, E. C. (2023). Plant Growth-Promoting Rhizobacteria for Sustainable Agricultural Production. *Microorganisms*, 11(4), 1088. DOI: <https://doi.org/10.3390/microorganisms11041088>
- [9] Naher, U. A., Biswas, J. C., Maniruzzaman, Md., Khan, F. H., Sarkar, Md. I. U., Jahan, A., ... Kabir, Md. S. (2021). Bio-Organic Fertilizer: A Green Technology to Reduce Synthetic N and P Fertilizer for Rice Production. *Frontiers in Plant Science*, 12. DOI: <https://doi.org/10.3389/fpls.2021.602052>
- [10] Gómez-Godínez, L. J., Aguirre-Noyola, J. L., Martínez-Romero, E., Arteaga-Garibay, R. I., Ireta-Moreno, J., & Ruvalcaba-Gómez, J. M. (2023). A Look at Plant-Growth-Promoting Bacteria. *Plants*, 12(8), 1668. DOI: <https://doi.org/10.3390/plants12081668>
- [11] Dasgupta, D., Kumar, K., Miglani, R., Mishra, R., Panda, A. K., & Bisht, S. S. (2021). Microbial biofertilizers: Recent trends and future outlook. In: *Recent Advancement in Microbial Biotechnology* (pp. 1–26). Elsevier. DOI: <https://doi.org/10.1016/B978-0-12-822098-6.00001-X>

- [12] Khan, A., Panthari, D., Sharma, R. S., Punetha, A., Singh, A. V., & Upadhyay, V. K. (2023). Biofertilizers: a microbial-assisted strategy to improve plant growth and soil health. In: Advanced Microbial Techniques in Agriculture, Environment, and Health Management (pp. 97–118). Elsevier. DOI: <https://doi.org/10.1016/B978-0-323-91643-1.00007-7>
- [13] Wang, Z., Lu, K., Liu, X., Zhu, Y., & Liu, C. (2023). Comparative Functional Genome Analysis Reveals the Habitat Adaptation and Biocontrol Characteristics of Plant Growth-Promoting Bacteria in NCBI Databases. *Microbiology Spectrum*, 11(3). DOI: <https://doi.org/10.1128/spectrum.05007-22>
- [14] Dias, G. M., Sousa Pires, A., Grilo, V. S., Castro, M. R., Figueiredo Vilela, L., & Neves, B. C. (2019). Comparative genomics of *Paraburkholderia kururiensis* and its potential in bioremediation, biofertilization, and biocontrol of plant pathogens. *MicrobiologyOpen*, 8(8). DOI: <https://doi.org/10.1002/mbo3.801>
- [15] Eastman, A. W., Heinrichs, D. E., & Yuan, Z.-C. (2014). Comparative and genetic analysis of the four sequenced *Paenibacillus polymyxa* genomes reveals a diverse metabolism and conservation of genes relevant to plant-growth promotion and competitiveness. *BMC Genomics*, 15(1), 851. DOI: <https://doi.org/10.1186/1471-2164-15-851>
- [16] Narayanan, Z., & Glick, B. R. (2022). Secondary Metabolites Produced by Plant Growth-Promoting Bacterial Endophytes. *Microorganisms*, 10(10), 2008. DOI: <https://doi.org/10.3390/microorganisms10102008>
- [17] Jin, T., Ren, J., Li, Y., Bai, B., Liu, R., & Wang, Y. (2022). Plant growth-promoting effect and genomic analysis of the *P. putida* LWPZF isolated from *C. japonicum* rhizosphere. *AMB Express*, 12(1), 101. DOI: <https://doi.org/10.1186/s13568-022-01445-3>
- [18] Jiang, L., Seo, J., Peng, Y., Jeon, D., Park, S. J., Kim, C. Y., ... Lee, J. (2023). Genome insights into the plant growth-promoting bacterium *Saccharibacillus brassicae* ATSA2T. *AMB Express*, 13(1), 9. DOI: <https://doi.org/10.1186/s13568-023-01514-1>
- [19] Costa, S. de S. (2023). Metagenômica e Amplicon: perguntas frequentes e respostas essenciais. In: BIOINFO – Revista Brasileira de Bioinformática e Biologia Computacional. [cited 2023 Jul 25]. Available from: <https://bioinfo.com.br/a-bioinformatica-na-metagenomica-e-amplicon-perguntas-frequentes-e-respostas-essenciais/>
- [20] Thomas, T., Gilbert, J., & Meyer, F. (2012). Metagenomics – a guide from sampling to data analysis. *Microbial Informatics and Experimentation*, 2(1), 3. DOI: <https://doi.org/10.1186/2042-5783-2-3>
- [21] Nwachukwu, B. C., & Babalola, O. O. (2022). Metagenomics: A Tool for Exploring Key Microbiome With the Potentials for Improving Sustainable Agriculture. *Frontiers in Sustainable Food Systems*, 6. DOI: <https://doi.org/10.3389/fsufs.2022.886987>

- [22] Lemos, L. N., Mendes, L. W., Baldrian, P., & Pylro, V. S. (2021). Genome-Resolved Metagenomics Is Essential for Unlocking the Microbial Black Box of the Soil. *Trends in Microbiology*, 29(4), 279–282. DOI: <https://doi.org/10.1016/j.tim.2021.01.013>
- [23] Romero, M. F., Gallego, D., Lechuga-Jiménez, A., Martínez, J. F., Barajas, H. R., Hayano-Kanashiro, C., ... Alcaraz, L. D. (2021). Metagenomics of mine tailing rhizospheric communities and its selection for plant establishment towards bioremediation. *Microbiological Research*, 247, 126732. DOI: <https://doi.org/10.1016/j.micres.2021.126732>
- [24] Ji, S. H., Paul, N. C., Deng, J. X., Kim, Y. S., Yun, B.-S., & Yu, S. H. (2013). Biocontrol Activity of *Bacillus amyloliquefaciens* CNU114001 against Fungal Plant Diseases. *Mycobiology*, 41(4), 234–242. DOI: <https://doi.org/10.5941/MYCO.2013.41.4.234>
- [25] Yi, Y., Shan, Y., Liu, S., Yang, Y., Liu, Y., Yin, Y., ... Li, R. (2021). Antagonistic Strain *Bacillus amyloliquefaciens* XZ34-1 for Controlling Bipolaris sorokiniana and Promoting Growth in Wheat. *Pathogens*, 10(11), 1526. DOI: <https://doi.org/10.3390/pathogens10111526>
- [26] Kumar, A., Singh, S., Gaurav, A. K., Srivastava, S., & Verma, J. P. (2020). Plant Growth-Promoting Bacteria: Biological Tools for the Mitigation of Salinity Stress in Plants. *Frontiers in Microbiology*, 11. DOI: <https://doi.org/10.3389/fmicb.2020.01216>
- [27] Mishra, P., Mishra, J., & Arora, N. K. (2021). Plant growth promoting bacteria for combating salinity stress in plants – Recent developments and prospects: A review. *Microbiological Research*, 252, 126861. DOI: <https://doi.org/10.1016/j.micres.2021.126861>
- [28] Gupta, A., Mishra, R., Rai, S., Bano, A., Pathak, N., Fujita, M., ... Hasanuzzaman, M. (2022). Mechanistic Insights of Plant Growth Promoting Bacteria Mediated Drought and Salt Stress Tolerance in Plants for Sustainable Agriculture. *International Journal of Molecular Sciences*, 23(7), 3741. DOI: <https://doi.org/10.3390/ijms23073741>
- [29] Garcia-Lopez, E., Alcazar, P., & Cid, C. (2021). Identification of Biomolecules Involved in the Adaptation to the Environment of Cold-Loving Microorganisms and Metabolic Pathways for Their Production. *Biomolecules*, 11(8), 1155. DOI: <https://doi.org/10.3390/biom11081155>
- [30] Somers, E., Ptacek, D., Gysegom, P., Srinivasan, M., & Vanderleyden, J. (2005). *Azospirillum brasilense* Produces the Auxin-Like Phenylacetic Acid by Using the Key Enzyme for Indole-3-Acetic Acid Biosynthesis. *Applied and Environmental Microbiology*, 71(4), 1803–1810. DOI: <https://doi.org/10.1128/AEM.71.4.1803-1810.2005>
- [31] Fukami, J., Cerezini, P., & Hungria, M. (2018). Azospirillum: benefits that go far beyond biological nitrogen fixation. *AMB Express*, 8(1), 73. DOI: <https://doi.org/10.1186/s13568-018-0608-1>

- [32] Alkema, W., Boekhorst, J., Wels, M., & van Huijum, S. A. F. T. (2016). Microbial bioinformatics for food safety and production. *Briefings in Bioinformatics*, 17(2), 283–292. DOI: <https://doi.org/10.1093/bib/bbv034>
- [33] Nascimento, F. dos S., Rocha, A. de J., Soares, J. M. da S., Mascarenhas, M. S., Ferreira, M. dos S., Morais Lino, L. S., ... Amorim, E. P. (2023). Gene Editing for Plant Resistance to Abiotic Factors: A Systematic Review. *Plants*, 12(2), 305. DOI: <https://doi.org/10.3390/plants12020305>
- [34] Hamdan, M. F., Karlson, C. K. S., Teoh, E. Y., Lau, S.-E., & Tan, B. C. (2022). Genome Editing for Sustainable Crop Improvement and Mitigation of Biotic and Abiotic Stresses. *Plants*, 11(19), 2625. DOI: <https://doi.org/10.3390/plants11192625>
- [35] Batley, J., & Edwards, D. (2016). The application of genomics and bioinformatics to accelerate crop improvement in a changing climate. *Current opinion in plant biology*, 30, 78–81. DOI: <https://doi.org/10.1016/J.PBI.2016.02.002>
- [36] Marsh, J. I., Hu, H., Gill, M., Batley, J., & Edwards, D. (2021). Crop breeding for a changing climate: integrating phenomics and genomics with bioinformatics. *Theoretical and Applied Genetics*, 134(6), 1677–1690. DOI: <https://doi.org/10.1007/s00122-021-03820-3>

# 8

## A BIOINFORMÁTICA NA ERA PRÉ E PÓS-PANDEMIA

### Autores 8.1

Bibiana Sampaio de Oliveira Fam 

Revisão: Ana Paula Abreu , Aline Sampaio Cremonesi 

### Cite este artigo 8.1

Fam, BSO. **A bioinformática na era pré e pós-pandemia.** BIOINFO. ISSN: 2764-8273. Vol. 3. p.08 (2023). doi: 10.51780/bioinfo-03-08

## Resumo 8.1

### Opiniões & Perspectivas

**A BIOLOGIA COMPUTACIONAL** teve sua origem na década de 1950 [1], junto à descoberta da estrutura de DNA veio a necessidade da utilização de computadores para a análise de sequência de ácidos nucleicos e proteínas. Porém, foi na década de 1980 que a utilização de computadores para a análise de sequências biológicas começou a ser amplamente utilizada por pesquisadores de diferentes áreas de pesquisa [2]. Nos anos seguintes houve um crescimento exponencial na aplicação da biologia computacional em diferentes áreas de conhecimento, e na década de 1990 surgiram diferentes algoritmos que ainda hoje são considerados grandes avanços no campo da bioinformática. Como exemplo temos ferramentas de análise de sequências, como o BLAST, que permitem a comparação rápida e eficiente de sequências de DNA e proteínas [3] os *microarrays* implementados para análise de dados de expressão gênica [2,4]. Ganharam força também análises de dados de evolução molecular, que permitem entender como as espécies evoluíram ao longo do tempo e também como diferentes modificações em genes e proteínas levam a novas funções. Assim, podemos dizer que os anos 1990 foram fundamentais para o salto que viria a partir dos anos 2000 [2].

Com o avanço das tecnologias de sequenciamento de DNA e o aumento exponencial de dados genômicos disponíveis, surgem desafios e oportunidades no campo da bioinformática. Assim, os anos 2000 foram uma década de grande progresso na área, surgindo novos desafios e oportunidades na análise de dados biológicos. Observamos o aprimoramento de antigas ferramentas e o desenvolvimento de novas ferramentas para a análise de dados, como algoritmos de alinhamento de sequência genômicas entre diferentes espécies e indivíduos. A consolidação das ferramentas “ômicas” para análises de genômica, proteômica, transcriptômica e metabolômica [5]. Além disso, o grande volume de dados genômicos permitiu avanços nas áreas de evolução humana e molecular jamais imagináveis a anos atrás.

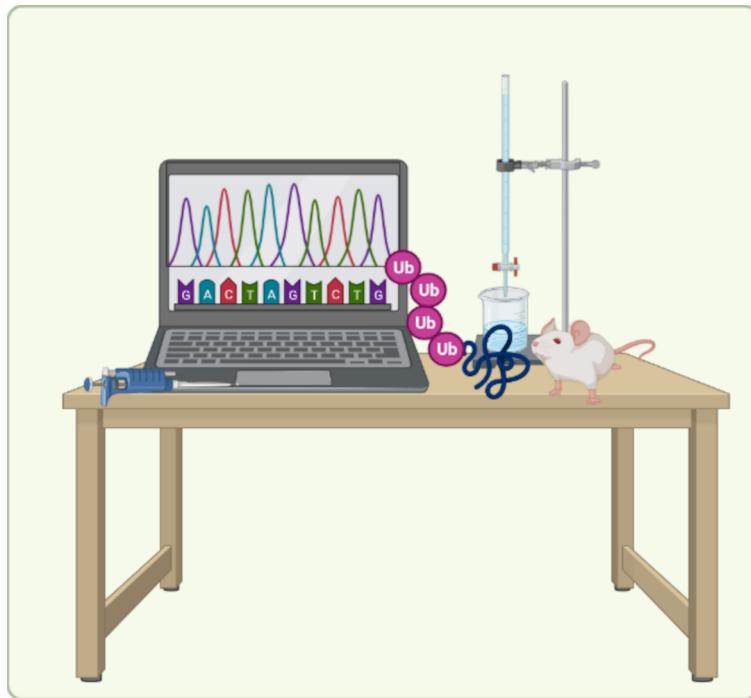


Figura 8.1: Sequenciamento. Fonte: autoria própria.

Todo este conhecimento consolidado na área da biologia computacional permitiu com que a humanidade respondesse de maneira rápida e certeira em um momento crucial de sua história: a **pandemia de COVID-19**. E assim podemos afirmar que uma das áreas que teve um papel crucial no combate à pandemia foi a bioinformática, que é a aplicação da tecnologia da informação para análise de dados biológicos. Nunca em um momentos críticos como este a ciência mundial teve uma resposta tão rápida em identificar um patógeno. Através das ferramentas de bioinformática tivemos capacidade de realizar a vigilância genômica e epidemiológica de SARS-CoV-2 e desenvolver vacinas em tempo recorde. Por isso, podemos dizer que a pandemia de Covid-19 trouxe à tona a importância da pesquisa científica e de tecnologia no enfrentamento de crises globais de saúde. Na era pós-pandemia, a bioinformática continuará sendo uma ferramenta essencial para a pesquisa em saúde, em diversas áreas além das já consolidadas ciências "ômicas" [6,7].

A bioinformática tem sido utilizada para entender melhor a resposta imune ao vírus e identificar fatores de risco para a Covid-19 e diferentes graus de suscetibilidade. Outra área em que a bioinformática pode ser aplicada é na identificação de variantes do vírus que possam escapar da imunidade já conferida pelas vacinas atuais, promovendo cada vez mais uma medicina personalizada, aprimorando a farmacogenética e a medicina de precisão, na descoberta de medicamentos, diagnóstico de doenças, desenvolvimento de vacinas [6,7,8].

Através da bioinformática, é possível analisar grandes quantidades de dados biológicos e encontrar informações relevantes para o desenvolvimento de novas terapias e tratamentos. Portanto, na era pós-pandemia, a bioinformática continuará sendo uma ferramenta crucial para a pesquisa em saúde no desenvolver novas terapias e tratamentos, auxiliando na melhoria de diagnóstico de doenças e entender melhor as interações entre genes, doenças e ambiente.

#### Saiba mais 8.1

Este artigo está disponível em <https://bioinfo.com.br/a-bioinformatica-na-era-pre-e-pos-pandemia/>

## 8.1 Referências

- [1] Gauthier, J.; Vincent, A. T.; Charette, S. J.; Derome, N. A brief history of bioinformatics. Brief Bioinform. 2019; 20:1981–1996. DOI: 10.1093/bib/bby063
- [2] Hogeweg, P. The Roots of Bioinformatics in Theoretical Biology. PLoS Comput Biol, 2011. DOI: 10.1371/journal.pcbi.1002021
- [3] Wheeler, D.; Bhagwat, M. BLAST QuickStart: Example-Driven Web-Based BLAST Tutorial. In: Bergman NH, editor. Chapter 9. Comparative Genomics: Volumes 1 and 2. Totowa (NJ): Humana Press; 2007.
- [4] Pertea, M.; The Human Transcriptome: An Unfinished Story. Genes, 2012. 3:344–360. DOI: 10.3390/genes3030344
- [5] Freitas, A. S.; Pinto, H. B. Sequenciamento NGS: Status e Perspectivas In: BIOINFO – Revista Brasileira de Bioinformática. Ed. 1. Vol. 1. Julho, 2021. DOI: 10.51780/978-6-599-275326-04

[6] Santos, L. Docagem molecular: em busca do encaixe perfeito e acessível. In: BIOINFO – Revista Brasileira de Bioinformática. Ed. 1. Vol. 1. Julho, 2021. DOI: 10.51780/978-6-599-275326-09

[7] Silva, L. X.; Bastos, L. L.; Santos, L. H. Modelagem computacional de proteínas. In: BIOINFO – Revista Brasileira de Bioinformática. Ed. 1. Vol. 1. Julho, 2021. DOI: 10.51780/978-6-599-275326-08

[8] Rosa, R. S. L., Esteves, M. E. A., Bulla, A. C. S., Silva, M. L. Preditores farmacocinéticos e toxicológicos in silico para via oral: conheça e análise ADMETox. BIOINFO- Revista Brasileira de Bioinformática 01. Julho, 2021. DOI: 10.51780/978-6-599-275326-08

# 9

## USO DO BLASTN NA CONSTRUÇÃO DE PRIMERS

### Autores 9.1

Bárbara Rebeca de Macedo Pinheiro 

Revisão: Savio Costa , Izabela Mamede 

### Cite este artigo 9.1

Pinheiro, BRM. **Uso do BLASTn na construção de primers.** BIOINFO. ISSN: 2764-8273. Vol. 3. p.09 (2023). doi: 10.51780/bioinfo-03-09

## Resumo 9.1

### Opiniões & Perspectivas

O ESTUDO da genética vem evoluindo nas últimas décadas e uma das técnicas revolucionárias para o estudo do DNA foi a **Reação em Cadeia da Polimerase (PCR)**. Essa técnica consiste na replicação, em milhares de cópias, de um fragmento de DNA através das etapas de desnaturação, anelamento e extensão. Na etapa de desnaturação ocorre a separação entre as duas fitas de DNA para que, durante a fase de anelamento, os *primers* se liguem a essas fitas e permitam a ação da Taq polimerase em criar uma nova fita de DNA no decorrer da fase de extensão. Os *primers* são oligonucleotídeos curtos, em média com 20 nucleotídeos, que são construídos e utilizados em pares, em que um se liga na fita sense e outro na fita antisense do DNA, como mostra a Figura 9.1. São chamados também de iniciadores pois são eles que indicam para a enzima Taq polimerase onde iniciar a formação da nova fita de DNA e flanqueiam a região alvo para que durante os ciclos da PCR somente essa região de interesse seja replicada. Sendo assim, a construção dos primers é de extrema importância e deve ser realizada com cuidado para que, durante a PCR, não ocorra a formação de dímeros ou ligações em regiões que não sejam a de interesse [1]. Nesse contexto, o uso de ferramentas de bioinformática para construção desses iniciadores é recomendado e amplamente utilizado, visando a diminuição de erros que possam acontecer durante o experimento.



Figura 9.1: Primers se ligando às fitas de DNA e delimitando o segmento de interesse. Fonte: Fábio Madonini, 2016 [3].

Desse modo, destaca-se aqui a utilização de uma ferramenta de bioinformática, entre outras como o Primer3 e o PrimerQuest, para que haja uma melhor construção de primers: o BLAST (*Basic Local Alignment Search Tool*). O BLAST pode ser utilizado tanto pela web, via servidor do NCBI (*National Center for Biotechnology Information*), quanto pode ser instalado localmente. Geralmente, os pesquisadores utilizam a versão *web* devido sua praticidade e rapidez ao enviar suas sequências para análise no servidor remoto do NCBI. O **BLAST** realiza a comparação entre sequências indicadas pelo pesquisador e sequências biológicas armazenadas em um banco de dados [2], sendo utilizado para se ter a certeza de que a sequência de nucleotídeos da região do seu interesse corresponde àquela encontrada no seu organismo de estudo. Para a PCR se usa o **BLASTn**, que pesquisa a sequência de nucleotídeos de interesse. Entre os benefícios em se utilizar o BLASTn é que essa ferramenta é executada no NCBI, uma instituição que apresenta um dos maiores bancos de dados de sequências genéticas, com atualizações regulares, além de oferecer uma interface amigável e apresentar fácil integração com outras ferramentas do NCBI. Nessa função, insere-se no programa a sequência, em formato FASTA, e se configura os parâmetros para a análise, como qual o organismo a ser estudado e qual o banco de dados escolhido para a busca. Caso a sequência indicada tenha similaridade com a sequência apontada no banco de dados, há a certeza de que aquela região do DNA corresponde com a real e a construção de primers pode ser iniciada efetivamente.

Essa ferramenta é extremamente necessária para quem trabalha com polimorfismos, que são variações genéticas que aparecem como consequências de mutações, podendo ter diferentes classificações a depender da mutação original. Logo, é necessário conhecer quais as sequências de nucleotídeos antes e depois daquela mutação, para que haja a construção de *primers* que se liguem somente naquela região flankeadora. O uso do BLAST é bem mais vasto do que o apresentado, podendo ser utilizado para pesquisar também sobre similaridades de aminoácidos e entre nucleotídeos e aminoácidos.

### Saiba mais 9.1

Este artigo está disponível em <https://bioinfo.com.br/uso-do-blastn-na-construcao-de-primers/>

## 9.1 Referências

- [1] EL-SAMAD, Hana et al. FORMAS DE TRABALHAR COM CÉLULAS: analisando células, moléculas e sistemas. In: ALBERTS, Bruce et al. Biologia Molecular da Célula. 6. ed. Porto Alegre: Artmed, 2017. Cap. 8. p. 473-477.
- [2] LIBÓRIO, Leandro; RESENDE, Victor Hugo. Introdução aos bancos de dados biológicos. In: Bioinfo – Revista Brasileira de Bioinformática e Biologia Computacional. Vol. 1. Ed. 1. Jul. 2021. Alfahelix. doi: 10.51780/978-6-599-275326-16
- [3] MANDONINI, Fabio. Primer per PCR. 2016. Disponível em: [https://commons.wikimedia.org/wiki/File:Primer\\_per\\_PCR.png](https://commons.wikimedia.org/wiki/File:Primer_per_PCR.png). Acesso em: 27 jul. 2023.

# 10

## BIOINFORMÁTICA NA LUTA CONTRA O CÂNCER: OS BANCOS DE DADOS NA PESQUISA ONCOLÓGICA

### Autores 10.1

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

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

### Cite este artigo 10.1

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

### Resumo 10.1

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

## 10.1 Introdução

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

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

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

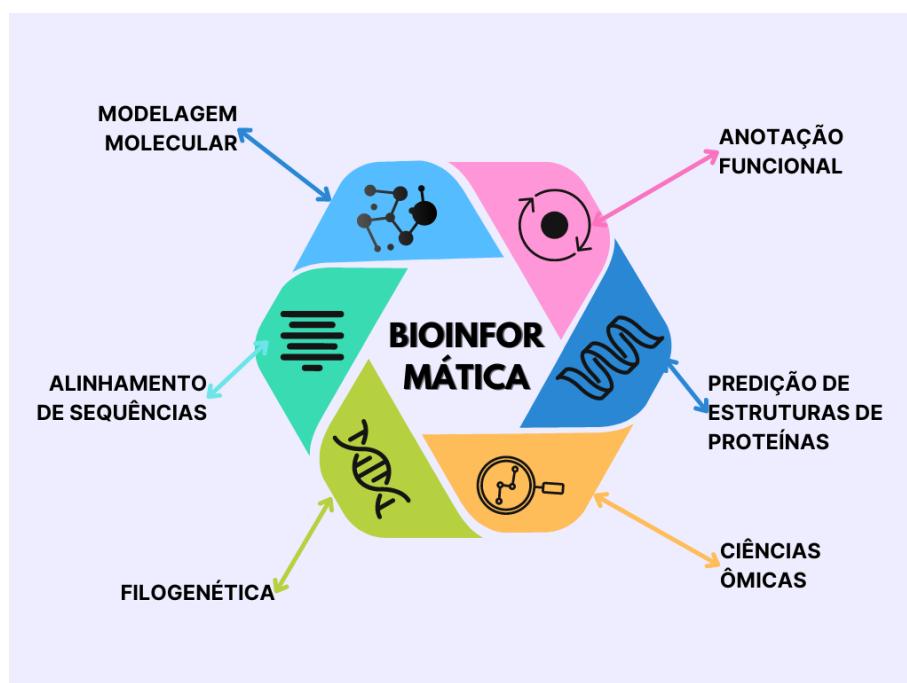


Figura 10.1: Aplicações da bioinformática e suas áreas de forma geral, ciências ômicas, filogenética, alinhamento de sequências, modelagem molecular, anotação funcional e predição de estruturas de proteínas. Fonte: adaptado de [14].

## 10.2 A Bioinformática na Oncologia

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

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

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

#### **10.2.1 Análise de expressão gênica:**

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

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

Algumas etapas envolvidas nesse processamento são:

- Pré-processamento dos dados, no qual, os dados brutos do sequenciamento de RNA são processados.
- Análise diferencial da expressão gênica que compara os níveis de expressão entre diferentes grupos de amostras com o intuito de identificar genes que apresentam alterações significativas na expressão.
- Análise funcional, quando os genes diferencialmente expressos são submetidos a análises funcionais para entender os processos biológicos afetados pelo mesmo.

### **10.2.2 Análises em proteômica:**

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

### **10.2.3 Predição e Mineração de Dados:**

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

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

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

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

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

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

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

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

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

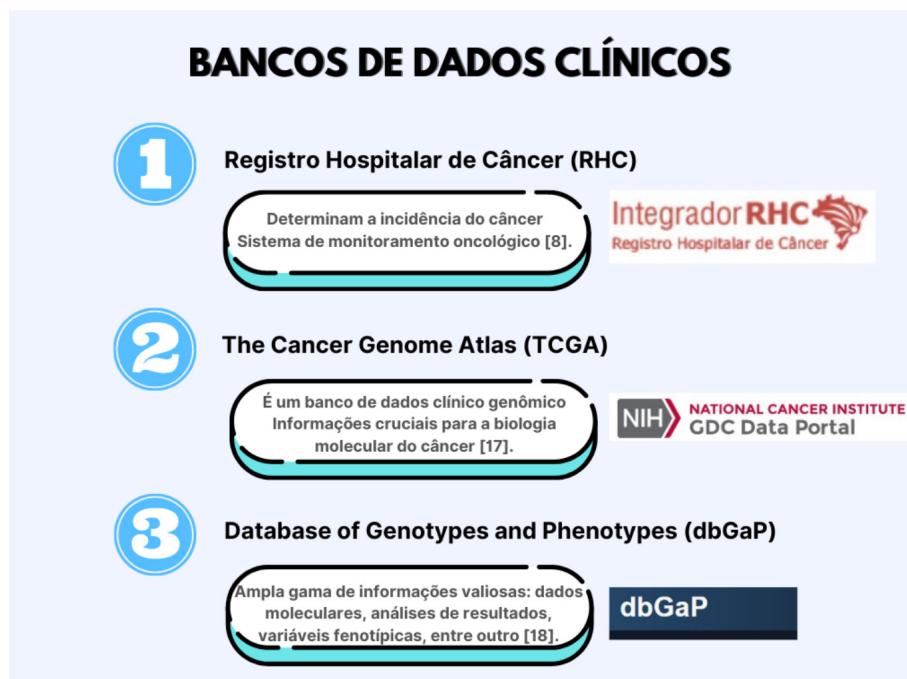


Figura 10.2: Bancos de dados clínicos. Fonte: [8; 17-18].

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

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

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

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

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

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

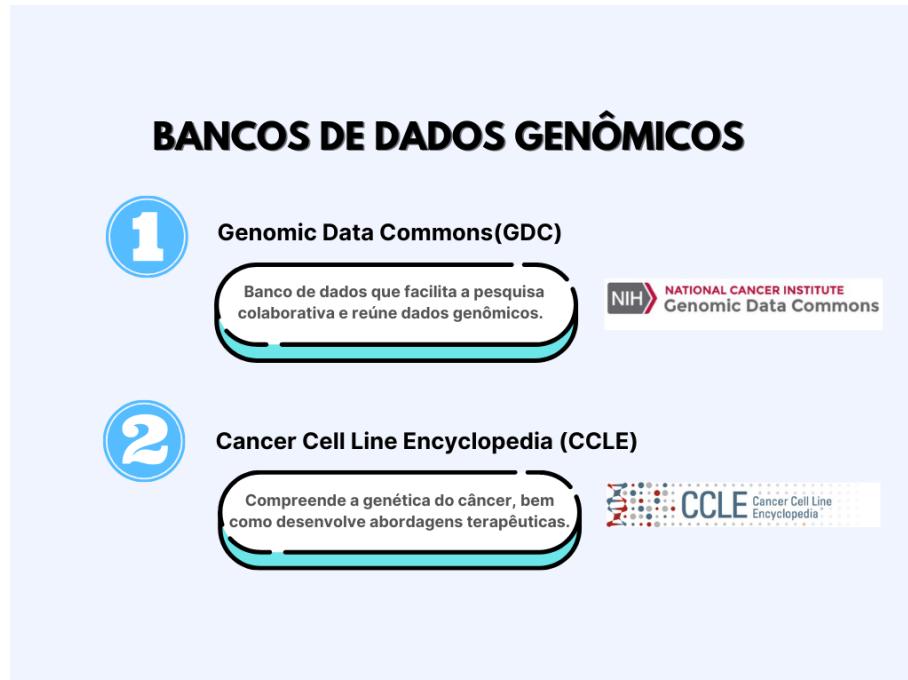


Figura 10.3: Bancos de dados genômicos. Fonte: autoria própria.

Saiba mais 10.1

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

## 10.4 Referências

- [1] CARVALHO, Mayra Costa da Cruz Gallo de; SILVA, Danielle Cristina Gregorio da. Sequenciamento de DNA de nova geração e suas aplicações na genômica de plantas. Ciência Rural, v. 40, p. 735-744, 2010.
- [2] CIEŚLIK, Marcin; CHINNAIYAN, Arul M. Cancer transcriptome profiling at the juncture of clinical translation. Nature Reviews Genetics, v. 19, n. 2, p. 93-109, 2018.
- [3] DAS, Tonmoy et al. Integration of online omics-data resources for cancer research. Frontiers in Genetics, v. 11, p. 578345, 2020.
- [4] GUIDO, Rafael VC; ANDRICOPULO, Adriano D.; OLIVA, Glaucius. Planejamento de fármacos, biotecnologia e química medicinal: aplicações em doenças infecciosas. Estudos avançados, v. 24, p. 81-98, 2010.

- [5] HAGEN, Joel B. The origins of bioinformatics. *Nature Reviews Genetics*, v. 1, n. 3, p. 231-236, 2000.
- [6] JOTHI, Neesha et al. Data mining in healthcare—a review. *Procedia computer science*, v. 72, p. 306-313, 2015.
- [7] KAUR, Ishleen; DOJA, M. N.; AHMAD, Tanvir. Data mining and machine learning in cancer survival research: an overview and future recommendations. *Journal of Biomedical Informatics*, v. 128, p. 104026, 2022.
- [8] KLIGERMAN, Jacob. Registro hospitalar de câncer no Brasil. *Revista Brasileira de Cancerologia*, v. 47, n. 4, p. 357-359, 2001.
- [9] LÜHRS, L. et al. Identificação de microssatélites e síntese de primers para erva-mate a partir de programas de bioinformática. 2021.
- [10] MINISTÉRIO DA SAÚDE (BR). INSTITUTO NACIONAL DE CâNCER JOSÉ ALENCAR GOMES DA SILVA. Informação dos registros hospitalares de câncer como estratégia de transformação: perfil do Instituto Nacional de Câncer José Alencar Gomes da Silva em 25 anos. 2012.
- [11] NEURALMED. Big Data na saúde: Por que a importância de olhar para esse universo? Disponível em: <https://www.neuralmed.ai/blog/big-data-na-saude>. Acesso em: 23 maio 2023.
- [12] NUSINOW, David P. et al. Quantitative proteomics of the cancer cell line encyclopedia. *Cell*, v. 180, n. 2, p. 387-402. e16, 2020
- [13] ORACLE. O que é um Banco de Dados? Disponível em: <https://www.oracle.com/br/database/what-is-database/>. Acesso em: 23 maio 2023.
- [14] SAFADY, Nágela G. Bioinformática: união entre ciência e tecnologia. Disponível em: <https://blog.varsomecs.com/bioinformatica/>. Acesso em: 22 maio 2023.
- [15] SANTOS, M. de O.; LIMA, F. C. da S. de; MARTINS, L. F. L.; OLIVEIRA, J. F. P.; ALMEIDA, L. M. de; CANCELA, M. de C. Estimativa de Incidência de Câncer no Brasil, 2023-2025. *Revista Brasileira de Cancerologia*, [S. l.], v. 69, n. 1, p. e-213700, 2023. DOI: 10.32635/2176-9745.RBC.2023v69n1.3700. Disponível em: <https://rbc.inca.gov.br/index.php/revista/article/view/3700>. Acesso em: 11 jul. 2023.
- [16] SHRUTHI, Basavaradhya Sahukar et al. Proteomics: A new perspective for cancer. *Advanced biomedical research*, v. 5, 2016.
- [17] TOMCZAK, Katarzyna; CZERWIŃSKA, Patrycja; WIZNEROWICZ, Maciej. Review The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemporary Oncology/Współczesna Onkologia*, v. 2015, n. 1, p. 68-77, 2015.

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

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

# 11

## INTRODUÇÃO À BIOLOGIA ESTRUTURAL DE PROTEÍNAS

### Autores 11.1

Rafael Pereira Lemos , Paulo Henrique dos Santos , Aline Rocha 

Revisão: Bibiana Fam , Filipe Teixeira , Carlos Capelini 

### Cite este artigo 11.1

Lemos, R; Santos, PH; Rocha, A. **Introdução à Biologia Estrutural de Proteínas.** BIOINFO.

ISSN: 2764-8273. Vol. 3. p.11 (2023). doi: 10.51780/bioinfo-03-11

### Resumo 11.1

**A**s proteínas são as macromoléculas mais abundantes nos sistemas biológicos, estando presentes em todas as células e tecidos. Apresentam uma diversidade de funções biológicas em nosso organismo, como: estruturação de células e tecidos, transporte e armazenamento de outras moléculas, receptores, hormônios, anticorpos, fatores de transcrição e enzimas. Esta gama de funções é possível devido à diversidade estrutural das proteínas. As proteínas são polímeros cujos monômeros compreendem resíduos dos **20 aminoácidos** que estão naturalmente presentes nestas moléculas. Estes aminoácidos são combinados em sequências diferentes para dar origem à diversidade de proteínas existentes nos diferentes organismos. As proteínas são produzidas no processo de tradução do RNA mensageiro (RNAm), a partir da união de unidades de aminoácidos carreados por RNAs transportadores (RNAt). Este carreamento é feito com base na sequência do molde de RNAm que por sua vez é produzido a partir do processo de transcrição da sequência codificante de DNA. Desta forma, as proteínas correspondem às moléculas pelas quais a informação genética é expressa [1].

## 11.1 Aminoácidos

Cada **aminoácido** apresenta um grupo amino terminal básico (-NH<sub>2</sub>), e um grupo carboxílico terminal ácido (-COOH), além de uma cadeia lateral variável (-R) que determina o tipo de aminoácido, e um hidrogênio que completa as quatro ligações ao carbono central, também chamado carbono alfa (C- $\alpha$ ) (Figura 11.1). Para a maioria dos aminoácidos, com exceção da glicina, o carbono alfa é quiral. Dessa forma, esses aminoácidos podem existir na forma de estereoisômeros (D- ou L-), sendo que nas proteínas são encontrados quase exclusivamente os estereoisômeros L- [1,2].

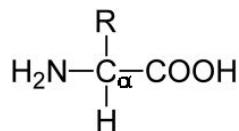


Figura 11.1: Estrutura geral de um aminoácido. Fonte: próprio autor.

Os aminoácidos podem ser representados em códigos de uma ou três letras (Tabela 11.1) [3].

Tabela 11.1: Classificação, códigos e propriedades dos 20 aminoácidos que constituem as proteínas. Fonte: adaptado de [1,3].

Grupo	Aminoácido	Código (3 letras)	Código (1 letra)	Massa Molecular (Da)	pKa (COOH)	pKa (NH3+)	pKa (R)	pI
Apolares	Glicina	Gly	G	75	2,34	9,60	–	5,97
	Alanina	Ala	A	89	2,34	9,69	–	6,01
	Prolina	Pro	P	115	1,99	10,96	–	6,48
	Valina	Val	V	117	2,32	9,62	–	5,97
	Leucina	Leu	L	131	2,36	9,60	–	5,98
	Isoleucina	Ile	I	131	2,36	9,68	–	6,02
	Metionina	Met	M	149	2,28	9,21	–	5,74
	Fenilalanina	Phe	F	165	1,83	9,13	–	5,48
Aromáticos	Tirosína	Tyr	Y	181	2,20	9,11	10,07	5,66
	Triptofano	Trp	W	204	2,38	9,39	–	5,89
	Serina	Ser	S	105	2,21	9,15	–	5,68
Polares não carregados	Treonina	Thr	T	119	2,11	9,62	–	5,87
	Cisteína	Cys	C	121	1,96	10,28	8,18	5,07
	Asparagina	Asn	N	132	2,02	8,80	–	5,41
	Glutamina	Gln	Q	146	2,17	9,13	–	5,65
	Lisina	Lys	K	146	2,18	8,95	10,53	9,74
Básicos	Histidina	His	H	155	1,82	9,17	6,00	7,59
	Arginina	Arg	R	174	2,17	9,04	12,48	10,76
	Aspartato	Asp	D	133	1,88	9,60	3,65	2,77
Ácidos	Glutamato	Glu	E	147	2,19	9,67	4,25	3,22

Eles podem ser classificados em essenciais, ou seja, aqueles que não são produzidos pelo nosso corpo e devem ser obtidos da dieta (histidina, isoleucina, leucina, lisina, metionina, fenilalanina, treonina e triptofano), ou não essenciais, que podem ser sintetizados pelo nosso organismo [4]. Porém, mais importante para a biologia estrutural é a classificação dos aminoácidos de acordo com as propriedades da cadeia lateral, sendo divididos em: alifáticos ou apolares, aromáticos, polares não carregados, carregados positivamente ou básicos e carregados negativamente ou ácidos (Figura 11.2).

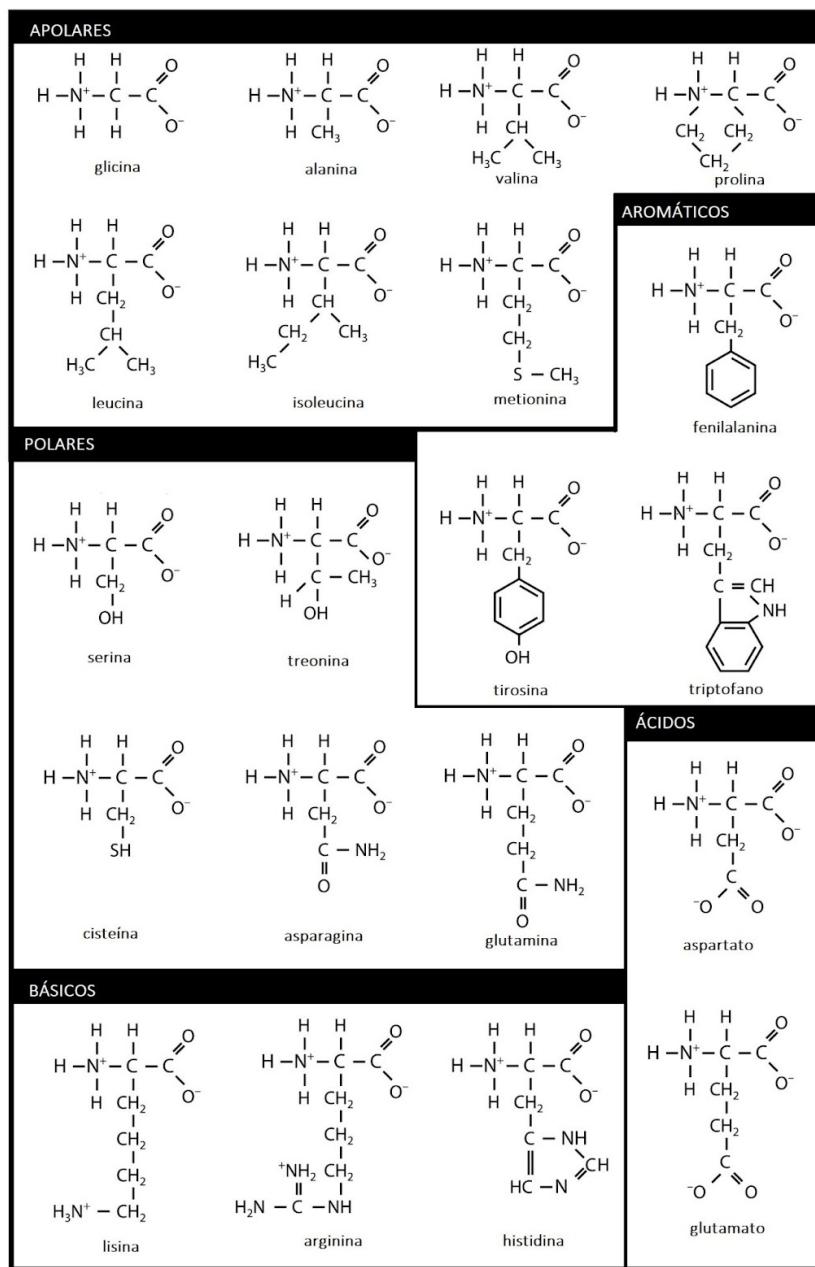


Figura 11.2: Estrutura dos aminoácidos. Fonte: próprio autor.

## 11.2 Estrutura proteica

Os peptídeos e proteínas são formados por meio de ligações peptídicas entre o grupo carboxílico de um aminoácido e o grupo amino do aminoácido seguinte,

com liberação de uma molécula de água. A ligação peptídica (OC–NH) (Figura 11.3) tem caráter intermediário entre uma ligação simples e uma dupla, sendo uma ligação planar e rígida. Já as ligações entre  $C\alpha$  e NH ou entre  $C\alpha$  e CO são ligações simples e podem fazer rotações. Os ângulos de rotação dessas ligações são denominados phi ( $\phi$ ) para a ligação  $C\alpha$ , e psi ( $\psi$ ) para a ligação  $C\alpha$ -CO [5]. Nas proteínas, os valores de  $\phi$  e  $\psi$  são limitados pela repulsão estérica entre as cadeias laterais (R) dos resíduos de aminoácidos.

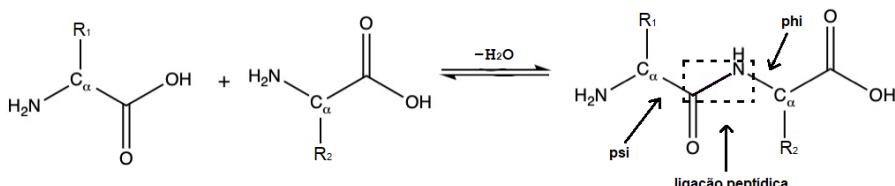


Figura 11.3: A ligação peptídica e os ângulos diedros phi ( $\phi$ ) e psi ( $\psi$ ). Fonte: Próprio Autor.

A estrutura proteica é de fundamental importância para a sua função, e é classificada de forma hierárquica em quatro níveis: estrutura primária, secundária, terciária e quaternária [6]. A **estrutura primária** corresponde à **sequência** dos resíduos de aminoácidos da proteína, unidos pela ligação peptídica, além das pontes de dissulfeto que podem ser formadas pelas cisteínas. A **estrutura secundária** corresponde ao arranjo espacial dos átomos adjacentes na **cadeia principal** da proteína, e é definida pela ligação peptídica e pelos ângulos diedros  $\phi$  e  $\psi$ . Certos padrões regulares de ângulos são observados nas estruturas de proteínas, gerando estruturas secundárias comuns, como a **alfa hélice**, as **folhas beta** e as **voltas beta**. Já a **estrutura terciária** corresponde ao arranjo tridimensional dos átomos da proteína devido ao seu enovelamento ou dobramento, incluindo também as interações das cadeias laterais, mesmo entre resíduos de aminoácidos que estão distantes na sequência proteica, mas próximos no espaço. Quando a proteína apresenta duas ou mais cadeias enoveladas distintas, que interagem entre si formando complexos, a **estrutura** é denominada **quaternária**.

Conforme a sua estrutura, as proteínas podem ser classificadas ainda em [7]: **proteínas globulares**, quando possuem cadeias bem enoveladas, contendo diversos tipos de estruturas secundárias, sendo assim adaptadas a diversas

funções, incluindo proteínas solúveis; **proteínas fibrosas**, quando possuem cadeias arranjadas em filamentos ou folhas, contendo geralmente um único tipo de estrutura secundária, estando adaptadas a funções de estruturação de células e tecidos; **proteínas intrinsecamente desordenadas**, que não possuem estrutura tridimensional definida; e as **proteínas de membrana**, que contém estruturas específicas com aminoácidos hidrofóbicos expostos para interação com as membranas celulares.

As proteínas enoveladas geralmente apresentam padrões de enovelamento identificáveis [6]. Quando duas ou mais estruturas secundárias são conectadas na forma de um padrão estrutural identificável, forma-se uma unidade estrutural denominada **motivo** (ex.: alça  $\beta$ - $\alpha$ - $\beta$  e barril  $\beta$ ). Já um **domínio** é uma região da proteína independentemente estável, ou seja, que conserva sua estrutura terciária, formada pela combinação de estruturas secundárias ou motivos, e geralmente tem uma função específica. Proteínas que estão relacionadas evolutivamente e pertencem a uma mesma **família** apresentam domínios semelhantes.

Além da cadeia polipeptídica, outros elementos podem estar presentes na estrutura proteica, como **modificações pós-traducionais**, que são adições de grupos químicos ou moléculas em resíduos de aminoácidos específicos após a tradução proteica (ex.: sítios de glicosilação, fosforilação, etc.) [8]. Adicionalmente, a ação de **cofatores** como **grupos prostéticos e coenzimas** [9], que são componentes não proteicos ligados à cadeia polipeptídica, podem ser necessários para a função de proteínas (ex.: heme, íons, NADH). As proteínas exercem sua função via alterações em sua dinâmica conformacional, provocada pela interação com outras moléculas, sejam elas receptores, ligantes, substratos, etc. A região de interação da proteína com seu ligante é denominada **sítio de ligação**, que em enzimas também é denominado **sítio ativo ou catalítico**. Algumas proteínas podem ter sua função modulada por moléculas que se ligam em regiões diferentes do sítio ativo, denominadas **sítios alostéricos**.

### **11.3 Bases de dados usadas em biologia estrutural**

Por fim, diversas bases de dados e ferramentas de bioinformática são frutos da contribuição de estudos na área de biologia estrutural e têm favorecido o avanço de estudos no entendimento da estrutura e função de proteínas. A identificação da sequência e identidade de proteínas, além da obtenção de informações relevantes a respeito de sua estrutura e função, podem ser obtidas com ferramentas como o BLAST [10] e bases de dados como o Uniprot [11] e o PDB [12]. O portal Expasy [13] reúne um conjunto de recursos para a predição de características físico-químicas e análises de sequências. A predição de estruturas secundárias também pode ser feita com ferramentas como o PSIPRED [14] e o DeepTMHMM [15]. Nos últimos anos, a predição de estruturas terciárias e quaternárias através da inteligência artificial (AlphaFold) [16] tem sido amplamente aplicada, e a visualização destas estruturas pode ser realizada por meio de programas como o PyMOL (Schrödinger, LLC) [17] e ChimeraX [18].

O objetivo deste artigo foi apresentar uma breve introdução à biologia estrutural de proteínas e outras moléculas. Nos próximos artigos, apresentaremos um roteiro para a utilização dessas bases e ferramentas bioinformáticas para a extração de informações a partir de sequências ou estruturas proteicas. Recomenda-se o senso crítico no uso das ferramentas, e sua combinação com a validação experimental sempre que possível.

#### **Nota de transparência 11.1**

Este material foi originalmente produzido para um minicurso ministrado durante o Curso de Inverno em Bioinformática da UFMG, realizado em 4 de Julho de 2023, na Universidade Federal de Minas Gerais, Belo Horizonte, Brasil.

#### **Saiba mais 11.1**

Este artigo está disponível em <https://bioinfo.com.br/introducao-a-biologia-estrutural-de-proteinas/>

## 11.4 Referências

- [1] Nelson, D. L.; Cox, M. M. Princípios de bioquímica de Lehninger. 6. Ed. Porto Alegre: Artmed, 2014.
- [2] Fujii, N.; Takata, T.; Fujii, N.; Aki, K.; Sakaue, H. D-Amino Acids in Protein: The Mirror of Life as a Molecular Index of Aging. *Biochimica et Biophysica Acta (BBA) – Proteins and Proteomics*, v. 1866, n. 7, p. 840–847, jul. 2018.
- [3] IUPAC-IUB Joint Commission on Biochemical Nomenclature (JCBN). Nomenclature and Symbolism for Amino Acids and Peptides. Recommendations 1983. *European Journal of Biochemistry*, v. 138, n. 1, p. 9–37, jan. 1984.
- [4] Wu, G. Amino Acids: Metabolism, Functions, and Nutrition. *Amino Acids*, v. 37, n. 1, p. 1–17, maio 2009.
- [5] Ramachandran, G. N.; Venkatachalam, C. M.; Krimm, S. Stereochemical Criteria for Polypeptide and Protein Chain Conformations. *Biophysical Journal*, v. 6, n. 6, p. 849–872, nov. 1966.
- [6] Sun, P. D.; Foster, C. E.; Boyington, J. C. Overview of Protein Structural and Functional Folds. *Current Protocols in Protein Science*, v. 35, n. 1, fev. 2004.
- [7] Andreeva, A.; Kulesha, E.; Gough, J.; Murzin, A. G. The SCOP Database in 2020: Expanded Classification of Representative Family and Superfamily Domains of Known Protein Structures. *Nucleic Acids Research*, v. 48, n. D1, p. D376–D382, 8 jan. 2020.
- [8] Ramazi, S.; Zahiri, J. Post-Translational Modifications in Proteins: Resources, Tools and Prediction Methods. *Database*, v. 2021, p. baab012, 7 abr. 2021.
- [9] De Bolster, M. W. G. Glossary of Terms Used in Bioinorganic Chemistry (IUPAC Recommendations 1997). *Pure and Applied Chemistry*, v. 69, n. 6, p. 1251–1304, 1 jan. 1997.
- [10] Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic Local Alignment Search Tool. *Journal of Molecular Biology*, v. 215, n. 3, p. 403–410, out. 1990.
- [11] The Uniprot Consortium; Bateman, A.; Martin, M.-J.; Orchard, S.; Magrane, M.; Ahmad, S. *et al.* UniProt: The Universal Protein Knowledgebase in 2023. *Nucleic Acids Research*, v. 51, n. D1, p. D523–D531, 6 jan. 2023.
- [12] Burley, S. K.; Berman, H. M.; Bhikadiya, C.; Bi, C.; Chen, L.; Di Costanzo, L. *et al.* RCSB Protein Data Bank: Biological Macromolecular Structures Enabling Research and Education in Fundamental Biology, Biomedicine, Biotechnology and Energy. *Nucleic Acids Research*, v. 47, n. D1, p. D464–D474, 8 jan. 2019.

- [13] Duvaud, S.; Gabella, C.; Lisacek, F.; Stockinger, H.; Ioannidis, V.; Durinx, C. Expasy, the Swiss Bioinformatics Resource Portal, as Designed by Its Users. *Nucleic Acids Research*, v. 49, n. W1, p. W216–W227, 2 jul. 2021.
- [14] Buchan, D. W. A.; Jones, D. T. The PSIPRED Protein Analysis Workbench: 20 Years On. *Nucleic Acids Research*, v. 47, n. W1, p. W402–W407, 2 jul. 2019.
- [15] Hallgren, J.; Tsirigos, K. D.; Pedersen, M. D.; Armenteros, J. J. A.; Marcatili, P.; Nielsen, H.; Krogh, A.; Winther, O. DeepTMHMM predicts alpha and beta transmembrane proteins using deep neural networks. *Bioinformatics*, 10 abr. 2022.
- [16] Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O. *et al.* Highly Accurate Protein Structure Prediction with AlphaFold. *Nature*, v. 596, n. 7873, p. 583–589, ago. 2021.
- [17] Schrodinger, LLC. 2010. The PyMOL Molecular Graphics System.
- [18] Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Meng, E. C.; Couch, G. S.; Croll, T. I.; Morris, J. H.; Ferrin, T. E. UCSF ChimeraX: Structure Visualization for Researchers, Educators, and Developers. *Protein Science*, v. 30, n. 1, p. 70–82, jan. 2021.

# 12 EXTRAÇÃO DE INFORMAÇÕES DE SEQUÊNCIAS E ESTRUTURAS DE PROTEÍNAS

Autores 12.1

Rafael Pereira Lemos , Paulo Henrique dos Santos , Aline Rocha 

Revisão: Izabela Mamede , Marcos Antonio Nobrega de Sousa , Wylerson Nogueira 

Cite este artigo 12.1

Lemos, R; Santos, PH; Rocha, A. **Extração de Informações de Sequências e Estruturas de Proteínas.** BIOINFO. ISSN: 2764-8273. Vol. 3. p.12 (2023). doi: 10.51780/bioinfo-03-12

### Resumo 12.1

ESTE manuscrito apresenta uma simulação do processo computacional para obtenção de informações estruturais e físico-químicas de proteínas, partindo apenas de uma sequência proteica. Todas as ferramentas utilizadas neste roteiro estão disponíveis de forma online e gratuita (com a exceção do programa ChimeraX, que deve ser baixado, mas possui licença acadêmica), necessitando assim apenas de conexão com a internet para serem utilizadas.

Uma sequência modelo é apresentada ao longo do roteiro, e todas as análises são realizadas a partir dela. No entanto, os participantes são encorajados a testar e realizar o roteiro com seus próprios objetos de estudo ou interesse, além de explorar os demais parâmetros e funcionalidades disponíveis nas ferramentas que não são mencionados aqui. Em alguns momentos, também são apresentados materiais suplementares (tutoriais ou artigos de documentação) para auxiliar os alunos.

Neste artigo, abordaremos como métodos computacionais podem ser utilizados para a obtenção de informações sobre a identidade, sequência, estrutura e propriedades das proteínas.

## 12.1 Parte 1 – Identificação da Sequência

Para fins didáticos, utilizaremos a seguinte sequência proteica para todas as atividades subsequentes (você pode copiar e colar o conteúdo da caixa abaixo). Recomendamos que você salve a sequência em um bloco de notas, para facilitar a execução das atividades.

MLKRYLVLSVATAAFSLPSLVNAQQNILSVHILNQQTGK  
PAADVTVTLEKKADNGWLQLNTAKTDKGRIKALWPEQTA  
TTGDYRVVFKTGDYFKQNLESFFPEIPVEFHINKVNEHY  
HVPLLLSQYGYSTYRGS

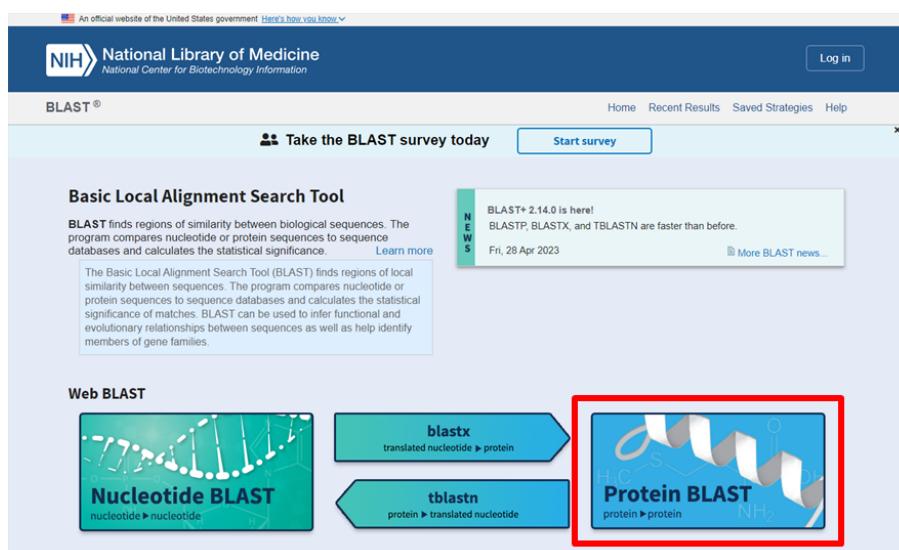
Suponhamos que alguém te entregou a sequência acima para analisar, mas você ainda não tem nem ideia de qual é a proteína codificada por essa sequência. Como podemos então obter essa informação?

O programa *BLAST* [1,2] é capaz de identificar regiões de similaridade entre sequências biológicas (nucleotídeos e proteínas) fornecidas pelo usuário, compará-las com bancos de dados, e realizar o cálculo das significâncias estatísticas (neste caso, o cálculo da similaridade entre as sequências com a probabilidade de pareamento ao acaso, também chamado de *E-value*).

Como o nosso objetivo é identificar uma sequência proteica comparando-a com dados de outras sequências proteicas, utilizaremos a função *Protein BLAST*. Também é possível realizar análises de nucleotídeos para nucleotídeos (*Nucleotide BLAST*), nucleotídeo para proteína (*blastx*), e proteína para nucleotídeo (*tblastn*).

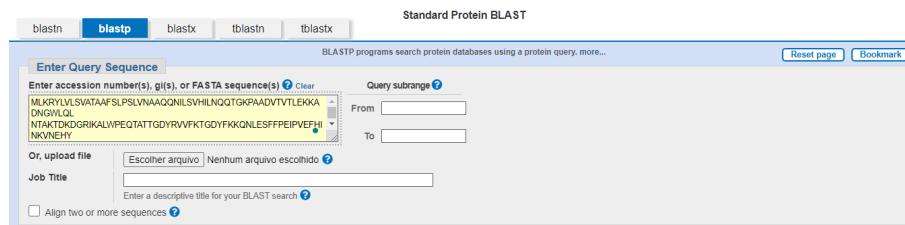
**Passo 1:** Acesse o link da suíte do *BLAST* – <https://blast.ncbi.nlm.nih.gov/Blast.cgi>

**Passo 2:** Selecione a ferramenta *Protein BLAST*



**Passo 3:** Cole a sequência na caixa localizada na parte superior da página (“*Enter Query Sequence*”). Também é possível escolher um arquivo com a sequência ou um número de acesso de outro banco de dados (como o Uniprot, por exemplo).

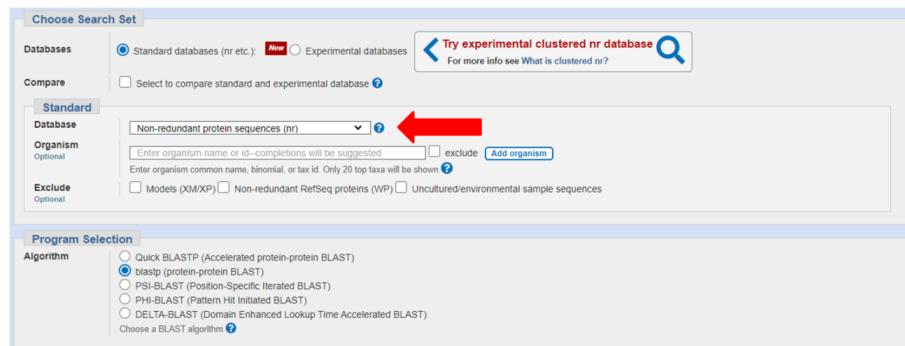
Múltiplas sequências simultâneas também são permitidas, desde que estejam em formato FASTA (<https://www.ncbi.nlm.nih.gov/genbank/fastaformat/>).



The screenshot shows the 'Standard Protein BLAST' search interface. In the 'Enter Query Sequence' section, there is a text input field containing a FASTA sequence: MURKYLVLVATAAFSLPSLVNAAQQNLISVHLNQQTGKPAADVTITLEKKAA... The interface includes tabs for 'blastn', 'blastp', 'tblastx', and 'tblastn'. There are also fields for 'Query subrange', 'Job Title', and a checkbox for 'Align two or more sequences'.

**Dica:** clique nos ícones de “?” para obter ajuda em relação às funcionalidades.

**Passo 4:** Na opção “Database”, dentro de “Choose Search Set”, escolhemos a opção de banco de dados de sequências de proteínas não-redundantes (nr), que deverá ser usada como padrão. Caso você queira buscar apenas por sequências que já tiveram suas estruturas resolvidas, poderá escolher a opção do banco de dados do *Protein Data Bank* (PDB), por exemplo.



The screenshot shows the 'Choose Search Set' interface. Under the 'Standard' tab, the 'Database' dropdown is set to 'Non-redundant protein sequences (nr)'. A red arrow points to this dropdown. Other options include 'Experimental databases' and 'Select to compare standard and experimental database'. Below the dropdown, there are fields for 'Organism' and 'Exclude' with various checkboxes. At the bottom, the 'Program Selection' section shows 'blastp (protein-protein BLAST)' selected under the 'Algorithm' tab.

**Passo 5:** Clique no botão “BLAST”, no fim da página, e aguarde a execução.

Após a execução do programa, a tela de resultados conterá as informações da sua busca na parte superior, e uma lista na parte inferior. Vamos analisar esta lista com mais calma. Caso tenha alguma dúvida, você também pode clicar nas páginas de ajuda, que contam com um tutorial de leitura

da página, assim como vídeos no YouTube (a *playlist* está disponível em <https://youtube.com/playlist?list=PL7dF9e2qSW0azL2xOKAtxDW7QI8UU4XZ6>).

**Passo 6:** Na aba de descrições, você pode conferir um sumário de todos os “*hits*” encontrados pelo *BLAST*. A primeira coluna representa a descrição da proteína encontrada, a segunda o nome da espécie associada, e as outras representam métricas de qualidade do programa.

Descriptions		Graphic Summary	Alignments	Taxonomy	Download	Select columns	Show	100	?
<b>Sequences producing significant alignments</b>									
<input type="checkbox"/> select all	0 sequences selected	GenPept	Graphics	Distance tree of results	Multiple alignment	MSA Viewer			
Description	Scientific Name	Max Score	Total Cover	E value	Per. Ident.	Acc Len	Accession		
<input type="checkbox"/> hydroxyisourate hydrolase [Enterobacteriaceae]	Enterobacteriaceae	285	285	100%	5e-97	100.00%	137	WP_000920120_1	
<input type="checkbox"/> hydroxyisourate hydrolase [Escherichia coli]	Escherichia coli	285	285	100%	7e-97	100.00%	141	MBS8952411_1	
<input type="checkbox"/> hydroxyisourate hydrolase [Escherichia coli]	Escherichia coli	285	285	100%	8e-97	99.27%	137	EFA4795043_1	
<input type="checkbox"/> hydroxyisourate hydrolase [Escherichia coli]	Escherichia coli	284	284	100%	1e-96	99.27%	137	QK43475_1	
<input type="checkbox"/> hydroxyisourate hydrolase [Escherichia coli]	Escherichia coli	284	284	100%	1e-96	99.27%	137	EFG1770639_1	
<input type="checkbox"/> hydroxyisourate hydrolase [Escherichia coli]	Escherichia coli	284	284	100%	1e-96	99.27%	137	WP_138810039_1	
<input type="checkbox"/> hydroxyisourate hydrolase [Escherichia coli]	Escherichia coli	284	284	100%	1e-96	99.27%	137	MBA8348281_1	
<input type="checkbox"/> hydroxyisourate hydrolase [Escherichia coli]	Escherichia coli	284	284	100%	2e-96	99.27%	137	MSL29023_1	
<input type="checkbox"/> hydroxyisourate hydrolase [Escherichia coli]	Escherichia coli	284	284	100%	2e-96	99.27%	137	WP_097315440_1	
<input type="checkbox"/> hydroxyisourate hydrolase [Escherichia coli]	Escherichia coli	284	284	100%	2e-96	99.27%	137	WP_112027313_1	
<input type="checkbox"/> hydroxyisourate hydrolase [Escherichia sp. 4_1_40B]	Escherichia sp. 4_1_40B	284	284	100%	2e-96	99.27%	137	WP_000920121_1	

É possível perceber que todos os “*hits*” estão ordenados pelo “*E-value*”, que é a principal métrica de confiança utilizada, e mede a probabilidade de um hit aleatório ocorrer com essa sequência. Ou seja, quanto menor o valor, maior a confiança de que a sua sequência inserida é realmente a informada.

As colunas de cobertura e porcentagem de identidade (à esquerda e direita da coluna de “*E-value*”, respectivamente), também trazem informações importantes. Suas descrições detalhadas podem ser visualizadas no link: <https://blast.ncbi.nlm.nih.gov/doc/blast-help/FAQ.html>.

**Passo 7:** Na aba de alinhamentos, é possível comparar, resíduo a resíduo, a sequência inserida como entrada e aquelas encontradas pela busca do BLAST. Por padrão, essas sequências também estão ordenadas de acordo com o “E-value”.

The screenshot shows a BLAST search results page with the following details:

- Descriptions**, **Graphic Summary**, **Alignments** (selected), **Taxonomy**
- Alignment view**: Pairwise
- Restore defaults**
- Download**
- 100 sequences selected**
- MULTISPECIES: hydroxyisourate hydrolase [Enterobacteriaceae]**
- Sequence ID: WP\_000920120.1 Length: 137 Number of Matches: 1**
- See 41 more title(s) ▾ See all Identical Proteins(IPG)**
- Range 1: 1 to 137 GenPept Graphics**
- Score: 285 bits(730) Expect: 5e-97 Compositional matrix adjust.: 137/137(100%) Identities: 137/137(100%) Positives: 137/137(100%) Gaps: 0/137(0%)**
- Query 1** (Sequence: MLKRYLVLVTAATAFSLPLSLVNAQONILSVHILNQQGTGPAADVTITLEKKADNGHLQL) vs **Sbjct 1** (Sequence: MLKRYLVLVTAATAFSLPLSLVNAQONILSVHILNQQGTGPAADVTITLEKKADNGHLQL) with 100% identity over 137 positions.
- Query 61** (Sequence: NTAKTDXDKGRKALNPQEATTGGDYRVFKFTGDYFKKQNLLESFFPEIPVEFHINKVNEHY) vs **Sbjct 61** (Sequence: NTAKTDXDKGRKALNPQEATTGGDYRVFKFTGDYFKKQNLLESFFPEIPVEFHINKVNEHY) with 100% identity over 137 positions.
- Query 121** (Sequence: HVPLLSSQYGYSTYRGs) vs **Sbjct 121** (Sequence: HVPLLSSQYGYSTYRGs) with 100% identity over 137 positions.

No exemplo acima, podemos perceber que a sequência com a maior confiança possui um “E-value” de 5e-97 (ou seja, bem baixo), 100% de identidade, 100% de alinhamento positivo, e 0% de gaps em relação à nossa sequência de entrada.

**Passo 8:** De posse da análise do BLAST, e selecionando o melhor resultado obtido, podemos assumir então que nossa sequência de entrada se trata da proteína Hidroxiisourato Hidrolase do gênero de Bactérias *Enterobacteriaceae*, ou mais especificamente, de *Escherichia coli* (como pode ser observado nos demais resultados da lista). O identificador da sequência (identificado pela seta vermelha), também é importante, como veremos a seguir.

**Passo 9:** Copie a sequência identificada na imagem acima (WP\_000920120.1).

Mas e agora, que proteína é essa? O que ela faz?

## 12.2 Parte 2 – Uniprot

O Uniprot é um banco de dados mantido pelo Instituto Europeu de Bioinformática (EBI), e que contém informações sobre sequências, estruturas, funções e anotações biológicas de proteínas.

Você pode pesquisar por diversas palavras chave, como o próprio nome ou sigla da proteína, algum organismo de preferência, ou por um número de acesso individual. O número de acesso principal do Uniprot é padronizado no formato de um prefixo “P”, seguido de seis caracteres (ex: P123456), mas diversos outros formatos também são aceitos (como o que acabamos de copiar do BLAST, no passo 9 da parte anterior).

**Passo 1:** Acesse o link do Uniprot (<https://www.uniprot.org/>). A tela inicial é mostrada abaixo.



**Passo 2:** Como acabamos de fazer uma análise no BLAST, nossa primeira tarefa agora é identificar e entender mais sobre os resultados obtidos. Para isso, vamos copiar o código do identificador da sequência com maior confiança (WP\_000920120.1), e colá-lo e pesquisá-lo no Uniprot.

## UniProtKB 6 results

A screenshot of the UniProtKB search results page. The title "UniProtKB 6 results" is at the top. Below it, there is a toolbar with links for BLAST, Align, Map IDs, Download, Add, View, Cards, Table, and Share. The search results are displayed in a grid format. The first result is for "P76341 · HIUH\_ECOLI", which is a 5-hydroxyisourate hydrolase from Escherichia coli (strain K12). It has an annotation score of 55 and is associated with #Hydrolase and #Purine metabolism. It has 3 3D structures and 9 reviewed publications. The second result is for "A0A370VB20 · A0A370VB20\_9ESCH", which is a 5-hydroxyisourate hydrolase from Escherichia marmotae. It has an annotation score of 25 and is associated with #Hydrolase and #Purine metabolism. It has 1 domain and 1 publication. On the right side of the results, there are buttons for "Feedback" and "Help".

Como você pode ter percebido, 6 resultados foram encontrados, mas o primeiro é diferente dos demais por algumas razões:

- O primeiro resultado possui um código de acesso padrão do Uniprot (P76341), ou seja, esta proteína foi revisada e anotada no SwissProt. Apenas sequências de alta qualidade e confiáveis possuem essa característica.
- O score de anotação em relação aos demais é superior (5/5), o que demonstra alta evidência experimental dessa proteína. Mais informações sobre os scores de anotação do Uniprot podem ser encontrados aqui: <https://www.ebi.ac.uk/training/online/courses/uniprot-exploring-protein-sequence-and-functional-info/annotation-score/>
- A quantidade de informações sobre estruturas 3D e publicações a respeito dessa proteína é superior às demais, evidenciando a maior quantidade de dados biológicos disponíveis.

**Passo 3:** Clique no primeiro resultado, e você será redirecionado para a página completa dessa proteína, contendo todas as informações disponíveis no UniProt sobre a mesma.

The screenshot shows the UniProt protein details page for P76341 · HIUH\_ECOLI. At the top, there's a navigation bar with links for BLAST, Align, Peptide search, ID mapping, SPARQL, UniProtKB, Advanced, List, Search, Help, and a feedback button. The main content area has a sidebar on the left with categories like Function, Names & Taxonomy, Subcellular Location, Phenotypes & Variants, PTM/Processing, Expression, Interaction, Structure, Family & Domains, Sequence, and Similar Proteins. The central part of the page displays the protein's name (P76341 · HIUH\_ECOLI), its function (5-hydroxisourate hydrolase), gene (hiuH), status (UniProtKB reviewed (Swiss-Prot)), organism (Escherichia coli (strain K12)), amino acids (137), protein existence (Evidence at protein level), and annotation score (5/5). Below this, there are tabs for Entry, Variant viewer, Feature viewer, Publications, External links, and History. The Publications tab is active, showing one publication related to catalyzing the hydrolysis of 5-hydroxisourate to 2-oxo-4-hydroxy-4-carboxy-5-ureidoimidazoline (OHCU). The Feature viewer tab shows catalytic activity: 5-hydroxisourate + H<sub>2</sub>O = 5-hydroxy-2-oxo-4-ureido-2,5-dihydro-1H-imidazole-5-carboxylate + H<sup>+</sup>. The bottom right corner has a help button.

Aqui você pode encontrar informações sobre a função, mecanismo enzimático (apenas para enzimas), sítios de ligação e de modificações pós-traducionais

(fosforilações, glicosilações, etc.), taxonomia, localização celular, domínios e outras anotações biológicas. Voltaremos aqui algumas vezes ao longo deste artigo, então não feche a aba. Enquanto isso, fique à vontade para explorar!

Agora que sabemos que nossa sequência desconhecida pertence a uma enzima da via de degradação do ácido úrico em *E. coli*, quais outras informações podemos obter?

### 12.3 Parte 3 – Características Físico-Químicas

Além do conhecimento da função e outras propriedades que obtivemos a partir do Uniprot, informações físico-químicas das proteínas também são extremamente importantes para uma variedade de aplicações na Bioinformática.

Para esta etapa, utilizaremos a plataforma ExPASy (*Expert Protein Analysis System*; disponível em: <https://www.expasy.org/>)<sup>[3]</sup>, mantida pelo Instituto Suíço de Bioinformática (SIB). Essa plataforma disponibiliza gratuitamente diversas ferramentas para ajudar na análise e caracterização de proteínas.



Como não conseguiremos cobrir todas as ferramentas disponíveis (e nem todas são aplicáveis a todos os casos e perguntas), focaremos apenas em uma, embora todas tendam a manter o mesmo padrão de uso. Fique à vontade para explorar as outras!

O ProtParam <sup>[4]</sup> é uma das ferramentas mais simples disponíveis no ExPASy, mas que fornece informações físico-químicas cruciais para o estudo de proteínas específicas. A documentação da ferramenta pode ser acessada aqui: <https://web.expasy.org/protparam/protparam-doc.html>.

**Passo 1:** Acesse o link do ProtParam (<https://web.expasy.org/protparam/>).

Please note that you may only fill out **one** of the following fields at a time.

Enter a Swiss-Prot/TrEMBL accession number (AC) (for example **P05130**) or a sequence identifier (ID) (for example **KPC1\_DROME**):  
P76341

Or you can paste your own amino acid sequence (in one-letter code) in the box below:  
MILKRYLVLSVATAAFSLPSLVNIAAQNLILSVHILNQQTGPAAADVTVTLEKKADWKGILQL  
NTIATKTDKDGRIKALIPIEQATTGDRYRVFKTGDFKKQNLESFFPEIPVEFHINKVNEHY  
HVPLLLSQQGYSTYRGs

RESET Compute parameters

**Passo 2:** Insira o código Uniprot obtido no passo anterior (P76341) OU a sequência da proteína nas caixas iniciais, e clique em “compute parameters”.

**Nota:** caso você insira o código Uniprot, uma tela aparecerá para que você selecione qual porção da sequência da proteína quer analisar. É possível perceber que a proteína madura compreende apenas os resíduos de 24 a 137, com os 23 primeiros correspondendo a um peptídeo sinal.

Selection of endpoints on the sequence

**HIUH\_ECOLI** (P76341)

5-hydroxyisourate hydrolase precursor (EC 3.5.2.17) (HIU hydrolase) (HIUHase) (Transthyretin-like protein) (TLP) (Transthyretin-related protein) (TRP)  
Escherichia coli (strain K12)

Please select one of the following features by clicking on a pair of endpoints, and the computation will be carried out for the corresponding sequence fragment. By default, the complete sequence is used.

**Note:** Only the features corresponding to subsequences of at least 5 residues are highlighted.

FT	SIGNAL	1-23
FT	CHAIN	<b>24-137 5-hydroxyisourate hydrolase</b>
FT	STRAND	29-35
FT	TURN	36-39
FT	STRAND	46-52
FT	STRAND	54-64
FT	STRAND	69-72
FT	STRAND	82-89
FT	HELIX	91-97
FT	STRAND	107-115
FT	STRAND	120-127
FT	STRAND	130-134

Por curiosidade, podemos ir no Uniprot conferir a informação da localização do peptídeo sinal. Na aba lateral “PTM/Processing”, o primeiro resultado indica

justamente a divisão da sequência da proteína em uma parte sinal (resíduos 1-23) e a cadeia da proteína em si (resíduos 24-137). Obviamente, nem todas as proteínas já terão todas suas informações anotadas no Uniprot, por isso a importância de também se realizarem análises teóricas como as que estamos vendo aqui.

## PTM/Processing<sup>1</sup>

### Features

Showing features for signal<sup>1</sup>, chain<sup>1</sup>.



Por agora, apenas clique em “SUBMIT” no final da página do ProtParam, para que a análise seja feita com a sequência completa de 137 resíduos (a mesma que utilizamos para fazer as análises do BLAST).

Or, if you wish to select a different sequence fragment (at least 5 amino acids long), you can enter the desired endpoints on the sequence here (by default, the computation will be carried out for the complete sequence).

N-terminal:   
C-terminal:

The sequence HIUH\_ECOLI consists of 137 amino acids.

A janela de resultados do ProtParam mostra informações teóricas como o peso molecular da proteína, ponto isoelétrico (pI), composição atômica e de aminoácidos. Além disso, são calculados o coeficiente de extinção molar, meia-vida da proteína, e índices de estabilidade e hidropaticidade, parâmetros bastante úteis em ensaios experimentais.

Number of amino acids:	137	Total number of negatively charged residues (Asp + Glu):	12
Molecular weight:	15459.61	Total number of positively charged residues (Arg + Lys):	15
Theoretical pI:	9.10	Atomic composition:	
Amino acid composition:	<a href="#">CSV format</a>	Carbon	C 704
Ala (A)	11	Hydrogen	H 1094
Arg (R)	4	Nitrogen	N 186
Asn (N)	8	Oxygen	O 204
Asp (D)	6	Sulfur	S 1
Cys (C)	0	Formula:	$C_{704}H_{1094}N_{186}O_{204}S_1$
Gln (Q)	8	Total number of atoms:	2189
Glu (E)	6	Extinction coefficients:	
Gly (G)	7	Extinction coefficients are in units of $M^{-1} cm^{-1}$ , at 280 nm measured in water.	
His (H)	4	Ext. coefficient	21430
Ile (I)	5	Abs 0.1% ( $\approx 1 g/l$ )	1.386
Leu (L)	15	Estimated half-life:	
Lys (K)	11	The N-terminal of the sequence considered is M (Met).	
Met (M)	1	The estimated half-life is: 30 hours (mammalian reticulocytes, in vitro).	
Phe (F)	6	>20 hours (yeast, in vivo).	
Pro (P)	6	>10 hours (Escherichia coli, in vivo).	
Ser (S)	8	Instability index:	
Thr (T)	11	The instability index (II) is computed to be 38.98.	
Trp (W)	2	This classifies the protein as stable.	
Tyr (Y)	7		
Val (V)	11	Aliphatic index:	88.25
Pyl (O)	0	Grand average of hydropathicity (GRAVY):	-0.328
Sec (U)	0		

Além de parâmetros físico-químicos, que podem ser obtidos diretamente a partir da sequência, também é possível predizer de forma prática a estrutura secundária de uma proteína apenas a partir dessa informação.

## 12.4 Parte 4 – Predição de Estruturas Secundárias

A análise de estruturas secundárias serve como a base do entendimento estrutural de proteínas, e pode fornecer informações importantes apenas com base em uma sequência. Normalmente, estruturas secundárias são divididas em -hélices, folhas- e espirais aleatórias [5].

Existem diversas ferramentas para predição computacional de estruturas secundárias, e utilizaremos aqui o servidor *web PSIPRED Workbench* [6,7].

**Passo 1:** Acesse o link do PSIPRED (<http://bioinf.cs.ucl.ac.uk/psipred/>)

O PSIPRED fornece uma gama de análises disponíveis, utilizando apenas uma sequência proteica como entrada. Algumas delas são a predição de estruturas secundárias, a predição de desordem estrutural, análises de contatos, reconhecimento de padrões de dobras e de domínios estruturais. Pode-se obter

mais detalhes sobre cada análise passando o cursor em cima de cada opção, ou acessando o tutorial da própria ferramenta ([http://bioinfadmin.cs.ucl.ac.uk/UCL-CS\\_Bioinformatics\\_Server\\_Tutorial.html](http://bioinfadmin.cs.ucl.ac.uk/UCL-CS_Bioinformatics_Server_Tutorial.html)).

Como algumas das análises demandam algum tempo e/ou dependem de perguntas biológicas mais específicas, realizaremos neste tutorial apenas a predição de estruturas secundárias, predição de hélices transmembrana e de desordem estrutural (essas duas últimas análises serão mencionadas em mais detalhes na próxima parte). Fique à vontade para explorar as outras opções!

**Passo 2:** Selecione a opção do PSIPRED 4.0, para predição de estruturas secundárias, MEMSAT-SVM para predição de hélices transmembrana, e DISOPRED3 para predição de desordens estruturais.

Data Input

Select input data type

Sequence Data  PDB Structure Data

Choose prediction methods (hover for short description)

Popular Analyses

PSIPRED 4.0 (Predict Secondary Structure)  DISOPRED3 (Disopred Prediction)  
 MEMSAT-SVM (Membrane Helix Prediction)  pGenTHREADER (Profile Based Fold Recognition)

Contact Analysis

DeepMetaPSICOV 1.0 (Structural Contact Prediction)  MEMPACK (TM Topology and Helix Packing)

Fold Recognition

GenTHREADER (Rapid Fold Recognition)  pDomTHREADER (Protein Domain Fold Recognition)

Structure Modelling

Bioserf 2.0 (Automated Homology Modelling)  Domserf 2.1 (Automated Domain Homology Modelling)  
 DMPfold 1.0 Fast Mode (Protein Structure Prediction)

Single Sequence Prediction

S4Pred 1.2 (Single Sequence SS prediction)

**Passo 3:** Na aba de “*Submission details*”, insira a sequência da proteína que estamos trabalhando, um nome para a análise (de preferência algum que torne mais fácil você relembrar o que foi feito), e um email para que os resultados sejam enviados (esta parte é opcional mas recomendada caso você faça análises mais demoradas).

**Submission details**

**Protein Sequence**

```
MLKRYLVLSVATAAFSLPSLVNAQQNILSVHILNQQTGKPAADVTVTLEKKADNGWLQL
NTAKTDKGRIKALWPEQTATTGDYRVVFKTGDYFKQNLESFFPEIPVEFHINKVNEHY
HVPLLLSQYGYSTYRGS
```

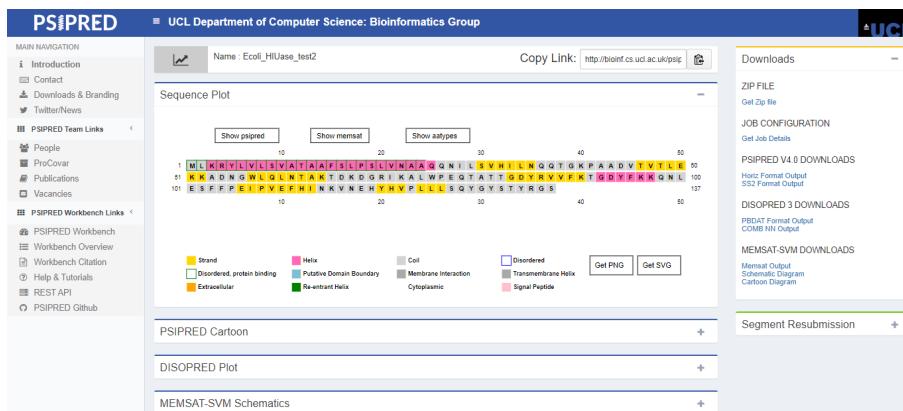
[Help...](#)  
If you wish to test these services follow this link to retrieve a test fasta sequence.

**Job name**  
Ecoli\_HIUase

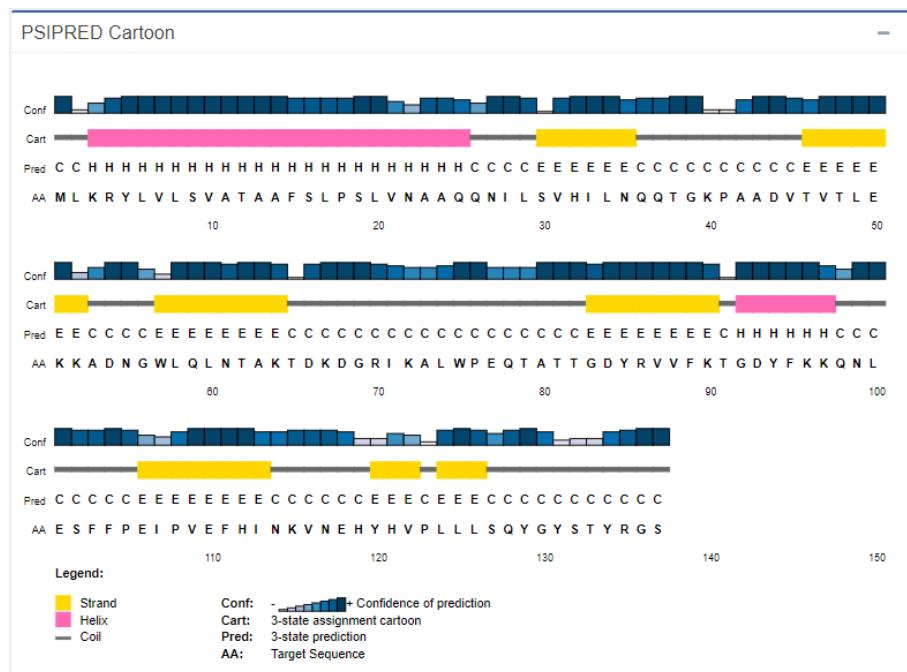
**Email (optional)**  
Email (optional)

**Reset** **Submit**

**Passo 4:** Após finalizadas as análises, os resultados serão mostrados diretamente na ferramenta. Na primeira caixa (“*sequence plot*”), você pode selecionar para que os dados de uma determinada análise sejam mostrados. No exemplo abaixo, é mostrada a visualização dos resultados do próprio PSIPRED na sequência.

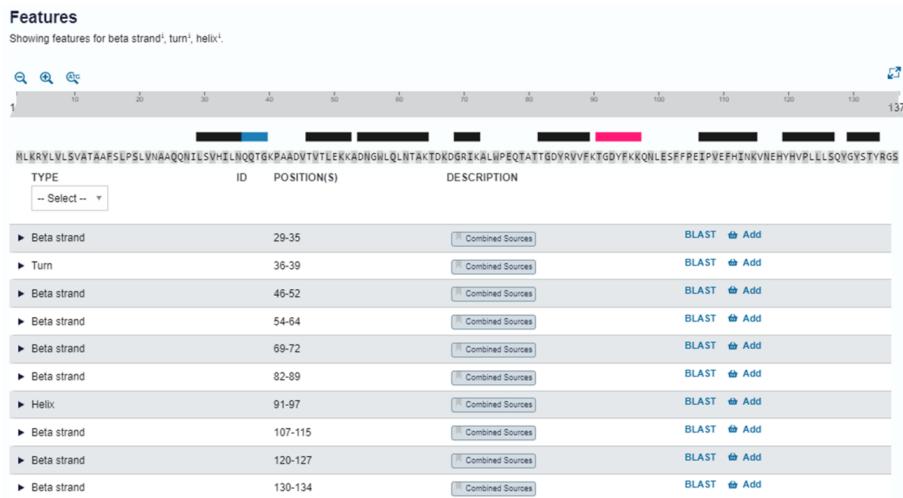


**Passo 5:** Expandindo cada uma das caixas inferiores, você pode visualizar os resultados individuais de cada análise. Por ora, iremos analisar os resultados da predição de estruturas secundárias em nossa proteína.



Como pode ser percebido, aparentemente temos duas -hélices (caixas rosas) e sete folhas- (caixas amarelas). O restante da estrutura é composto por espirais aleatórias (linha cinza). As barras azuis em cima das caixas indicam a confiança de predição aminoácido-específica (quanto mais alta e mais escura a barra, maior a confiança).

**Passo 6:** Já que nossa proteína é bem anotada no Uniprot, podemos comparar os dados obtidos de forma teórica pelo PSIPRED com aqueles já depositados e revisados. Na aba lateral “*Structure*” do Uniprot, na seção “*Features*”, podemos analisar as estruturas secundárias obtidas experimentalmente para essa proteína, além de identificar a origem dessas informações (clicando nos botões “*combined sources*” em cada entrada).

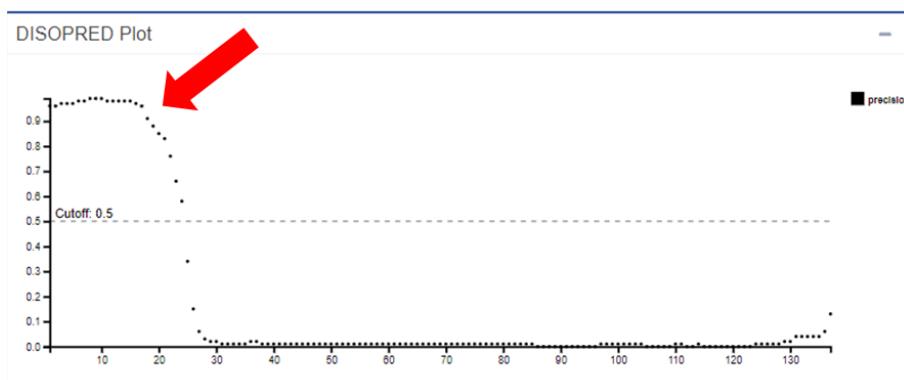


## 12.5 Parte 5 – Identificação de Desordem Estrutural e Domínios Transmembrana

Além de informações como a estrutura secundária, também é possível predizer computacionalmente potenciais regiões desordenadas ou transmembranares apenas a partir de uma sequência proteica.

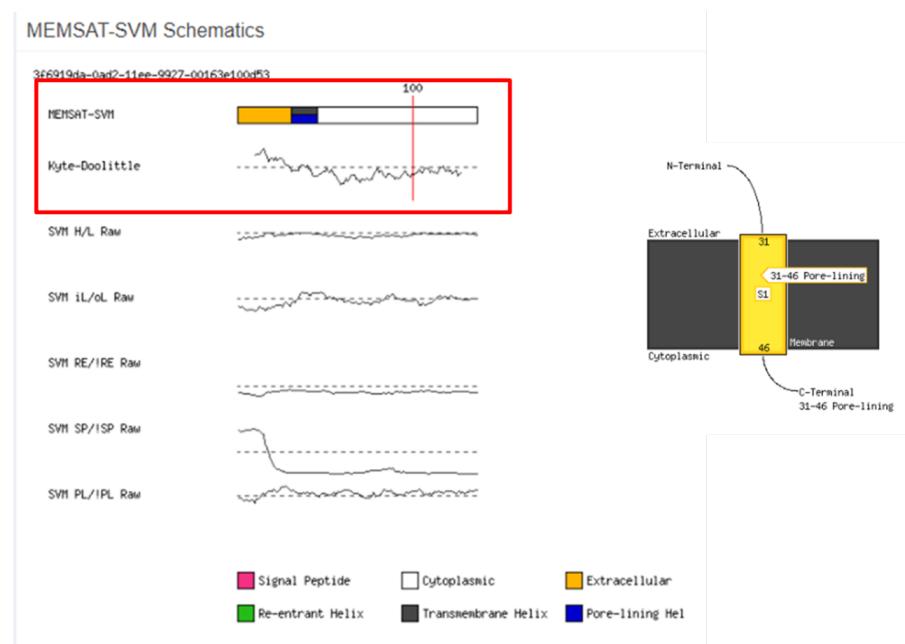
Nesta parte, usaremos os dois outros resultados obtidos anteriormente pelo PSIPRED em conjunto com a ferramenta DeepTMHMM [8], que utiliza *deep learning* para predizer regiões transmembranares.

**Passo 1:** No PSIPRED, abra a caixa do resultado do DISOPRED.



É possível identificar uma região com alta taxa de desordem na proteína (seta vermelha). Essa região pertence justamente ao peptídeo sinal presente nos primeiros resíduos da HIUase. Como peptídeos sinal normalmente são clivados após o endereçamento de uma proteína, essas regiões tendem a ser intrinsecamente desordenadas.

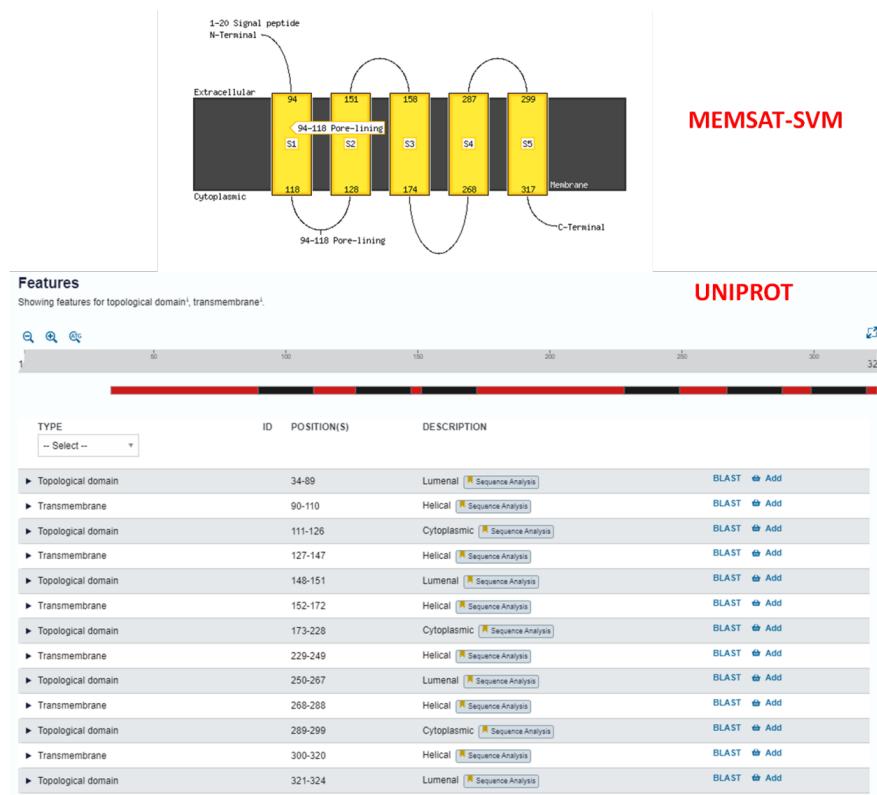
**Passo 2:** Abra a caixa de resultado do MEMSAT-SVM.



Para esta análise, é possível perceber que a ferramenta identificou os primeiros resíduos da sequência como extracelulares (caixa amarela), além de uma possível região transmembranar/poro entre os resíduos 31 e 46 (caixa preta e azul, e esquema à direita). O peptídeo sinal não foi identificado.

Se olharmos no Uniprot como anteriormente, não há nenhuma menção à regiões transmembranares nessa proteína, assim como regiões extracelulares (a proteína se localiza no periplasma celular, como é possível ver na seção “Subcellular Location” do Uniprot). Sendo assim, a predição desta proteína pela ferramenta MEMSAT-SVM não foi acurada.

**Nota:** analisando uma proteína verdadeiramente transmembranar (Proteína Transmembranar 165, de Homo sapiens – ID Uniprot: Q9HC07), os resultados são acurados. Para acessar a sequência dessa proteína, use o seguinte link: <https://rest.uniprot.org/uniprotkb/Q9HC07.fasta>



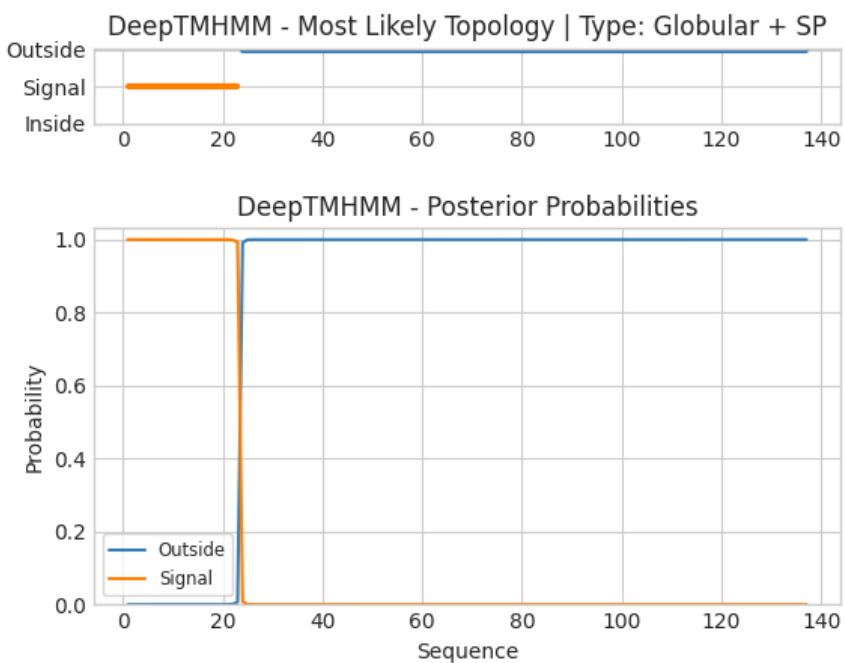
Como podemos então obter previsões mais acuradas sobre regiões transmembranares, tendo em mãos apenas a sequência da proteína?

**Passo 3:** Abra o link do DeepTMHMM (<https://dtu.biolib.com/DeepTMHMM>).

**Passo 4:** Insira a sequência da proteína a ser analisada, e clique em “Run”.

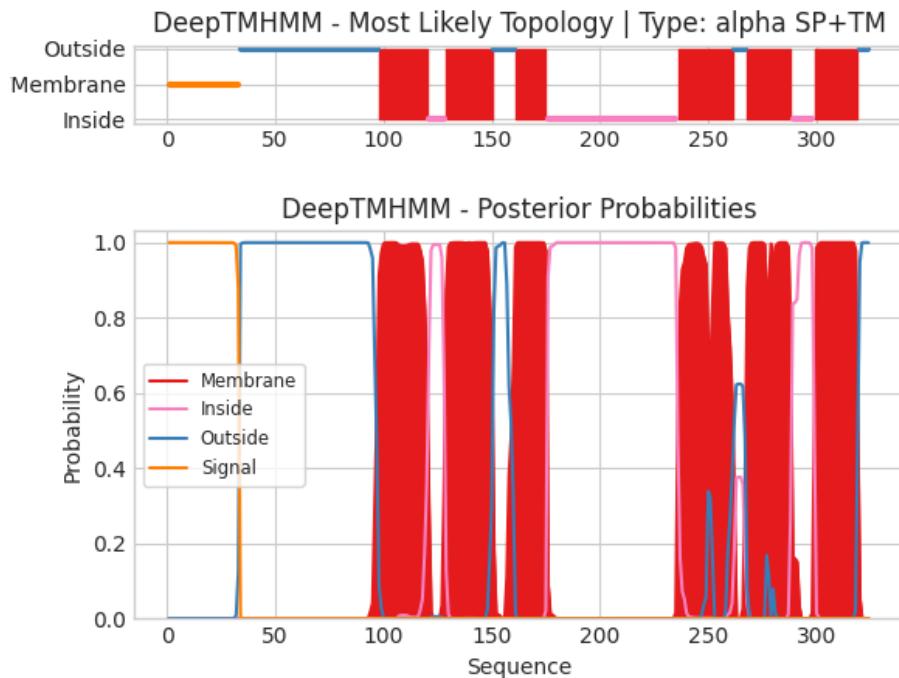
Screenshot of the DeepTMHMM web application interface. The main header includes a logo, search bar, and navigation links (Docs, Explore, Results, Build). Below the header, the URL shows "Technical University of Denmark > DeepTMHMM". The interface has three main sections: "Input" (highlighted with a red box), "Compute", and "Results". The "Input" section contains a text area for pasting protein sequences in FASTA format, a "Select File" button, and a "Clear File" button. A "Run" button is located below the input area. To the left of the input section, there is a sidebar with "Version" (1.0.24), "Developer" (Technical University of Denmark), "Cite Application" (with APA, ISO 690, BibTeX options), and "Run Application from Script" (Python, Terminal). The central content area is titled "DeepTMHMM" and contains release notes, a brief description of the model, and a summary of its performance.

**Passo 5:** Desta vez, os resultados se mostraram muito mais condizentes com os dados encontrados no Uniprot, evidenciando a necessidade da utilização de múltiplas ferramentas complementares para obtenção de dados computacionais, especialmente teóricos.



A linha laranja indica região de peptídeo sinal, e a azul indica uma região externa à uma membrana (ou seja, citoplasmática). A topologia da proteína também é predita como globular acompanhada de um peptídeo sinal (SP).

**Passo 6:** Podemos também realizar uma análise no DeepTMHMM com uma proteína verdadeiramente transmembranar (Q9HC07, a mesma utilizada anteriormente).



Desta vez, além das linhas laranja e azul (peptídeo sinal e região externa, respectivamente), também podemos identificar regiões internas à membrana (linha rosa), e transmembranares (manchas vermelhas). A topologia identificada para a proteína nesse caso é de -hélices transmembranares, além de um peptídeo sinal.

## 12.6 Parte 6 – Obtenção de Estruturas 3D

A obtenção de estruturas tridimensionais de proteínas pode ser feita através de diferentes metodologias experimentais ou computacionais [9,10]. Ao resolver as

estruturas, os pesquisadores depositam as coordenadas espaciais dos átomos em um banco de dados chamado PDB (*Protein Data Bank*). Lá você pode ter acesso a uma variedade de estruturas de proteínas depositadas pela comunidade científica ao seu dispor, para utilizá-las em suas análises.

Neste banco de dados é possível fazer a busca utilizando o código da estrutura, que possui como padrão 4 dígitos alfanuméricicos (ex: 2H1X, 3IWU, 7KCN, etc). Há também como filtrar pela função da proteína, como Phosphatase, HIUase, Glycosidase, etc. Após realizar a busca você vai receber diversas estruturas com várias informações.

**Passo 1:** No Uniprot que estamos trabalhando, na seção “*Structure*”, você pode encontrar todas as estruturas tridimensionais que já foram resolvidas para essa determinada proteína. Não se preocupe se nenhuma estrutura já tiver sido depositada no PDB. Falaremos do AlphaFold mais à frente!

The screenshot shows the UniProt entry page for P00953. In the left sidebar, under the 'Structure' category, the 'PDB' option is highlighted with a red box. Below the sidebar, there's a table of structures. The first row, which has 'PDB' selected in the 'SOURCE' dropdown, is highlighted with a yellow background. This row contains the identifier '2G2N', method 'X-ray', resolution '1.65 Å', chain 'A/B/C/D 24-137', and links to PDBe, RCSB-PDB, PDBj, and PDBsum, each with a 'Foldseek' link. A red arrow points to the 'SOURCE' dropdown.

SOURCE	IDENTIFIER	METHOD	RESOLUTION	CHAIN	POSITIONS	LINKS
PDB	2G2N	X-ray	1.65 Å	A/B/C/D 24-137		PDBe · RCSB-PDB · PDBj · PDBsum
PDB	2G2P	X-ray	2.10 Å	A/B/C/D 24-137		PDBe · RCSB-PDB · PDBj · PDBsum
PDB	2IGL	X-ray	1.80 Å	A/B/C/D 24-137		PDBe · RCSB-PDB · PDBj · PDBsum
AlphaFold	AF-P76341-F1	Predicted		1-137		AlphaFold

Procure pelo código do PDB da estrutura com melhor resolução. Voltaremos a trabalhar com esse código daqui a pouco.

**Passo 2:** Acesse o banco de dados do PDB (<https://www.rcsb.org/>).

**Passo 3:** Antes de pesquisarmos nossa proteína, vale conferir a parte de busca avançada que pode ser vista abaixo da região de busca.



**Passo 4:** Na busca avançada existe a possibilidade de especificar detalhes estruturais, experimentais e químicos desejados a fim de filtrar as estruturas desejadas. Outros filtros podem ser acessados também após realizar a busca.

**Passo 5:** Caso seja feita uma busca pela função, por exemplo, é interessante observar os filtros de métodos experimentais, organismo de origem, taxonomia e a resolução das estruturas. Este último é um parâmetro importante para a validação das estruturas. Então, façam o experimento de buscar por “HIUase” utilizando os filtros de “Refinements” para buscar estruturas de diferentes níveis de resolução, organismos, taxonomia, etc.

Structure Determination Methodology
<input checked="" type="checkbox"/> experimental (7)

Scientific Name of Source Organism
<input type="checkbox"/> Danio rerio (5)
<input type="checkbox"/> synthetic construct (1)
<input type="checkbox"/> unidentified (1)

Taxonomy
<input type="checkbox"/> Eukaryota (5)
<input type="checkbox"/> other sequences (1)
<input type="checkbox"/> unclassified sequences (1)

<b>Experimental Method</b>
<input type="checkbox"/> X-RAY DIFFRACTION (7)
<b>Polymer Entity Type</b>
<input type="checkbox"/> Protein (7)
<b>Refinement Resolution (Å)</b>
<input type="checkbox"/> 1.0 - 1.5 (1)
<input type="checkbox"/> 1.5 - 2.0 (5)
<input type="checkbox"/> 2.0 - 2.5 (1)
<b>Release Date</b>
<input type="checkbox"/> 2005 - 2009 (2)
<input type="checkbox"/> 2010 - 2014 (3)
<input type="checkbox"/> 2020 - 2024 (2)

Vale também ressaltar informações importantes para se avaliar nas estruturas, como a resolução em angstroms (quanto menor, melhor), assim como o organismo possuidor da proteína e os ligantes já descritos na literatura.

**3IWU**



Crystal structure of Y116T/I16A double mutant of 5-hydroxyisourate hydrolase  
Cendron, L., Ramazzina, I., Berni, R., Percudani, R., Zanotti, G.  
(2011) J Mol Biol **409**: 504-512

Released: 2010-09-01  
Method: X-RAY DIFFRACTION 2.3 Å  
Organisms: *Danio rerio*  
Macromolecule: 5-hydroxyisourate hydrolase (protein)

[3D View](#)

**3Q1E**



Crystal structure of Y116T/I16A double mutant of 5-hydroxyisourate hydrolase in complex with T4  
Cendron, L., Ramazzina, I., Percudani, R., Zanotti, G., Berni, R.  
(2011) J Mol Biol **409**: 504-512

Released: 2011-05-04  
Method: X-RAY DIFFRACTION 1.95 Å  
Organisms: *Danio rerio*  
Macromolecule: 5-hydroxyisourate hydrolase (protein)  
Unique Ligands: T44

[3D View](#)

← Resolução em Angstrons

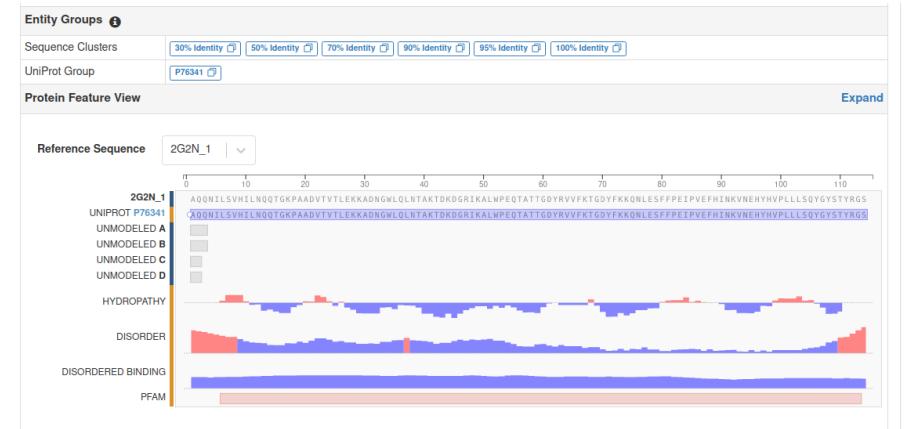
**Passo 6:** Voltando para o código PDB que obtivemos anteriormente no Uniprot (2G2N), iremos pesquisá-lo na busca normal do PDB.

**Passo 7:** Após abrir a estrutura 2G2N, já é possível visualizar diversas informações e validações importantes sobre a estrutura obtida experimentalmente, como os *outliers* do gráfico de Ramachandran (que será detalhado mais para frente).



Os parâmetros de validação são de extrema importância para a escolha de seu PDB em análises, sejam de dinâmica molecular, modelagem de estruturas ou como objeto de estudo em geral. Ou seja, é um campo que deve ser analisado com cuidado e sendo sempre comparado com outras estruturas similares.

**Passo 8:** Aqui, também é possível observar parâmetros como hidropaticidade dos resíduos, grau de desordem (mobilidade dos resíduos) e também informações da sequência no Uniprot. Lembre-se que já calculamos a maioria desses parâmetros anteriormente, mas de forma teórica. Fique à vontade para compará-los com as informações experimentais!



Para a obtenção de informações estruturais confiáveis, algumas metodologias experimentais necessitam do uso de pequenas moléculas, seja para estabilizar a estrutura ou para melhorar a resolução do experimento. Algumas perguntas também exigem a obtenção da proteína com seus ligantes biológicos, assim como inibidores. Nesta seção do PDB é possível verificar a presença das moléculas que foram obtidas juntamente à proteína depositada.

Small Molecules				
Ligands (2 Unique)				
ID	Chains	Name / Formula / InChI Key	2D Diagram	3D Interactions
SO4 <a href="#">Query on SO4</a>	O [auth B], T [auth C]	SULFATE ION O <sub>4</sub> S QAOWNCCODCNURD-UHFFFAOYSA-L		<a href="#">Ligand Interaction</a>
ZN <a href="#">Query on ZN</a>	E [auth A], F [auth A], G [auth A], H [auth A], I [auth A]	ZINC ION Zn PTFCDOFLOPIGGS-UHFFFAOYSA-N		<a href="#">Ligand Interaction</a>

## 12.7 Parte 7 – O Formato PDB

Todas as informações sobre a estrutura tridimensional de uma proteína estão disponíveis no formato “.pdb”. Esse é um arquivo de texto (ou seja, pode ser aberto em qualquer editor, como o Word, Notepad, etc.), contendo as coordenadas tridimensionais de cada átomo presente na estrutura. Além disso, há um cabeçalho com informações adicionais, como as condições experimentais para obtenção da estrutura, referências, etc.

**Passo 1:** Você pode visualizar o arquivo .pdb da proteína clicando em “*Display Files*” e depois em “*PDB Format*”. Também é possível baixar o arquivo clicando em “*Download Files*” e depois em “*PDB Format*”.

**Nota:** Para visualizar diretamente o .pdb da estrutura 2G2N, acesse em “<https://files.rcsb.org/view/2G2N.pdb>”.

**Passo 2:** Abra o arquivo .pdb da estrutura 2G2N em algum editor de texto (ou no próprio navegador). No cabeçalho é possível ver algumas informações como o organismo na qual a proteína foi expressa, o sistema de expressão utilizado, assim como a metodologia usada para resolver a estrutura.

```

HEADER      UNKNOWN FUNCTION          16-FEB-06    2G2N
TITLE       CRYSTAL STRUCTURE OF E.COLI TRANSTHYRETIN-RELATED PROTEIN WITH BOUND
TITLE       2 ZN
COMPND     MOL_ID: 1;
COMPND     2 MOLECULE: TRANSTHYRETIN-LIKE PROTEIN;
COMPND     3 CHAIN: A, B, C, D;
COMPND     4 SYNONYM: TRANSTHYRETIN-RELATED PROTEIN;
COMPND     5 ENGINEERED: YES
SOURCE      MOL_ID: 1;
SOURCE      2 ORGANISM_SCIENTIFIC: ESCHERICHIA COLI;
SOURCE      3 ORGANISM_TAXID: 562;
SOURCE      4 EXPRESSION_SYSTEM: ESCHERICHIA COLI;
SOURCE      5 EXPRESSION_SYSTEM_TAXID: 562;
SOURCE      6 EXPRESSION_SYSTEM_VECTOR_TYPE: PLASMID;
SOURCE      7 EXPRESSION_SYSTEM_PLASMID: PET24D
KEYWDS     TRANSTHYRETIN, TRANSTHYRETIN-RELATED PROTEIN, UNKNOWN FUNCTION
EXPDTA    X-RAY DIFFRACTION
  
```

Também há informações sobre o artigo publicado referente ao depósito da estrutura no PDB. É interessante lê-lo para obter informações adicionais sobre a proteína.

```
JRNL      AUTH  E.LUNDBERG,S.BACKSTROM,U.H.SAUER,A.E.SAUER-ERIKSSON
JRNL      TITL  THE TRANSTHYRETIN-RELATED PROTEIN: STRUCTURAL INVESTIGATION
JRNL      TITL  2 OF A NOVEL PROTEIN FAMILY
JRNL      REF   J.STRUCT.BIOL.          V. 155  445 2006
JRNL      REFN   ISSN 1047-8477
JRNL      PMID  16723258
JRNL      DOI   10.1016/J.JSB.2006.04.002
```

Logo após, na seção de “*Remarks*”, você pode obter todas as outras informações sobre a proteína (antes das coordenadas tridimensionais). Mais informações sobre os *remarks* podem ser obtidas aqui: <https://www.wwpdb.org/documentation/file-format-content/format33/remarks1.html>.

Um *remark* bastante importante são os “*Missing Residues*”. É bastante comum que alguns resíduos com grande mobilidade não sejam detectados em metodologias como Cristalografia de Raios X. Mas, note que neste caso se trata apenas dos primeiros resíduos da proteína que possuem graus de liberdade maiores e são mais desordenados. Em seguida também poderemos observar outros parâmetros estruturais.

```
REMARK 465 MISSING RESIDUES
REMARK 465 THE FOLLOWING RESIDUES WERE NOT LOCATED IN THE
REMARK 465 EXPERIMENT. (M=MODEL NUMBER; RES=RESIDUE NAME; C=CHAIN
REMARK 465 IDENTIFIER; SSSEQ=SEQUENCE NUMBER; I=INSERTION CODE.)
REMARK 465
REMARK 465 M RES C SSSEQI
REMARK 465     ALA A    1
REMARK 465     GLN A    2
REMARK 465     GLN A    3
REMARK 465     ALA B    1
REMARK 465     GLN B    2
REMARK 465     GLN B    3
REMARK 465     ALA C    1
REMARK 465     GLN C    2
REMARK 465     ALA D    1
REMARK 465     GLN D    2
```

Na parte estrutural (após os *remarks*), podemos observar informações sobre resíduos que contribuem para estruturas secundárias da proteína, ou seja, aqueles resíduos que formam alfa hélices e folhas beta.

HELIX	1	1	LYS A	67	GLN A	75	1				9
HELIX	2	2	LYS B	67	GLN B	75	1				9
HELIX	3	3	LYS C	67	GLN C	75	1				9
HELIX	4	4	LYS D	67	GLN D	75	1				9
SHEET	1	A	8	LYS A	17	PRO A	18	0			
SHEET	2	A	8	LEU A	6	ASN A	12	-1	N	ASN A	12
SHEET	3	A	8	TYR A	97	SER A	104	1	O	LEU A	103
SHEET	4	A	8	GLY A	107	TYR A	111	-1	O	SER A	109
SHEET	5	A	8	GLY B	107	TYR B	111	-1	O	TYR B	108
SHEET	6	A	8	TYR B	97	SER B	104	-1	N	LEU B	102
SHEET	7	A	8	LEU B	6	ASN B	12	1	N	LEU B	11
SHEET	8	A	8	LYS B	17	PRO B	18	-1	O	LYS B	17
SHEET	1	B	8	ARG A	47	ILE A	48	0		ASN B	12
SHEET	2	B	8	LEU A	6	ASN A	12	-1	N	VAL A	8
SHEET	3	B	8	TYR A	97	SER A	104	1	O	LEU A	103
SHEET	4	B	8	GLY A	107	TYR A	111	-1	O	SER A	109
SHEET	5	B	8	GLY B	107	TYR B	111	-1	O	TYR B	108
SHEET	6	B	8	TYR B	97	SER B	104	-1	N	LEU B	102
SHEET	7	B	8	LEU B	6	ASN B	12	1	N	LEU B	11
SHEET	8	B	8	ARG B	47	ILE B	48	-1	O	LEU B	103
SHEET	1	C	4	TRP A	34	LYS A	41	0		VAL B	8
SHEET	2	C	4	THR A	23	LYS A	29	-1	N	VAL A	24
SHEET	3	C	4	GLY A	60	PHE A	66	-1	O	ASP A	61
SHEET	4	C	4	ILE A	84	ILE A	90	-1	O	VAL A	86
SHEET									N	VAL A	64

O cabeçalho também pode fornecer informações sobre sítios ativos das proteínas, podendo dar certa noção de resíduos importantes na função proteica.

SITE	1	AC1	4	HIS A	9	HIS A	98	HOH A1026	HOH A1027
SITE	1	AC2	4	HIS B	9	HIS B	98	HOH B1112	HOH B1113
SITE	1	AC3	4	HIS C	9	HIS C	98	HOH C1107	HOH C1108
SITE	1	AC4	4	HIS D	9	HIS D	98	HOH D1027	HOH D1028
SITE	1	AC5	3	HIS A	96	HIS A	98	SER A 114	
SITE	1	AC6	3	HIS B	96	HIS B	98	SER B 114	
SITE	1	AC7	3	HIS C	96	HIS C	98	SER C 114	
SITE	1	AC8	4	HIS D	96	HIS D	98	SER D 114	HOH D1034
SITE	1	AC9	2	HIS A	96	SER A	114		
SITE	1	BC1	2	HIS B	96	SER B	114		
SITE	1	BC2	3	ASP A	61	HIS A	89	HOH A1028	
SITE	1	BC3	4	ASP B	61	HIS B	89	HOH B1114	ASP D 31
SITE	1	BC4	3	ASP C	61	HIS C	89	HOH C1110	
SITE	1	BC5	4	ASP B	31	ASP D	61	HIS D 89	HOH D1030
SITE	1	BC6	4	GLU A	83	HOH A1030	GLU B 87		HOH B1117
SITE	1	BC7	3	GLU A	87	HOH A1029	GLU B 83		
SITE	1	BC8	3	GLU C	83	GLU D 87			HOH D1031
SITE	1	BC9	4	GLU C	87	GLU D 83			HOH D1032
SITE	1	CC1	5	TRP A	34	ARG A 63	TRP B 34		ARG B 63
SITE	2	CC1	5	HOH B1167					
SITE	1	CC2	7	TRP C	34	ARG C 63	HOH C1122		HOH C1141
SITE	2	CC2	7	HOH C1151		TRP D 34	ARG D 63		

**Passo 3:** Após o cabeçalho, chegamos na seção de "Atoms". Em geral as coordenadas dos átomos são dispostas da seguinte forma:

ATOM	1	N	ASN A	4	25.097	-28.095	33.555	1.00	14.18	N
ATOM	2	CA	ASN A	4	25.151	-29.305	32.679	1.00	14.36	C
ATOM	3	C	ASN A	4	24.503	-30.482	33.393	1.00	14.37	C
ATOM	4	O	ASN A	4	24.778	-30.759	34.559	1.00	14.19	O
ATOM	5	CB	ASN A	4	26.592	-29.652	32.288	1.00	14.40	C
ATOM	6	CG	ASN A	4	26.671	-30.715	31.183	1.00	14.39	C
ATOM	7	OD1	ASN A	4	26.428	-31.917	31.414	1.00	14.07	O
ATOM	8	ND2	ASN A	4	27.034	-30.281	29.986	1.00	14.91	N
ATOM	9	N	ILE A	5	23.643	-31.184	32.678	1.00	14.54	N
ATOM	10	CA	ILE A	5	22.780	-32.173	33.313	1.00	14.54	C
ATOM	11	C	ILE A	5	23.374	-33.574	33.512	1.00	14.33	C
ATOM	12	O	ILE A	5	22.700	-34.409	34.075	1.00	14.34	O
ATOM	13	CB	ILE A	5	21.453	-32.268	32.534	1.00	14.68	C

Cada átomo dispõe de diversas informações como: número, tipo atômico, o resíduo e cadeia na qual pertence, a posição do resíduo, as coordenadas cartesianas ( $x$ ,  $y$  e  $z$ ) e por fim a ocupância e fator-B.

Para exemplificar, tomemos o átomo de carbono marcado em vermelho na figura anterior. Ele é um carbono alfa descrito pelo tipo atômico CA de numeração 2 do resíduo de Asparagina 4 na cadeia A da estrutura. Logo temos as coordenadas  $x$ ,  $y$  e  $z$  e em seguida um valor de ocupação 1.00 e B-fator de 14.36. Vale verificar a documentação do formato de arquivo que pode ser obtida em <https://www.wwpdb.org/documentation/file-format>.

```
ATOM      2  CA  ASN A  4      25.151 -29.305  32.679  1.00 14.36          C
```

## 12.8 Parte 8 – Validação Estrutural

Após obter a sua estrutura de interesse (seja experimentalmente ou computacionalmente), é importante realizar a validação da mesma. Existem diversos parâmetros que podem ser facilmente analisados através de métodos como os gráficos de ERRAT e de Ramachandran, que serão detalhados nesta parte.

**Passo 1:** Acesse o servidor SAVES em <https://saves.mbi.ucla.edu/>.

## UCLA-DOE LAB — SAVES v6.0

UCLA

To run any or all programs:  
upload your structure, in PDB format only

Escolher arquivo Nenhum arquivo escolhido

Run programs

**Passo 2:** Submeta sua estrutura no formato “.pdb” e selecione “Run programs”.

**Passo 3:** Iremos utilizar duas das ferramentas do SAVES: a) ERRAT [11], que avalia a estabilidade e confiança estatística da conformação dos resíduos baseando-se em estruturas de referência; b) PROCHECK [12], que nos permite fazer o gráfico de Ramachandran [13], importante ferramenta de validação de estruturas por meio da análise de permissões dos rotâmetros *phi* () e *psi* () dos resíduos de aminoácidos.

## UCLA-DOE LAB — SAVES v6.0

UCLA

Job 1383190 has been created

New Job

job #1383190: 2g2n.pdb [job link] [3D Viewer]

<b>ERRAT</b> Analyzes the statistics of non-bonded interactions between different atom types and plots the value of the error function versus position of a 9-residue sliding window, calculated by a comparison with statistics from highly refined structures.  <a href="#">Start</a>	<b>Verify3D</b> Determines the compatibility of an atomic model (3D) with its own amino acid sequence (1D) by assigning a structural class based on its location and environment (alpha, beta, loop, polar, nonpolar etc) and comparing the results to good structures.  <a href="#">Start</a>	<b>PROVE</b> Temporarily down at the moment
<b>WHATCHECK</b> Derived from a subset of protein verification tools from the WHATIF program (Vriend, 1990), this does extensive checking of many stereochemical parameters of the residues in the model.  <a href="#">Start</a>	<b>PROCHECK</b> Checks the stereochemical quality of a protein structure by analyzing residue-by-residue geometry and overall structure geometry.  <a href="#">Start</a>	<b>OPEN</b> We are open to suggestions for a 6th program to operate in this window. If you know of a program that we could run locally on our server that would be most useful, please let us know: email holton at mbi dot ucla dot edu with your suggestion

**Passo 4:** O ERRAT irá retornar um valor de confiança de erro para a estrutura. Estruturas resolvidas experimentalmente tendem a ter valores superiores à 90,

como é o caso da estrutura utilizada. Assim como tudo na biologia, exceções existem e devem ser analisadas cuidadosamente.

## ERRAT

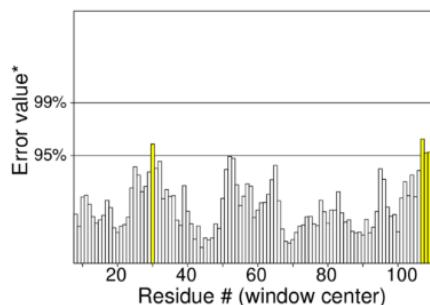
### Overall Quality Factor

**95.1338**

[Log](#) [PostScript](#) [PDF](#)

No gráfico do ERRAT é possível observar o nível de confiança de cada resíduo individualmente. Resíduos indicados em amarelo se encontram acima do intervalo de 95% de confiança de rejeição, e resíduos em vermelho acima de 99%. Os resíduos coloridos tendem a se encontrar em regiões de volta na proteína, ou são aqueles que possuem posições mais móveis e instáveis.

**Nota:** Caso sua proteína tenha mais de uma cadeia (como é o caso do nosso exemplo), o ERRAT irá calcular o gráfico para cada uma individualmente. Mesmo no caso de homotetrâmeros como a HIUase, é possível identificar pequenas variações entre as cadeias.



Passo 5: O PROCHECK pode demorar alguns minutos para ser calculado, mas depois de terminado, clique em “Results”.

## PROCHECK

Out of 8 evaluations

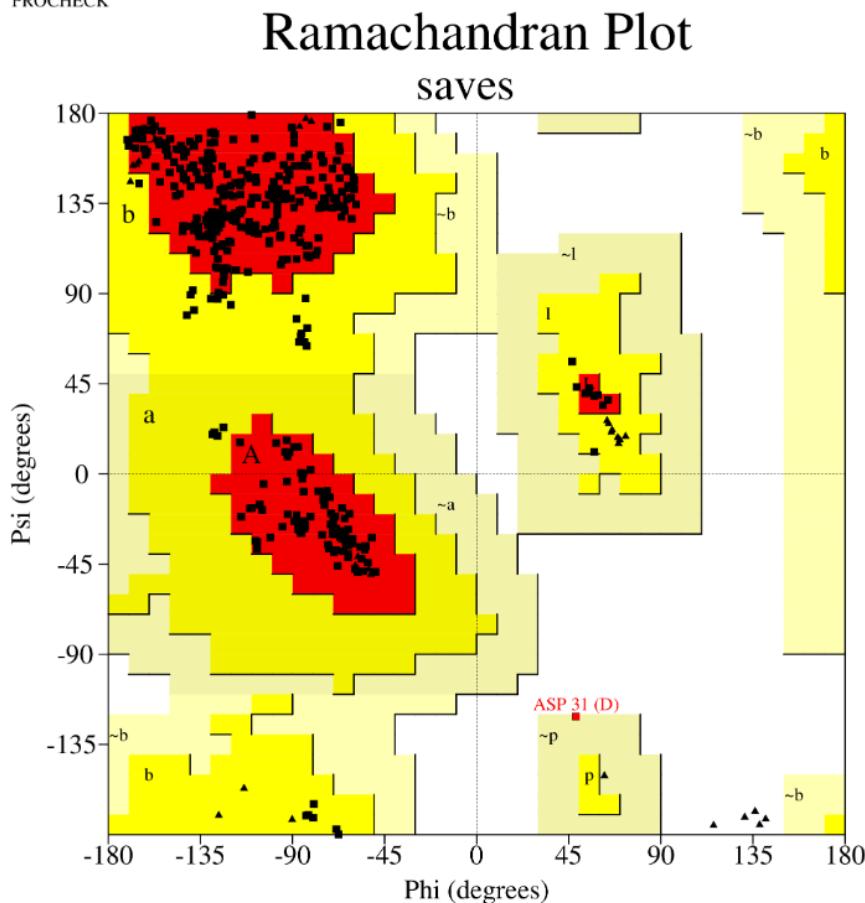
- Errors: 1
- Warning: 4
- Pass: 3

*The evaluations are the '+' (Warning) and '\*' (Error) in the summary. The categories on the left do not always correspond in number due to PROCHECK output documents.*

O programa pode mostrar avisos e erros que podem ser verificados na aba lateral. Não se preocupe tanto com eles, já que a principal métrica que analisaremos aqui é o gráfico de Ramachandran.



Como podemos ver, o aviso significa que outras avaliações podem ser necessárias para refinar a estrutura. Você pode acessar outras ferramentas que desempenham essas funções, como o PDB-REDO (mais detalhes na última parte deste manuscrito). Sabendo disso, podemos analisar o gráfico de Ramachandran.



Aqui podemos ver as regiões permitidas para a rotação dos ângulos *phi* e *psi*, onde as regiões ideais (ou mais permitidas) são aquelas em vermelho. As amarelas e beges são permitidas mas menos aceitas, e em branco são as regiões não permitidas. A quantidade de pontos fora da região vermelha será o seu parâmetro de avaliação da qualidade da estrutura.

No caso da nossa proteína, podemos ver que 92,3% dos resíduos se encontram na região vermelha, 7,4% na região amarela, 0,3% na região bege, e 0% na região branca. Porém, como dito anteriormente, o gráfico de Ramachandran é apenas uma evidência sobre a qualidade da estrutura, e outras análises mais profundadas podem se valer necessárias (você pode encontrar recursos adicionais na última parte deste manuscrito).

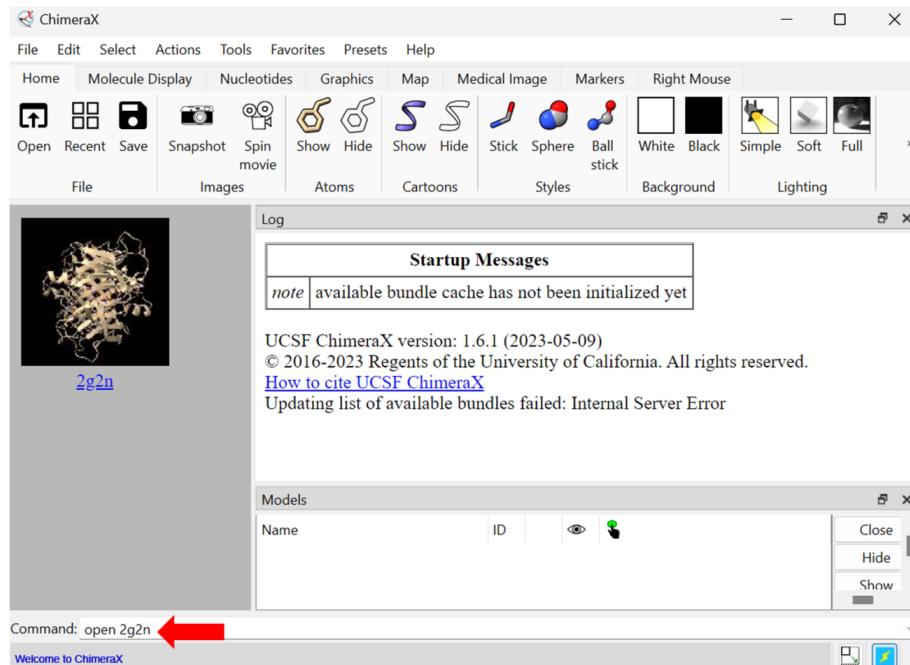
Residues in most favoured regions [A,B,L]	360	92.3%
Residues in additional allowed regions [a,b,l,p]	29	7.4%
Residues in generously allowed regions [-a,-b,-l,-p]	1	0.3%
Residues in disallowed regions	0	0.0%
	---	-----
Number of non-glycine and non-proline residues	390	100.0%
Number of end-residues (excl. Gly and Pro)	8	
Number of glycine residues (shown as triangles)	28	
Number of proline residues	20	
	---	
Total number of residues	446	

## 12.9 Parte 9 – Visualização de Estruturas

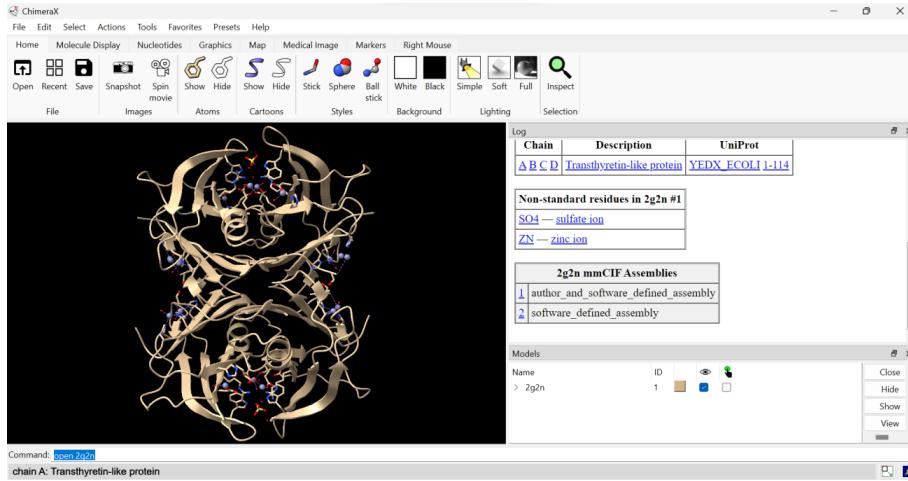
A partir da estrutura tridimensional obtida, o que fazer? Existem diversas informações que podem ser retiradas de suas estruturas através da inspeção visual, e para isto alguns softwares se destacam, como o ChimeraX [14], que será usado para exemplificação.

**Passo 1:** Faça o download do programa ChimeraX em “<https://www.cgl.ucsf.edu/chimerax/download.html>”.

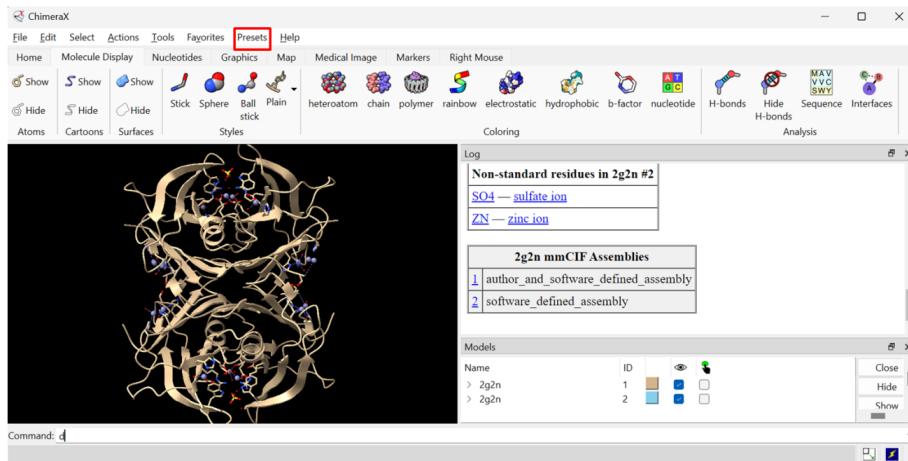
**Passo 2:** Após instalação, podemos abrir uma estrutura do PDB da seguinte maneira:



Com sua proteína aberta, a interface fica da seguinte forma:

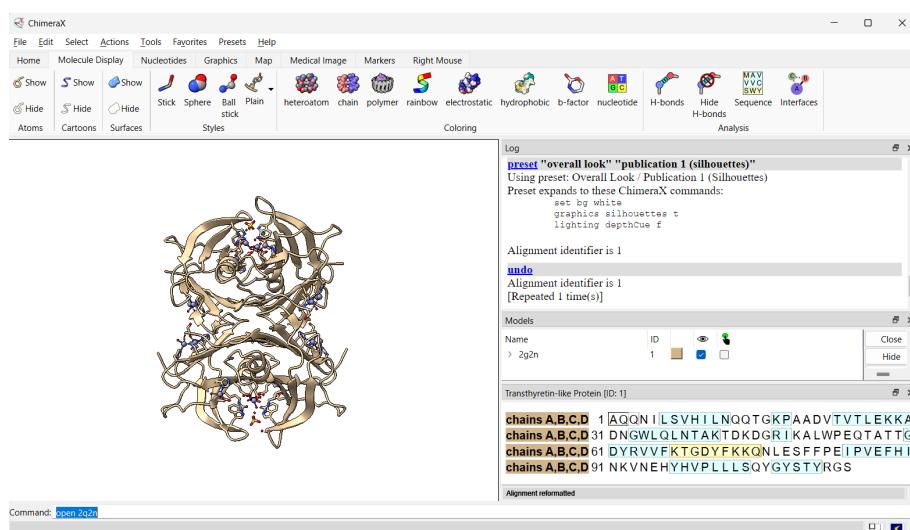
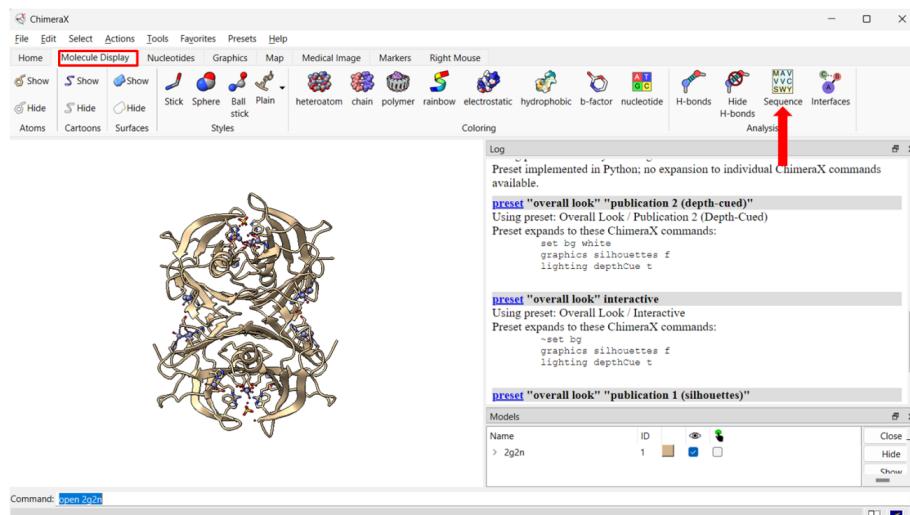


Agora, vamos navegar por algumas funções importantes, mas antes, vamos mudar o fundo.



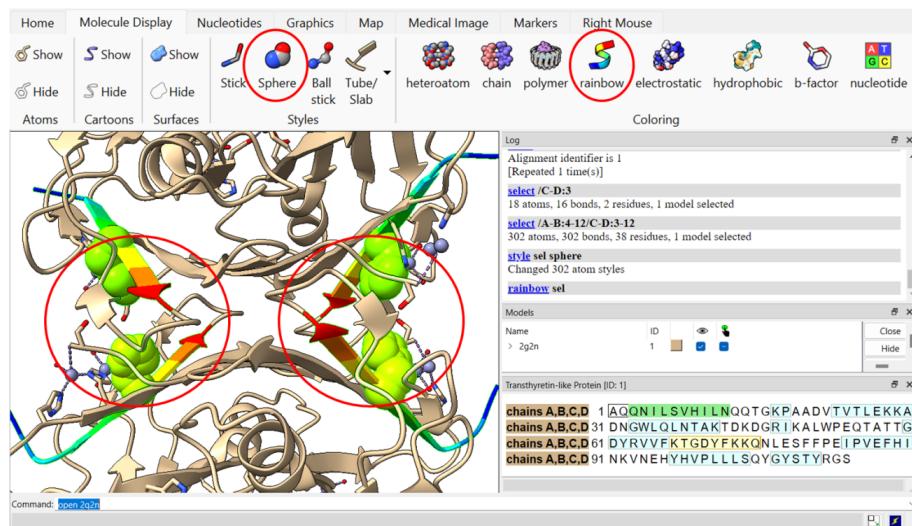
**Passo 3:** Em *Presets*, selecione a opção *Publication 1 (Silhouettes)*.

**Passo 4:** Na seção *Molecule Display* podemos acessar a sequência da estrutura carregada.

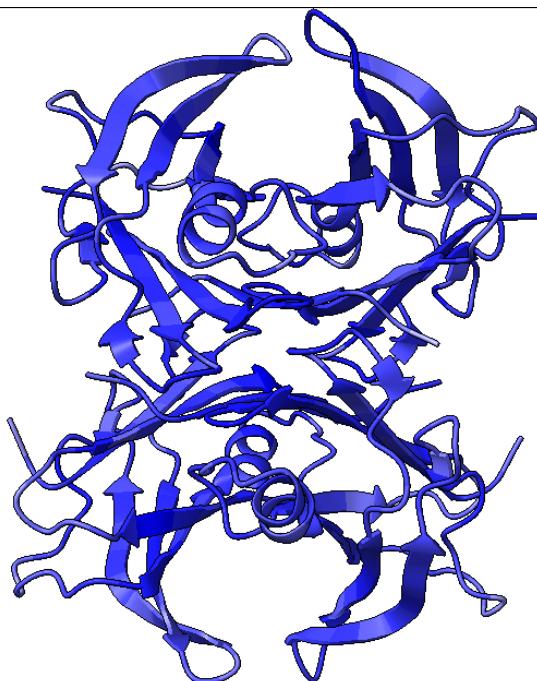


Com o botão esquerdo é possível selecionar resíduos das cadeias, e após selecionado algumas ações podem ser feitas, como mudar a representação gráfica desses resíduos.

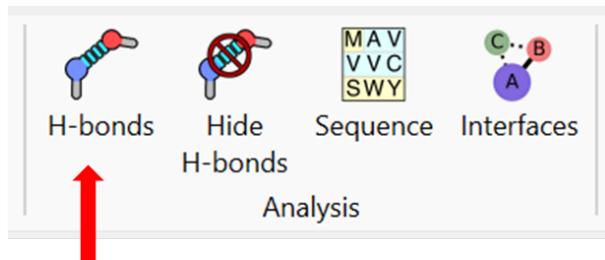
**Passo 5:** Selecione alguns resíduos na sequência e altere a visualização dos mesmos para “Sphere”, e a cor para “Rainbow”.



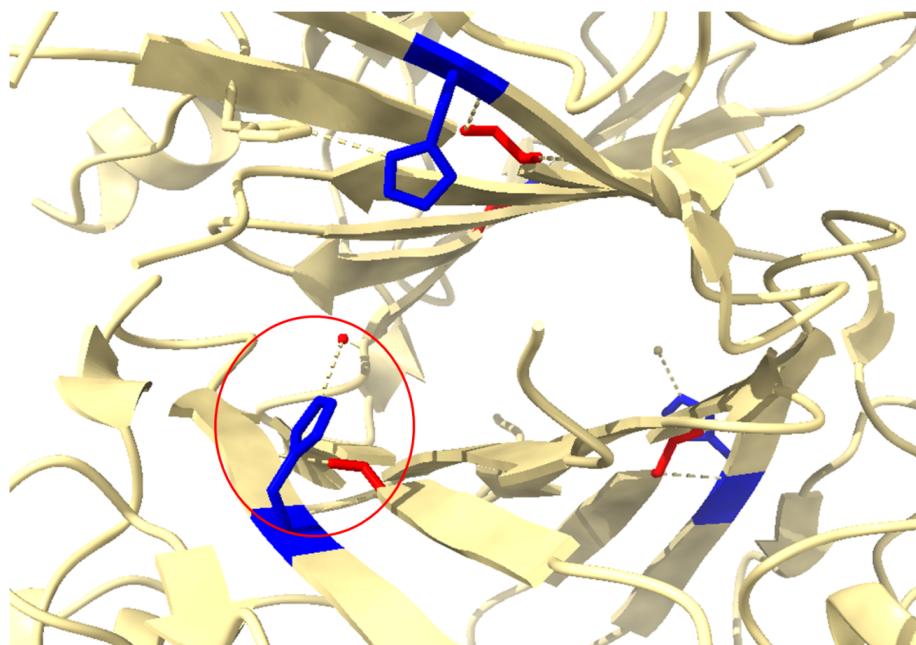
**Passo 6:** Também é possível colorir toda sua estrutura de acordo com o nível de B-fator dos resíduos, que irá nos dar a informação das regiões de maior mobilidade da estrutura.



**Passo 7:** Podemos analisar as ligações de hidrogênio (próximo da seção *Sequence* que foi usada anteriormente) feitas pelos resíduos da proteína.



É possível selecionar aqueles resíduos de interesse como mostrado anteriormente e verificar as interações.



Nesta seleção pode-se observar as interações de hidrogênio dos resíduos 9 (His) e 101(Leu). Note a presença da interação da Histidina com uma molécula de oxigênio representando a água (estruturas cristalográficas não possuem hidrogênios). Esta é uma informação estrutural de extrema importância para esta família de proteínas, visto que a sua ação catalítica parte da transferência de um próton da água para um resíduo de histidina no sítio ativo. E na sua proteína de interesse, quais interações podem ser vistas apenas pela visualização da molécula pelo ChimeraX? Fique à vontade para explorar!

## 12.10 Parte 10 – Introdução ao AlphaFold

Em 2021, foi publicada pela empresa DeepMind a ferramenta AlphaFold, que realiza a predição estrutural e modelagem de proteínas, utilizando apenas a sequência como entrada [15]. Essa ferramenta se baseia em uma abordagem de inteligência artificial, redes neurais, e aprendizado profundo. Mais detalhes podem ser obtidos em um artigo da própria revista BIOINFO [16].

Contudo, nosso foco neste manuscrito não é discutir o AlphaFold em detalhes ou mesmo ensiná-lo a usá-lo para modelar proteínas, e sim discutir as potenciais aplicações da ferramenta e quais informações podem ser aproveitadas de sua predição. Mais informações sobre o AlphaFold podem ser encontradas aqui: <https://www.deepmind.com/research/highlighted-research/alphafold>.

**Nota:** Caso você tenha interesse no funcionamento do AlphaFold e quiser testá-lo, foi publicada uma versão do programa aberta, gratuita e disponível em nuvem, denominada ColabFold [17]. O ColabFold utiliza o GoogleColab e possui diversas versões do AlphaFold disponíveis. Você pode acessá-lo por aqui: <https://github.com/sokrypton/ColabFold>.

Vamos ao que interessa: o AlphaFold se mostrou uma ferramenta tão poderosa para a predição de estruturas de proteínas que não haviam sido ainda depositadas que um grande esforço foi feito para que todas as proteínas do Uniprot fossem resolvidas por meio da ferramenta (isso mesmo, todas as mais de 200 milhões de sequências protéicas conhecidas pela ciência). O resultado se mostrou assustadoramente confiável para a grande maioria delas, mas obviamente nem todos os modelos possuem uma boa qualidade. Mais informações sobre a empreitada podem ser encontradas aqui: <https://www.deepmind.com/blog/alphafold-reveals-the-structure-of-the-protein-universe>

E agora, como saber se os modelos são confiáveis ou não?

**Passo 1:** Acesse novamente o ID Uniprot da proteína que estamos trabalhando (P76341). Na aba lateral esquerda, clique em “Structure”.

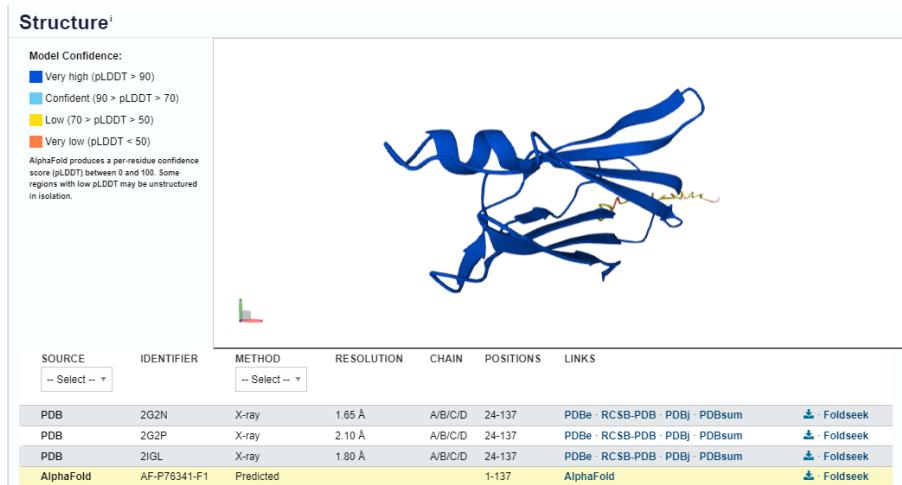
**Structure<sup>i</sup>**

SOURCE	IDENTIFIER	METHOD	RESOLUTION	CHAIN	POSITIONS	LINKS
PDB	2G2N	X-ray	1.65 Å	A/B/C/D	24-137	PDBe RCSB-PDB PDBj PDBsum
PDB	2G2P	X-ray	2.10 Å	A/B/C/D	24-137	PDBe RCSB-PDB PDBj PDBsum
PDB	2IGL	X-ray	1.80 Å	A/B/C/D	24-137	PDBe RCSB-PDB PDBj PDBsum
AlphaFold	AF-P76341-F1	Predicted			1-137	AlphaFold

Como nossa estrutura já possui modelos experimentais depositados no PDB, você deve perceber, abaixo da visualização da estrutura, que existem diversas entradas. Inclusive, a primeira delas é a 2G2N, que acabamos de trabalhar e analisar nos passos anteriores. Você pode identificar a origem, ID, método de obtenção, resolução, cadeias e resíduos, etc.

SOURCE	IDENTIFIER	METHOD	RESOLUTION	CHAIN	POSITIONS	LINKS
PDB	2G2N	X-ray	1.65 Å	A/B/C/D	24-137	PDBe RCSB-PDB PDBj PDBsum
PDB	2G2P	X-ray	2.10 Å	A/B/C/D	24-137	PDBe RCSB-PDB PDBj PDBsum
PDB	2IGL	X-ray	1.80 Å	A/B/C/D	24-137	PDBe RCSB-PDB PDBj PDBsum
AlphaFold	AF-P76341-F1	Predicted			1-137	AlphaFold

**Passo 2:** Suponhamos que sua proteína de interesse ainda não foi resolvida experimentalmente. Dessa forma, apenas a entrada modelada computacionalmente pelo AlphaFold estará disponível. Clique nela.



À esquerda da visualização da proteína, temos agora a principal métrica de confiança do AlphaFold, chamada  $p\text{LDDT}$  (“Predicted Local Distance Difference Test”)[18]. Essa métrica baseia-se em uma avaliação resíduo-específica da distância local entre todos os átomos, e como o ambiente de uma estrutura de referência é reproduzido em um modelo.

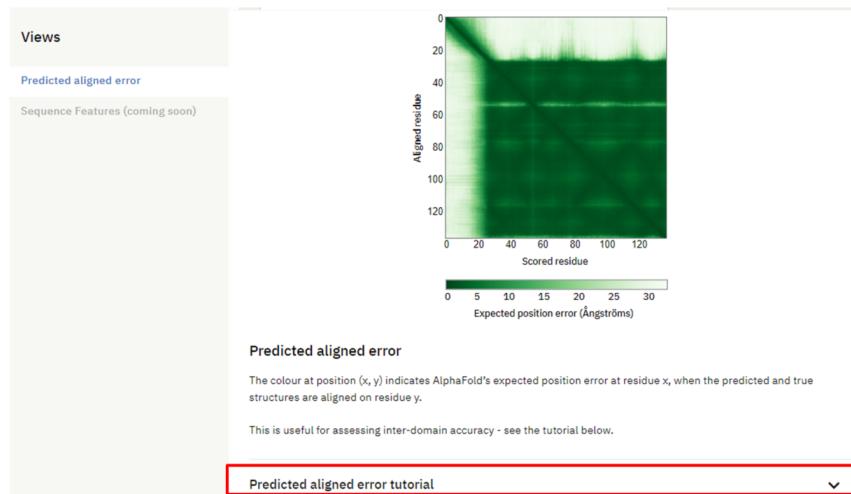
Valores azuis escuros de  $p\text{LDDT}$  ( $>90$ ) indicam alta confiança na modelagem daquele resíduo específico, enquanto valores laranjas indicam baixíssima confiança ( $<50$ ). Na nossa proteína, é possível observar que quase toda a estrutura do monômero se encontra azul escura (ou seja, modelada com alta confiança), enquanto uma pequena porção dos resíduos possui coloração amarelada/alaranjada. Esses últimos resíduos se referem justamente aos primeiros resíduos da sequência, que formam o peptídeo sinal e não possuem uma conformação definida. É importante lembrar que apenas uma das 4 cadeias da proteína foi modelada!

**Nota:** Você pode brincar com a visualização da proteína na própria página do Uniprot utilizando a roda do mouse para dar zoom e clicando e arrastando para rotacionar. Ao passar o mouse em cima de algum resíduo, as informações do mesmo irão aparecer no canto inferior direito.

**Passo 3:** Para obter mais informações sobre o modelo do AlphaFold, clique no link do AlphaFold disponível no próprio Uniprot.

SOURCE	IDENTIFIER	METHOD	RESOLUTION	CHAIN	POSITIONS	LINKS
PDB	2G2N	X-ray	1.65 Å	A/B/C/D	24-137	PDBe · RCSB-PDB · PDBj · PDBsum 
PDB	2G2P	X-ray	2.10 Å	A/B/C/D	24-137	PDBe · RCSB-PDB · PDBj · PDBsum 
PDB	2IGI	X-ray	1.80 Å	A/B/C/D	24-137	PDBe · RCSB-PDB · PDBj · PDBsum 
AlphaFold	AF-P76341-F1	Predicted			1-137	AlphaFold 

Aqui, além de poder visualizar melhor a estrutura da proteína, você pode obter outras informações de qualidade, como a métrica de PAE (“*Predicted Alignment Error*”), que mede a posição relativa das cadeias da proteína, utilizando a distância entre os resíduos. O gráfico de PAE é mostrado logo após a estrutura da proteína, e mais informações e interpretações podem ser obtidas logo abaixo, na caixa de tutorial.



**Passo 4:** Se quiser analisar a estrutura em alguma ferramenta externa (como ChimeraX ou PyMOL), baixe o arquivo PDB da proteína modelada pelo AlphaFold.

SOURCE	IDENTIFIER	METHOD	RESOLUTION	CHAIN	POSITIONS	LINKS
PDB	2G2N	X-ray	1.65 Å	A/B/C/D	24-137	PDBe · RCSB-PDB · PDBj · PDBsum 
PDB	2G2P	X-ray	2.10 Å	A/B/C/D	24-137	PDBe · RCSB-PDB · PDBj · PDBsum 
PDB	2IGI	X-ray	1.80 Å	A/B/C/D	24-137	PDBe · RCSB-PDB · PDBj · PDBsum 
AlphaFold	AF-P76341-F1	Predicted			1-137	AlphaFold 

## 12.11 Parte 11 – Ferramentas Extras

Por fim, apresentaremos, em ordem alfabética, outras ferramentas online e gratuitas que podem ser úteis para investigação computacional de sequências e estruturas de proteínas.

- **CATH:** Base de dados de classificação de estruturas e domínios proteicos do Protein Data Bank em superfamílias. Disponível em: <http://cathdb.info/>.
- **CAVER:** Programa para análise e visualização de cavidades e túneis em estruturas de proteínas. Possui uma versão para download (CAVER Analyst), e uma versão online como *plugin* para o visualizador de estruturas PyMOL. Disponível em: <https://www.caver.cz/>.
- **ClustalOmega:** Ferramenta para alinhamento múltiplo de sequências. Disponível em: <https://www.ebi.ac.uk/Tools/msa/clustalo/>.
- **HMMER:** Realiza alinhamentos de sequências e busca em bancos de dados por sequências homólogas de proteínas. O AlphaFold utiliza uma versão modificada (chamada jackHMMER) para realizar a busca de sequências para modelagem. Disponível em: <http://hmmer.org/>.
- **PDB REDO:** Esta plataforma valida e oferece versões otimizadas das estruturas depositadas no PDB, alterando certos parâmetros a fim de obter estruturas que melhor representem as formas ativas das proteínas. Disponível em: <https://pdb-redo.eu/>.
- **PDB Sum:** Fornece um “resumo” de uma entrada no PDB. Você pode obter informações rapidamente, como estrutura secundária, sítios de ligação, cavidades, etc. Disponível em: <https://www.ebi.ac.uk/thornton-srv/databases/pdbsum/>.
- **PFAM/InterPro:** Base de dados sobre famílias proteicas. Você pode obter informações como domínios, proteínas relacionadas por similaridade

de sequências ou características funcionais, etc. Disponível em: <https://www.ebi.ac.uk/interpro/>.

- **PreStO:** *Webservice* para encontrar ferramentas computacionais relacionadas à biologia estrutural. Você pode procurar por termos específicos ou na lista de quase 100 ferramentas listadas. Disponível em: <http://bioinfo.dcc.ufmg.br/presto/>.
- **PROPKA:** Ferramenta de predição de estados de protonação. O programa auxilia na escolha de protonação dos resíduos de uma proteína visto que algumas metodologias de resolução de estruturas, como a cristalografia, não resolvem a posição dos hidrogênios. Um *pipeline* oferecido pela plataforma prediz os valores de pKa dos resíduos e retorna um arquivo PDB com a protonação no pH desejado. Disponível em: <https://server.poissonboltzmann.org/pdb2pqr>.
- **SCOPe:** Base de dados de classificação de estruturas proteicas. Disponível em: <https://scop.berkeley.edu/>.
- **SMART:** Similar ao PFAM, fornece informações como estruturas de domínios e famílias proteicas. Disponível em <https://smart.embl-heidelberg.de/>.
- **VADAR:** Ferramenta especialmente direcionada para a identificação de resíduos expostos ou enterrados em estruturas. Diversas outras informações, como estruturas secundárias e métricas de qualidade também podem ser obtidas. Disponível em: <http://vadar.wishartlab.com/>.

#### Nota de transparência 12.1

Este material foi originalmente produzido para um minicurso ministrado durante o Curso de Inverno em Bioinformática da UFMG, realizado em 4 de Julho de 2023, na Universidade Federal de Minas Gerais, Belo Horizonte, Brasil.

### Saiba mais 12.1

Este artigo está disponível em <https://bioinfo.com.br/extracao-de-informacoes-de-sequencias-e-estruturas-de-proteinas/>

## 12.12 Referências

1. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic Local Alignment Search Tool. *J. Mol. Biol.* 1990, **215**, 403–410, doi:10.1016/S0022-2836(05)80360-2.
2. Altschul, S.F.; Madden, T.L.; Schäffer, A.A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D.J. Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs. *Nucleic Acids Res.* 1997, **25**, 3389–3402.
3. Gasteiger, E.; Gattiker, A.; Hoogland, C.; Ivanyi, I.; Appel, R.D.; Bairoch, A. ExPASy: The Proteomics Server for in-Depth Protein Knowledge and Analysis. *Nucleic Acids Res.* 2003, **31**, 3784–3788.
4. Gasteiger, E.; Hoogland, C.; Gattiker, A.; Duvaud, S.; Wilkins, M.R.; Appel, R.D.; Bairoch, A. Protein Identification and Analysis Tools on the ExPASy Server. In *The Proteomics Protocols Handbook*; Walker, J.M., Ed.; Springer Protocols Handbooks; Humana Press: Totowa, NJ, 2005; pp. 571–607 ISBN 978-1-59259-890-8.
5. Lyu, Z.; Wang, Z.; Luo, F.; Shuai, J.; Huang, Y. Protein Secondary Structure Prediction With a Reductive Deep Learning Method. *Front. Bioeng. Biotechnol.* 2021, **9**, 687426, doi:10.3389/fbioe.2021.687426.
6. McGuffin, L.J.; Bryson, K.; Jones, D.T. The PSIPRED Protein Structure Prediction Server. *Bioinformatics* 2000, **16**, 404–405, doi:10.1093/bioinformatics/16.4.404.
7. Buchan, D.W.A.; Jones, D.T. The PSIPRED Protein Analysis Workbench: 20 Years On. *Nucleic Acids Res.* 2019, **47**, W402–W407, doi:10.1093/nar/gkz297.
8. Hallgren, J.; Tsirigos, K.D.; Pedersen, M.D.; Armenteros, J.J.A.; Marcatili, P.; Nielsen, H.; Krogh, A.; Winther, O. DeepTMHMM Predicts Alpha and Beta Transmembrane Proteins Using Deep Neural Networks 2022, 2022.04.08.487609.
9. Silva, Letícia Xavier; Bastos, Luana Luiza; Santos, Lucianna Helene. Modelagem computacional de proteínas. In: BIOINFO - Revista Brasileira de Bioinformática e Biologia Computacional, v.1, 2021, n.8. doi: 10.51780/978-6-599-275326-08
10. Schaffer, J.E.; Kukshal, V.; Miller, J.J.; Kitainda, V.; Jez, J.M. Beyond X-Rays: An Overview of Emerging Structural Biology Methods. *Emerg. Top. Life Sci.* 2021, **5**, 221–230, doi:10.1042/ETLS20200272.

11. Colovos, C.; Yeates, T.O. Verification of Protein Structures: Patterns of Nonbonded Atomic Interactions. *Protein Sci.* 1993, 2, 1511–1519, doi:10.1002/pro.5560020916.
12. Laskowski, R.A.; MacArthur, M.W.; Moss, D.S.; Thornton, J.M. PROCHECK: A Program to Check the Stereochemical Quality of Protein Structures. *J. Appl. Crystallogr.* 1993, 26, 283–291, doi:10.1107/S0021889892009944.
13. Ramachandran, G.N.; Ramakrishnan, C.; Sasisekharan, V. Stereochemistry of Polypeptide Chain Configurations. *J. Mol. Biol.* 1963, 7, 95–99, doi:10.1016/S0022-2836(63)80023-6.
14. Pettersen, E.F.; Goddard, T.D.; Huang, C.C.; Meng, E.C.; Couch, G.S.; Croll, T.I.; Morris, J.H.; Ferrin, T.E. UCSF ChimeraX: Structure Visualization for Researchers, Educators, and Developers. *Protein Sci. Publ. Protein Soc.* 2021, 30, 70–82, doi:10.1002/pro.3943.
15. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* 2021, 596, 583–589, doi:10.1038/s41586-021-03819-2.
16. Mariano, D. AlphaFold e a busca pelo Santo Graal da Biologia Molecular. In: BIOINFO 02 - Revista Brasileira de Bioinformática e Biologia Computacional. Vol. 2. 2022. doi: 10.51780/978-65-992753-5-7-10
17. Mirdita, M.; Schütze, K.; Moriwaki, Y.; Heo, L.; Ovchinnikov, S.; Steinegger, M. ColabFold: Making Protein Folding Accessible to All. *Nat. Methods* 2022, 19, 679–682, doi:10.1038/s41592-022-01488-1.
18. Tunyasuvunakool, K.; Adler, J.; Wu, Z.; Green, T.; Zielinski, M.; Žídek, A.; Bridgland, A.; Cowie, A.; Meyer, C.; Laydon, A.; et al. Highly Accurate Protein Structure Prediction for the Human Proteome. *Nature* 2021, 596, 590–596, doi:10.1038/s41586-021-03828-1.

# 13 UM MUNDO DENTRO

## DE NÓS: EXPLORANDO A MICROBIOTA HUMANA ATRAVÉS DA METATRANSCRIPTÔMICA

Autores 13.1

Aline de Paula Dias da Silva , Monique Cristina dos Santos 

Revisão: Isaac Farias Cansanção , Mira Raya Paula de Lima 

Cite este artigo 13.1

Silva, APD; Snatos, MC. **Um mundo dentro de nós: explorando a microbiota humana através da metatranscriptômica.** BIOINFO. ISSN: 2764-8273. Vol. 3. p.13 (2023). doi: 10.51780/bioinfo-03-13

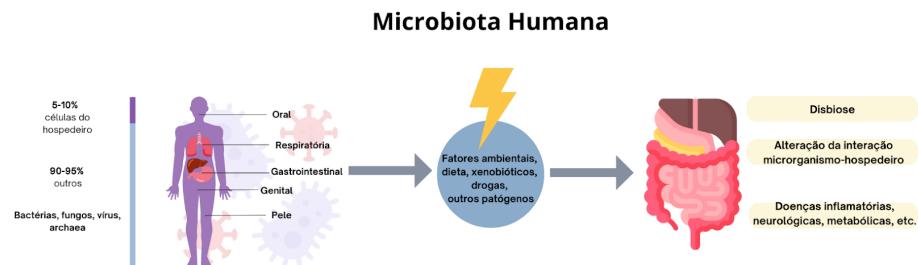
### Resumo 13.1

Neste artigo, você irá explorar a microbiota humana através da metranscriptômica.

VOCÊ já parou para pensar na quantidade de microrganismos existentes no mundo? E no nosso corpo? Claro, que como sempre, não estamos sozinhos no mundo e nem no nosso próprio corpo. Atualmente, sabe-se que mais de 1000 espécies de microrganismos habitam nosso corpo, fazendo parte do nosso microbioma [1]. O termo microbioma, cunhado pelo vencedor do prêmio Nobel, Joshua Lederberg, pode ser descrito como: “Um sistema ecológico de microrganismos comensais, simbióticos e também patogênicos que residem no corpo humano” [2]. Possuir um microbioma não é algo exclusivo do ser humano, já que podemos utilizar o termo “microbioma” para nos referir ao microbioma de plantas e de outros animais. Apesar dos termos “microbioma” e “microbiota” serem semelhantes, há algumas diferenças entre eles. Quando falamos de microbiota, estamos nos referindo ao conjunto dos microrganismos presentes em determinada região ou órgão, enquanto o microbioma é o conjunto de genes e materiais genéticos dos microrganismos que habitam determinado ambiente. Além disso, quando pensamos em microbiota, o primeiro pensamento que pode vir à sua mente, deve ser das bactérias intestinais. Porém, esses microrganismos podem ser bactérias, fungos, vírus ou archaea e não estão presentes somente no intestino, mas habitando outros locais do corpo, como boca, garganta, vias aéreas, estômago, intestino, sistema urogenital e pele [3].

Mas qual é o papel desses microrganismos no nosso corpo? Além de constituírem 90% do número total de células em nossos corpos [4], eles têm um papel crucial na manutenção da homeostase, com diversos benefícios ao hospedeiro, como: desenvolvimento do sistema imunológico e absorção de nutrientes [5]. Sabe-se que alterações na microbiota podem ocorrer por vários fatores, sejam fatores ambientais, dieta, xenobióticos, drogas e até mesmo outros patógenos podem alterar esse equilíbrio entre o microbioma e hospedeiro, o que

é chamado de disbiose. Atualmente se sabe que essas relações disbióticas estão correlacionadas no desenvolvimento de doenças inflamatórias e cânceres [6].



*Figura 13.1: Principais sítios da microbiota humana. Diversos fatores podem alterar a composição da microbiota humana e levar ao surgimento de diferentes condições. Fonte: autores.*

O principal objetivo do estudo do microbioma humano é de fato estudar a estrutura e a dinâmica das comunidades microbianas e a relação do microbioma e hospedeiro tanto em indivíduos saudáveis quanto na doença. Dessa forma, podemos estudar as interações entre os DNAs e RNAs que estão presentes no microbioma através de técnicas avançadas de sequenciamento de nova geração (NGS) e obter diversas respostas, pois cada técnica irá nos fornecer um tipo de resposta.

Existem algumas técnicas de sequenciamento que podem ser empregadas para o estudo do microbioma, entre elas: 1) o sequenciamento 16S que se baseia no sequenciamento das regiões hipervariáveis do gene rRNA 16S de bactérias. Esse tipo de sequenciamento é empregado devido à alta conservação evolutiva do gene 16S presente nas bactérias, presença de regiões variáveis nesse gene que permitem a diferenciação de diferentes grupos bacterianos e, além disso, pelo tamanho do gene ser menor, a amplificação e sequenciamento desse tipo de genoma é facilitada; 2) sequenciamento do DNA total (metagenoma); 3) sequenciamento do RNA total (metatranscriptoma).

### 13.1 Por que fazer metatranscriptômica?

A metatranscriptômica anda de mãos dadas com a metagenômica. A análise metagenômica identifica a abundância e a diversidade da comunidade microbiana. Enquanto isso, a análise do metatranscriptoma consegue identificar

os microrganismos e os transcritos que estão sendo expressos no microbioma (Figura 2). No estudo de microbiomas, a metatranscriptômica possui algumas vantagens já que ao sequenciar o RNA total (RNA-Seq) de uma determinada amostra:

- 1) Temos informações importantes sobre a expressão gênica, ou seja, podemos ver quais os microrganismos e genes que estão sendo ativados em determinado compartimento do corpo e quais são as suas funções que eles desempenham;
- 2) É possível realizar estudos *in vitro* e investigar as diferenças no microbioma de acordo com mudanças ambientais ou a presença de uma doença, por exemplo;
- 3) Caracterizar genes ativos de resistência a antibióticos e interações hospedeiro-microbioma;
- 4) Descoberta de novos genes e funções que ainda não foram descritas, além da descoberta de novos marcadores para diversas doenças;
- 5) Diferentemente do sequenciamento 16S, podemos verificar não somente a presença de bactérias, mas também de fungos, vírus e archaea.

## Estratégias para o estudo do microbioma humano

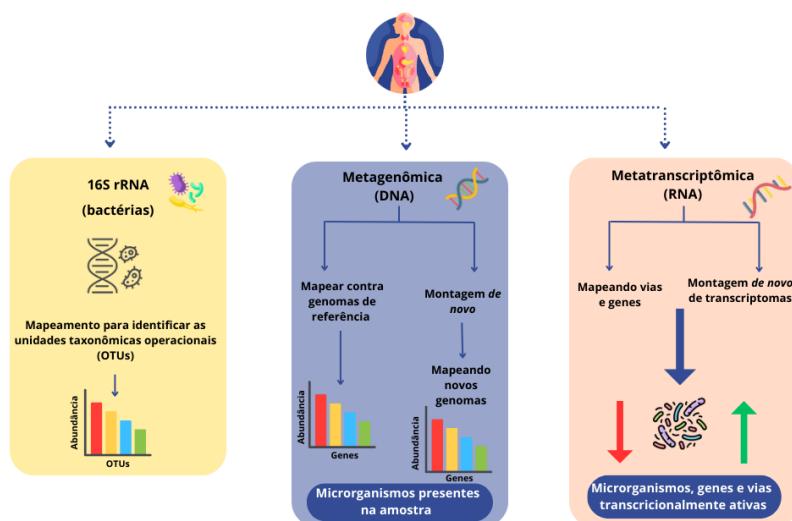


Figura 13.2: Principais estratégias de sequenciamento para o estudo do microbioma humano. Fonte: adaptado de [11].

## **13.2 E como estudar o microbioma através da Bioinformática?**

Agora que já compreendemos o que é a metatranscriptômica e quais são as suas vantagens, nos damos com a seguinte questão: Como analisar os dados gerados por esse tipo de estratégia? Assim como todos os dados brutos gerados por sequenciamento, eles devem ser pré-processados. Esse pré-processamento é necessário para podermos filtrar sequências que possuem uma qualidade muito baixa e remover adaptadores e possíveis contaminantes na nossa amostra. Um exemplo de programa muito útil para o pré-processamento que podemos citar é o Trimmomatic, o qual é muito simples de ser utilizado. O processo de montagem das sequências em metatranscriptômica pode ser desafiador, principalmente pela quantidade de microrganismos presentes e a complexidade das comunidades microbianas em uma amostra. Para este passo, podemos utilizar de duas estratégias distintas, como o mapeamento através de genomas e genes de referência e a montagem *de novo* de novos transcriptomas. Após mapearmos as leituras do RNA com base em genomas de referência, podemos identificar taxonomicamente os microrganismos presentes e qual é a funcionalidade do gene que está sendo expresso. Esse processo pode ser realizado por uma gama de softwares, como, por exemplo, Trinity e MEGAHIT [7-8]. Com esses dados, podemos utilizar o banco de dados KEGG para obter as vias cujos genes expressos estão regulados no microbioma, seja mais expresso ou menos expresso durante condições de saúde e doença [9]. Dessa forma, a partir dos dados de um RNA-Seq e dos diversos softwares disponíveis, podemos quantificar a expressão de genes e analisar o perfil dos microrganismos presentes.

## **13.3 Não sou Bioinformata! E agora?**

Se você deseja analisar o microbioma da sua amostra e não tem familiaridade com os softwares de bioinformática, não se desespere. A tecnologia vem avançando conforme os anos para nos auxiliar nessas questões e facilitar o nosso trabalho. Atualmente existem ferramentas online que podem realizar análises de metagenômica e metatranscriptômica. Um exemplo de ferramenta é o ID-

Seq, que é uma plataforma de bioinformática de código aberto baseada em nuvem que permite a detecção de microrganismos a partir de leituras brutas de sequenciamento de próxima geração (NGS) [10]. No pipeline do ID-Seq pode-se utilizar sequenciamento de RNA ou DNA de qualquer tipo de amostra, e o upload pode ser realizado pelo programa web ou por de linha de comando (CLI). A análise do sequenciamento passa pelas fases principais que citamos anteriormente: filtragem das sequências e controle de qualidade, montagem e também são gerados relatórios taxonômicos.

### **13.4 O futuro das pesquisas sobre o microbioma humano**

Vimos a importância de estudar o microbioma humano. Seja qual for a estratégia utilizada, podemos obter diversas informações sobre a interação e presença desses microrganismos no nosso corpo, identificando vias e genes que podem estar ativos em situação de saúde ou em diferentes doenças. Mas o que pode ser feito com essas informações?

A identificação das vias e genes ativos podem servir como alvo farmacológico e como tratamento para inúmeras doenças em que há uma disbiose no microbioma. O desenvolvimento de medicamentos probióticos é uma área que muito se interessa pelos estudos da microbiota humana, visto que esses medicamentos podem ajudar a reconstruir a comunidade microbiana residente e ser utilizado como tratamento para doenças.

Porém, ainda há muito a ser descoberto sobre os microrganismos que habitam em nós e o papel deles em nosso corpo. Com esses estudos, o potencial para melhorar a saúde humana é grande. Com o contínuo avanço da tecnologia e da bioinformática, estamos cada vez mais perto de desvendar os “segredos” desses microrganismos e utilizar desse conhecimento para prevenir doenças de maneira mais precisa e personalizada.

### Saiba mais 13.1

Este artigo está disponível em <https://bioinfo.com.br/um-mundo-dentro-de-nos-explorando-a-microbiota-humana-atraves-da-metatranscriptomica/>

## 13.5 Referências

[1] TURNBAUGH, P.J. et al. The human microbiome project. *Nature*, v. 449, p. 804–810, 2007.

[2] LEDERBERG, J. “Ome Sweet’Omics—A Genealogical Treasury of Words”. *The Scientist*, 2001.

[3] YAGI, K.; HUFFNAGLE, G.B.; LUKACS, N.W.; ASAI, N. The Lung Microbiome during Health and Disease. *Int J Mol Sci*, 2021.

[4] SAVAGE, D.C. Microbial ecology of the gastrointestinal tract. *Annu. Rev. Microbiol.*, v. 31, p. 107–133, 1977.

[5] MOENS, E.; VELDHOEN, M. Epithelial barrier biology: good fences make good neighbours. *Immunology*, 2012.

[6] RAJAGOPALA, S.V. et al. The human microbiome and cancer. *Cancer Prevention Research*, v. 10, n. 4, 2017.

[7] GRABHERR, M.G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol*, v. 29, n. 7, p. 644-652, 2011.

[8] LI, D. et al. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*, v. 31, n. 10, May 2015.

[9] KANEHISA, M.; GOTO, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, v. 28, n. 1, p. 27–30, 2000.

[10] KALANTAR, K.L. et al. IDseq—An open source cloud-based pipeline and analysis service for metagenomic pathogen detection and monitoring. *GigaScience*, v. 9, n. 10, October 2020.

[11] BIKEL, S. et al. Combining metagenomics, metatranscriptomics and viromics to explore novel microbial interactions: towards a systems-level understanding of human microbiome. *Comput Struct Biotechnol J*, 2015.

# 14 BIOINFORMÁTICA COMO UMA FERRAMENTA DIDÁTICA PARA O ENSINO DA GENÉTICA

## Autores 14.1

Marcos Antonio Nobrega de Sousa , Jeniffer Gabrielly de Sousa Pereira , Ana Luíza Vieira Soares , Ana Beatriz Braz dos Santos , Ricardo Henrique Pereira da Silva , Arthur Moraes de Medeiros , Bruna Lima de Araujo , Francisca Vitória Amaral Nobrega 

Revisão: Isaac Farias Cansanção , Diego Lucas Neres Rodrigues 

## Cite este artigo 14.1

Sousa, MAN *et al.* Bioinformática como uma ferramenta didática para o ensino da Genética. BIOINFO. ISSN: 2764-8273. Vol. 3. p.14 (2023). doi: 10.51780/bioinfo-03-14

### Resumo 14.1

O ensino do conteúdo de genética geralmente é um desafio tanto para os professores da educação básica quanto do ensino superior, devido ao seu alto nível de complexidade e abstração. Esse trabalho tem por objetivo mostrar a importância da bioinformática como uma ferramenta didática para o ensino da genética. Como um exemplo, para ilustrar este tema, foi realizada a aplicação de um questionário para coletar as respostas de estudantes de licenciatura em ciências biológicas de uma universidade pública brasileira após uma aula expositiva com utilização de ferramenta de bioinformática. Ao analisar os resultados foi comprovado pouco conhecimento dos discentes acerca do assunto. Portanto, verificou-se que ao ser utilizada em aulas de filogenia e genética, a bioinformática proporciona melhor entendimento das mesmas, e quebra a barreira da abstração, oferecendo dados facilmente acessíveis numa ferramenta didática que facilita o processo de ensino-aprendizagem de genética/biologia.

**Palavras-chaves:** Bioinformática, ferramenta didática, ensino de genética, biologia.

### 14.1 *Abstract*

*Teaching genetics content is usually a challenge for both basic and higher education teachers, due to its high level of complexity and abstraction. This work aims to show the importance of bioinformatics as a didactic tool for teaching genetics. As an example, to illustrate this theme, a questionnaire was applied to collect responses from undergraduate students in biological sciences at a Brazilian public university after an expository class using a bioinformatics tool. When analyzing the results, it was proved that the students had little knowledge about the subject. Therefore, it was found that when used in phylogeny and genetics classes, bioinformatics provides a better understanding of them, and breaks the barrier of abstraction, offering*

*easily accessible data in a didactic tool that facilitates the teaching-learning process of genetics/biology.*

**Keywords:** Bioinformatics, didactic tool, genetics teaching, biology.

## 14.2 Introdução

O ensino de genética e biologia molecular, apresenta-se como um desafio para os professores de ciências e biologia, pois apresentam um nível de abstração e descontextualização da realidade dos alunos [1]. Em decorrência disso, é comum os professores relatarem dificuldades no ensino da genética, pois aulas práticas são geralmente inviáveis, devido ao planejamento das atividades práticas-experimentais e dos custos que envolvem a utilização de materiais adequados para sua execução [2].

O avanço das ciências e o aprimoramento técnico científico, possibilitaram a introdução das tecnologias em diversas áreas de conhecimento. Apesar da valorização de livros didáticos, atualmente a educação básica necessita da inserção de tecnologias digitais nas suas metodologias, para assim, fugir do método tradicional de ensino, e poder facilitar o processo de ensino-aprendizagem [3].

A bioinformática surge como um meio alternativo para sanar essa dificuldade que ocorre na área da Biologia e Genética, pois pode ser definida como uma área de saber interdisciplinar, que tem por objetivo desenvolver e investigar sistemas que ajudem na compreensão do fluxo de informações, desde os genes até as estruturas moleculares [4]. A Bioinformática utiliza bancos de dados para armazenar um catálogo geral de informações sobre a variação do material genético, para atender a projetos de sequenciamento em larga escala [5]. Estas informações podem ser encontradas, por exemplo, por meio de sites, como: *Center for Biotechnology Information* (NCBI), que é um portal de informações sobre biotecnologia e bioinformática “<https://www.ncbi.nlm.nih.gov/>”, GenBank, que é um banco de dados de anotações de sequências de nucleotídeos e traduções de proteínas, disponíveis publicamente “<https://www.ncbi.nlm.nih.gov/genbank/>”, PubMed, que é um portal de literatura científica na área biomédica e de

ciências da vida “<https://www.ncbi.nlm.nih.gov/pmc/>”, e LocusLink e RefSeq, recursos do NCBI, que facilitam a recuperação de informações baseadas em genes e fornecem padrões de sequência de referência “<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC102393/>”, entre outros.

Estes dados podem ser utilizados para a execução de aulas experimentais em assuntos que envolvem genética, biologia celular e molecular e até mesmo filogenia. Porém, a maioria dos alunos e licenciados em ciências biológicas, desconhecem sobre as aplicações e o uso da bioinformática no curso. Por ela ser considerada uma ciência nova e ainda pouco difundida no meio acadêmico, mesmo frente a essencialidade da inovação de práticas para a utilização de novas tecnologias. Além disso, este assunto não foi abordado durante o processo de formação da maioria dos professores [6].

Ao ser utilizada como ferramenta didática, a bioinformática proporciona aos docentes uma forma de trabalhar e facilitar a compreensão de temas abstratos e conceituais, como o que se encontram na genética [7]. Neste sentido, este trabalho teve por objetivo mostrar a importância da bioinformática como uma ferramenta didática para o ensino de genética e filogenia. Além de analisar o conhecimento dos licenciandos em ciências biológicas, a respeito da bioinformática.

#### **14.2.1 Aspectos metodológicos da atividade didática**

Para exemplificar tal fato, foi explorado o conhecimento de 13 licenciandos em Ciências Biológicas, do curso de licenciatura em Ciências Biológicas, da Universidade Federal de Campina Grande – UFCG, campus Patos, PB acerca de conceitos básicos relacionados à bioinformática como uma ferramenta didática para o ensino da biologia. Os alunos foram amostrados de turmas que já assistiram aula da disciplina Genética Molecular. Os dados foram coletados como parte das atividades do componente curricular Genética Molecular, ministrado pelo professor da disciplina, que informou aos alunos, que os dados seriam utilizados em pesquisa, que a participação deles seria voluntária e que suas informações seriam utilizadas de forma sigilosa e anônima. Foram amostrados alunos das turmas 2020.2, 2021.2 e 2022.2. O critério de exclusão da amostragem foram os(as) alunos(as) não terem respondido as perguntas e/ou não ter aceito participar da

pesquisa, enquanto que os que aceitaram de forma livre, esclarecida e voluntária, foram incluídos na amostra .

Num primeiro momento, foi realizada uma aula expositiva sobre Bioinformática, na qual foi utilizado o site NCBI para transmitir informações aos entrevistados sobre o assunto. Este site apresenta um banco de dados de taxonomia, e é o repositório padrão de nomenclatura e classificação para o International Nucleotide Sequence Database Collaboration (INSDC), compreendendo os bancos de dados GenBank, ENA (EMBL) e DDBJ. Nele estão incluídos os nomes de organismos e linhagens taxonômicas para cada uma das sequências representadas nos bancos de dados de sequências de nucleotídeos e proteínas.

Parte deste artigo foi desenvolvido durante a disciplina de Bioinformática, ministrada como disciplina optativa no curso de graduação de Licenciatura em Ciências Biológicas, na Universidade Federal de Campina Grande- UFCG, no Campus de Patos, PB.

Para exemplificar como elucidar a origem do vírus SARS-COV2 (Coronavírus) , sua evolução (mutações, adaptações, etc.), são utilizados métodos que permitem entender as relações de parentesco entre organismos a partir de características encontradas no material genético dos organismos analisados. Desta forma, ao final da análise, é possível obter uma árvore filogenética, que é uma representação esquemática das relações de parentesco entre os taxa (linhagens taxonômicas), a partir dos critérios previamente utilizados.

Para realizar a análise filogenética molecular é necessário realizar as seguintes operações:

- 1) Obtenção das sequências ou genoma completo do vírus
- 2) Alinhamento das sequências;
- 3) Escolha de um método de substituição de nucleotídeos ou aminoácidos adequado;

4) Escolha do método de reconstrução filogenética

5) Construção da Árvore filogenética

### 14.2.2 1) Obtenção das sequências

Para este projeto, bastante simplificado, foram utilizadas cinco amostras de genomas completos de vírus retiradas do GenBank, para entender as relações entre genomas dos vírus coletados de diferentes amostras.

Os identificadores dos vírus, depositados no GenBank, foram: MN908947 (novo coronavírus “SARS-COV2” em humano oriundo de Wuhan), MN996532 (coronavírus de morcego), KY417146 (vírus relacionado à SARS de morcego), MT084071 (vírus relacionado à SARS de pangolim) e NC\_038294.1 (vírus relacionado a MERS-COV, que infecta humanos, proveniente de camelo).

#### 1.1) Procedimento:

a) Na página do NCBI ou GENBANK, identifique a palavra Acession, localizada na aba lateral esquerda da página. Para acesso ao campo de busca, clique no sinal de +, localizado à direita no nome Acession.

b) coloque os nomes dos identificadores dos genomas da seguinte forma: “MN908947, MN996532, KY417146, MT084071, NC\_038294.1” e, em seguida, clique no botão azul com a palavra submit. (Figura 14.1).

Refine Results		Reset
Virus		+
Accession		-
MN908947, MN996532, KY417146, MT084071, NC_038294.1		
<input type="button" value="Submit"/>		
Sequence Length		+

Nucleotide (10,487,008)		Protein (46,996,530)		RefSeq Genome (11,598)		Select Columns
Accession	Submitters	Organization	Level	Release Date	Isolate	Species
NC_067209	Shkoporov,A...	National Center for Biotech...	2022-10-21			Bohivirus oralis
NC_067210	Shkoporov,A...	National Center for Biotech...	2022-10-21			Aforbivirus intestinalis
NC_067211	Shkoporov,A...	National Center for Biotech...	2022-10-21			Buchavirus intestinalis
NC_067212	Shkoporov,A...	National Center for Biotech...	2022-10-21			Biripivirus hominis
NC_067213	Shkoporov,A...	National Center for Biotech...	2022-10-21			Blohavirus faecalis
NC_067214	Shkoporov,A...	National Center for Biotech...	2022-10-21			Buchavirus oralis
NC_067215	Shkoporov,A...	National Center for Biotech...	2022-10-21			Burzaovirus intestinalis

Figura 14.1: Acesso aos genomas no NCBI virus. FONTE: NCBI, 2023.

A partir deste momento, após clicar em submit, a página deve mudar e ficar com a aparência da figura 14.2, com apenas as cinco sequências selecionadas dispostas na tela.

The screenshot shows the NCBI Virus search interface. On the left, there is a sidebar with 'Refine Results' filters: Virus (+), Accession (+), Results limited to selected accessions (with a close button), Sequence Length (+), Ambiguous Characters (+), Sequence Type (+), RefSeq Genome Completeness (+), and Nucleotide Completeness (+). The main area displays a table titled 'Nucleotide (5)' with columns: Accession, Submitters, Organization, Release Date, Isolate, Species, and Molecule type. Five rows are listed, each corresponding to a selected accession: NC\_038294, MT084071, MN996532, MN908947, and KY417146. The 'Species' column indicates they are all 'Severe acute respiratory syndrome coronavirus 2'. The 'Molecule type' column shows 'ssRNA(+)'. A 'Select Columns' button is at the top right of the table. A 'Feedback' button is located on the right side of the page.

Figura 14.2: Genomas selecionados no NCBI virus. FONTE: NCBI, 2023.

### 14.2.3 2) Alinhamento das sequências

Para alinhar as sequências, deve-se clicar no primeiro quadrado branco da tabela ao lado da palavra accession em preto e depois clicar no quadro branco com o nome align em azul, localizado no canto superior direito da página.

Será aberta uma nova aba no navegador e será realizado o alinhamento múltiplo das sequências. Cujo resultado deve ser igual ao da Figura 14.3.

The screenshot shows the 'Multiple Alignment' tool on the NCBI Virus website. At the top, there is a header with the NIH logo, National Library of Medicine, National Center for Biotechnology Information, Log in, and Contact Us. Below the header, the 'NCBI Virus' logo and 'Sequences for discovery' are displayed. The main area shows a multiple sequence alignment of five genomes. The x-axis represents the sequence position from 0 to 32,079 bases. The y-axis lists the sequences: NC\_038294.1, MT084071.1, MN996532.1, MN908947.3, and KY417146.1. The alignment shows the conservation of nucleotides across the genome. A legend on the right identifies the colors for different organisms: Bat coronavirus England 1 (red), Bat SARS-like coronavirus (blue), Pangolin coronavirus (green), Bat coronavirus SarTGE13 (orange), and Severe acute respiratory syndrome coronavirus 2 (yellow). A 'Feedback' button is located on the right side of the page.

Figura 14.3: Resultado do alinhamento múltiplo das sequências utilizadas. FONTE: NCBI, 2023.

As etapas, 3, 4 e 5 seguintes serão realizadas voltando-se a aba inicial e clicando no quadro branco com o nome Build Phylogenetic Tree em azul, localizado no canto superior direito da página.

Então será gerada uma nova aba com a árvore filogenética de acordo com a Figura 14.4.

Neste segundo momento, foi realizado um Blast no site do National Center for Biotechnology Information (NCBI) ou GenBank, onde foram colocadas as cinco amostras dos vírus, com o objetivo de obter o alinhamento de sequências dos mesmos e por fim, resultar em uma árvore filogenética (Figura 14.4).

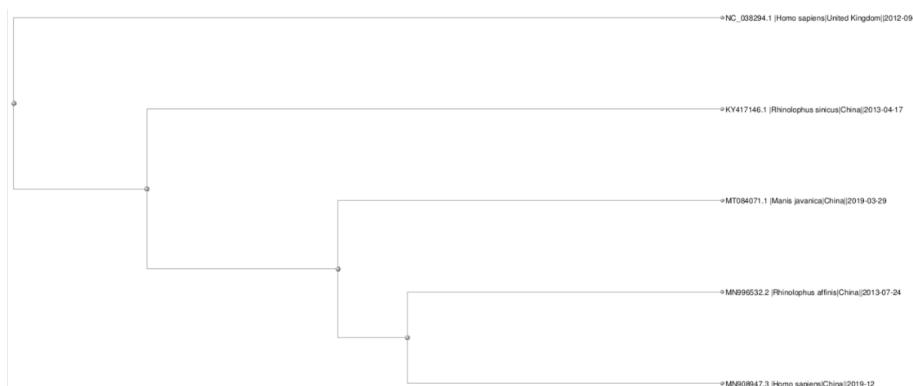


Figura 14.4: Árvore Filogenética gerada. FONTE: NCBI, 2023.

Esta árvore simplificada indica que o novo coronavírus (SARSCoV-2) é mais parecido com um coronavírus de morcego, e ambos são derivados do coronavírus de um pangolin. Isto sugere que o vírus provavelmente foi transmitido a partir de animais como morcegos ou pangolins. Este grupo tem uma relação próxima com outra espécie de morcego e que é proximamente relacionado com o vírus merscov, que infectou os humanos em 2012 e foi transmitido pelo camelo.

Após a finalização da aula expositiva, foi aplicado um questionário via Google Forms, com as seguintes perguntas: (1) O que é bioinformática?; (2) Você tinha conhecimento do Banco de Dados do *National Center for Biotechnology Information* (NCBI)?; (3) O que é Genoma?; (4) O que é Alinhamento de Sequências?; (5) O que é Filogenia Molecular?

Deste modo foram utilizadas quatro perguntas abertas e uma de múltipla escolha, selecionadas de acordo com o assunto da aula expositiva, a fim de medir os conhecimentos dos licenciandos acerca da temática.

*Tabela 14.1: Conceito de Bioinformática. FONTE: Autores, 2023.*

#	RESPOSTAS DOS PARTICIPANTES	Total
1	Biologia aplicada à informática ou tecnologia	2
2	A bioinformática é a junção da biologia e informática para a análise de dados biológicos, através de programas.	6
3	Ciência responsável pelo trabalho de informação de análises gênicas	4
4	Não sei	1

### 14.3 Resultados e discussão

Neste artigo, buscamos compreender como é a relação de conhecimento dos discentes com a bioinformática, e por conta disso, julgamos necessário investigar qual a concepção dos alunos, acerca do conceito de bioinformática. Visto que a Bioinformática pode ser definida como uma área de saber interdisciplinar, que tem por objetivo investigar e elaborar sistemas que colaborem para a compreensão do fluxo de informações de genes ou até estruturas moleculares[5]. E além disso, pode abranger outras vertentes, que envolvem o uso de banco de dados online, a genômica, proteômica e até os sistemas biológicos. Foi observado que os participantes responderam a pergunta sobre o conceito de bioinformática de três formas diferentes (Tabela 14.1).

A maioria dos alunos definiu a bioinformática de forma correta como sendo “A bioinformática é a junção da biologia e informática para a análise de dados biológicos, através de programas.”

**R10: É uma mescla entre ciências, que usa os dados biológicos aplicados a estatísticas, programas e informática para processar e analisar estes dados.**

Outros participantes definiram de maneira parcialmente correta como “biologia aplicada à informática ou tecnologia”

**R1: Biologia aplicada à informática**

**R3: Biologia interligada com a tecnologia**

Já outros definiram o conceito de bioinformática de forma incorreta “Ciência responsável pelo trabalho de informação de análises gênicas”

## R5: Ciência responsável pelo trabalho de informação de análises gênicas

Percebe-se que 46,16% dos alunos entrevistados conseguiram responder de forma correta à pergunta proposta, porém 30,77% responderam de forma parcialmente correta em relação ao conceito de “bioinformática”, pois só responderam que é a junção da biologia com a informática, porém os mesmos não souberam definir além disso. 15,38% responderam de forma incorreta, associando a bioinformática ao conceito de genômica, que é a ciência que trabalha com a análise de informações gênicas, e somente 7,69% não souberam responder.

O site do NCBI foi utilizado para a execução de uma aula expositiva inicial sobre Bioinformática. Após esta aula foi inserida uma pergunta de múltipla com o intuito de averiguar o conhecimento dos participantes no que diz respeito ao conceito desta área da ciência, e obtemos as seguintes respostas (Figura 14.5).

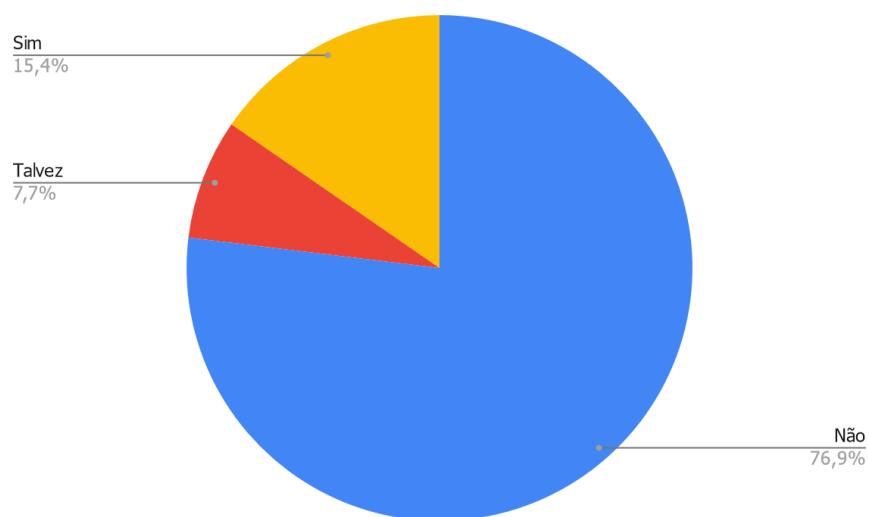


Figura 14.5: Respostas da pergunta: se os participantes conheciam o NCBI. FONTE: Autores, 2023.

Como resultado foi percebido que 76,9% dos participantes não conheciam o NCBI, e os mesmos não conheciam sua importância para a bioinformática e nunca tiveram contato com ele. Os 15,4% que conheciam, sabiam da existência do NCBI, mas nunca chegaram a utilizá-lo.

Tabela 14.2: O que é Genoma. Fonte: Autores, 2023.

#	RESPOSTAS DOS PARTICIPANTES	Total
1	Conjunto de todos os genes de uma espécie de ser vivo	10
2	São as características herdadas dos nossos pais	1
3	É o conjunto haplóide de cromossomos de uma espécie	1
4	Não sei	1

Também foi perguntado aos participantes, se eles detinham conhecimento sobre a definição de genoma, tendo em vista que o mesmo pode ser analisado na bioinformática, a partir de algoritmos computacionais para sua montagem e anotação [8]. As respostas dos alunos foram as seguintes, conforme a Tabela 14.2.

A maioria dos alunos respondeu corretamente, definindo genoma como: “Conjunto de todos os genes de uma espécie de ser vivo”.

**R8: “O Genoma são todos os conjuntos de genes presente em um organismo”.**

Outras pessoas responderam de forma incorreta, pois relacionaram genoma a “características herdadas dos pais”, ou “conjunto haplóide de cromossomo de uma espécie”

**R3: “São as características herdadas dos nossos pais”.**

**R6: “É o conjunto haplóide de cromossomos de uma espécie”.**

O curso de licenciatura em ciências biológicas, abrange disciplinas que envolvem conceitos relacionados à genética, como, por exemplo, o genoma. Por conta disso, 76,93% dos discentes responderam corretamente à pergunta e apenas 23,07% responderam de forma errada e apenas 7,69% não souberam responder.

Na aula expositiva, foi feito o alinhamento do genoma das 5 amostras de vírus, para assim resultar na árvore filogenética. Com base nisso, foi abordado se os discentes tinham conhecimento sobre a definição de alinhamento de sequências (Tabela 14.3).

*Tabela 14.3: O que é alinhamento de sequências. Fonte: Autores, 2023*

#	RESPOSTAS DOS PARTICIPANTES	Total
1	Processo de comparar duas sequências (de nucleotídeos ou proteínas) de forma a se observar seu nível de identidade.	4
2	Sequências de nosso genoma	1
3	Alinhamento e isolamento de duas sequências	3
4	Não sei	5

Menos da metade dos participantes responderam corretamente, afirmado que "é o processo de comparar duas sequências (de nucleotídeos ou proteínas) de forma a se observar seu nível de identidade"

**R6: "Alinhamento de sequências é uma forma de dispor as sequências de DNA, RNA ou proteínas, e tem o objetivo de identificar regiões similares que estão envolvidas em questões evolutivas".**

Poucas pessoas responderam parcialmente correto, dizendo que "é o Alinhamento e isolamento de duas sequências". A Ferramenta Básica de Pesquisa de Alinhamento Local (BLAST) encontra regiões de similaridade local entre sequências. O programa compara sequências de nucleotídeos ou proteínas com bancos de dados de sequências e calcula a significância estatística das correspondências. O BLAST pode ser usado para inferir relações funcionais e evolutivas entre sequências, bem como ajudar a identificar membros de famílias de genes [9].

**R7: "Não sei, acredito que esteja relacionado com o alinhamento de sequências do material genético".**

Apenas um participante respondeu de forma totalmente errada, afirmando que alinhamento de sequências são "Sequências de nosso genoma"

**R3: "São as sequências dos nossos genomas".**

Muitas pessoas relataram que desconheciam sobre o alinhamento de sequências e quando feito na aula expositiva, tiveram contato pela primeira vez com o mesmo. Por conta disso, observamos que apenas 30,77% souberam definir

*Tabela 14.4: O que é Filogenia Molecular. Fonte: Autores, 2023.*

	<b>RESPOSTAS DOS PARTICIPANTES</b>	<b>Total</b>
1	Analisa as diferenças moleculares hereditárias, principalmente em sequências de DNA	9
2	Não sei	4

corretamente, 23,08% responderam parcialmente correto, 38,46% não sabiam e apenas 7,69% responderam incorretamente.

Por fim, quando realizado o alinhamento de sequências, com a ferramenta Blast e demais disponíveis no site, foi obtida uma árvore filogenética, para informar o grau de similaridade entre os vírus. Visando isso, foi questionado se os alunos saberiam definir o conceito de Filogenia Molecular, e foram obtidas as seguintes respostas (Tabela 4).

A maioria dos participantes, responderam de forma correta, dizendo que” a filogenia molecular analisa as diferenças moleculares hereditárias, principalmente em sequências de DNA”.

**R1: “Uma forma recente de identificar a história evolutiva dos organismos com base na biologia molecular. Muito importante na construção de árvores filogenéticas utilizando genes e não só características morfológicas na construção de uma linha evolutiva.”**

Percebe-se que 69,23% dos participantes responderam de forma correta sobre a definição de filogenia molecular, pois os mesmos já tiveram contato com o termo, durante a graduação, porém, não sabiam que através de programas da bioinformática poderiam fazer essa filogenia molecular. 30,77% responderam que não sabiam, ou não lembraram.

A genética é uma área da biologia que é citada por muitos alunos de ensino médio como difícil, pois sentem dificuldades em assimilar seus conceitos em consequência da variedade de termos científicos [2]. Por conta disso, muitas vezes, a mesma acaba sendo taxada como uma disciplina “sem sentido”, o que gera uma certa dificuldade para os professores de ciências trabalharem com ela, em sala de aula.

Além disso, assuntos como filogenia, muitas vezes são repassados de forma inerte e descontextualizada, tornando a interpretação reducionista. Portanto, ferramentas como a bioinformática, contribuem para o processo ensino-aprendizagem, pois ao inserir dados e gerar as árvores de forma rápida, acaba tornando o ensino dinâmico e significativo [5].

Foi denotado que os licenciandos analisados não sabiam que a bioinformática continha um banco de dados, que poderia ser utilizado para ministrar aulas, seja por a mesma ainda ser considerada uma ciência nova, ou por não ser uma disciplina de caráter obrigatório, no curso de ciências biológicas da instituição em que eles estão inseridos.

#### 14.4 Conclusão

A bioinformática como ferramenta didática, se apresenta como um método eficaz para a contribuição do ensino de genética, pois a mesma, aborda uma área que por muitos é considerada desconexa com a realidade, por conta de sua complexidade. Visando isso, a bioinformática vem sanar essa lacuna e aparecer como um instrumento valioso para o ensino. Observa-se, no entanto, que, os docentes, geralmente, não apresentam treinamento necessário para trabalhar com a mesma. E a difusão deste conhecimento pode auxiliar a popularizar ainda mais o uso da bioinformática.

Saiba mais 14.1

Este artigo está disponível em <https://bioinfo.com.br/bioinformatica-como-uma-ferramenta-didatica-para-o-ensino-da-genetica/>

#### 14.5 Referências

- [1] LOPES, S. M. C. . Genetics Education in High School: challenges and new perspectives for quality of learning. Research, Society and Development, [S. l.], v. 12, n. 1, p. e7912139422, 2023. DOI: 10.33448/rsd-v12i1.39422. Disponível em: <https://rsdjurnal.org/index.php/rsd/article/view/39422>. Acesso em: 21 aug. 2023.

- [2] ARAUJO, Maurício dos Santos; FREITAS, Wanderson Lopes dos Santos; LIMA, Sintiane Maria de Sá; LIMA, Mara de Oliveira. A genética no contexto de sala de aula: dificuldades e desafios em uma escola pública de Floriano-PI. REnCiMa, v. 9, n.1, p. 19-30, 2018.
- [3] HAGEN, J.B. The origins of bioinformatics. *Nature Reviews Genetics*. Londres, Nature. v.1, p.231–236. dez. 2000. Disponível em: <https://doi.org/10.1038/35042090>. Acesso em: 21 jan. 2023.
- [4] ATWOOD, T.K; BLACKFORD, S.; BRAZAS, M.D.; DAVIES, A.; SCHNEIDER, M.V.A. Global Perspective on Evolving Bioinformatics and Data Science Training Needs. *Briefings in Bioinformatics*, Londres, OUP. v.20, n.2, p. 398–404, mar. 2019. Disponível em: <https://doi.org/10.1093/bib/bbx100>. Acesso em: 15 jan. 2023.
- [5] MENDES, Anna Carolina de Oliveira, RAMOS, Amanda Perse da Silva, BARBOSA, Luiz Miguel Viana, OLIVEIRA, Maria de Fátima Alves de. OLATCG: ferramenta de bioinformática para o ensino de genética no ensino médio. REAMEC – Rede Amazônica de Educação em Ciências e Matemática. Cuiabá, v. 10, n., 3, e22061, set./dez. 2022. Disponível em: <http://dx.doi.org/10.26571/reamec.v10i3.13954>. Acesso em: 05 fev. 2023.
- [6] SOUSA, F. B.; PEDRO, A. N., ARAÚJO, B. N. e COUTINHO, T.J.D. Potencialidades do uso da bioinformática como ferramenta de ensino. *Bioinfo*. 2022. Disponível em: <https://bioinfo.com.br/potencialidades-do-uso-da-bioinformatica-como-ferramenta-de-ensino/> Acesso em: 31 jul. de 2023.
- [7] MOARES, Isabelle de Oliveira. Bioinformática no ensino de Biologia: revisão bibliográfica e concepção de educadores. Monografia de especialização em Ensino de Ciências e Biologia. Colégio Pedro II. Rio de Janeiro, 2019.
- [8] FERREIRA, Mauricio Alexander de Moura Ferreira. Bioinformática como ferramenta no melhoramento genético de plantas. Trabalho de conclusão de curso para obtenção do grau de Bacharel em Ciências Biológicas. Vitória, 2017.
- [9] Disponível em: <https://blast.ncbi.nlm.nih.gov/Blast.cgi>. Acesso em 22 Ago 2023.

# 15 LÁ E DE VOLTA OUTRA VEZ: UM POUCO DO PASSADO, PRESENTE E FUTURO DA EVOLUÇÃO

Autores 15.1

Tiago Cabral Borelli 

Revisão: Filipe Teixeira 

Cite este artigo 15.1

Borelli, TC. **Lá e de volta outra vez: um pouco do passado, presente e futuro da evolução.**  
BIOINFO. ISSN: 2764-8273. Vol. 3. p.15 (2023). doi: 10.51780/bioinfo-03-15

### Resumo 15.1

#### Opiniões & Perspectivas

**E** se pudéssemos rebobinar a fita da evolução? Se reiniciarmos a história da terra a partir de um certo instante no passado, a vida se desenvolveria da mesma forma? Se você, assim como eu, é biólogo, então sabe que essas questões propostas por Stephen Jay Gould causam longas discussões durante as aulas de evolução. Há quem diga que sim pois as condições seriam as mesmas, há aqueles cuja opinião é fruto de efeito borboleta (sim, o filme) e argumentam que existe um componente caótico na evolução. “Bastaria uma mutação diferente e as mudanças se acumulariam até resultar em formas de vidas desconhecidas por nós”, um aluno poderia argumentar. De qualquer forma, até pouco tempo atrás esse assunto pertencia à teoria ou seria usado como um recurso didático. Afinal, não é possível um experimento para acompanhar a evolução, certo?

Errado! Bom, pelo menos em parte. O que Gould e outros biólogos do século XIX talvez nem puderam sonhar é que atualmente somos capazes de observar a evolução de uma forma tão íntima ao ponto de rastrearmos mutações genéticas no decorrer das gerações. Com experimentos chamados de ALE (Adaptive Laboratory Evolution ou Evolução Adaptativa em Laboratório) é possível expor bactérias à diversas condições e entender quais mutações foram vantajosas graças às análises bioinformáticas de sequenciamento genômico. Além disso, o mesmo experimento pode ser repetido várias e várias vezes. Ou seja, é possível rebobinar a fita da evolução! Os ALEs também são usados para evolução paralela, experimento no qual populações são desafiadas num mesmo tipo de condição em busca de padrões de seleção na evolução bacteriana.

Até agora você já entendeu o que podemos fazer com experimentos em laboratório. Porém, e se eu te contar que a bioinformática e a computação voltada à biologia podem nos levar ainda mais longe? E se for possível enxergar o(s) futuro(s)? Eu sei! Essa pergunta deixa espantando até o biólogo menos ortodoxo. Contudo, essa ideia tem se fixado na cabeça de evolucionistas que se esforçam para entender como a Teoria de Darwin explica o mundo natural. A revisão de Michael

Lässig chamada “Predicting Evolution” mostra os principais conceitos sobre esse tema de pesquisa [1].

Um dos alvos da predição evolutiva é a resistência antimicrobiana em bactérias. Em 2021, um artigo mostrou que a resistência antimicrobiana ou bacteriana matou mais pessoas que a AIDS e a Malária [2]. Então, entender como ela evolui e encontrar padrões nesse processo é uma questão fundamental. É onde eu e você entramos, jovem bioinformata! Modelos de inteligência artificial [3], modelos matemáticos e paisagens evolutivas são exemplos do que podemos usar para analisar a quantidade massiva de dados genômicos disponíveis.

Apesar dos grandes avanços e das mudanças que fizemos ainda existe muito a ser estudado. E eu espero que isso não o desanime, pois precisamos de pessoas interessadas que consigam conectar os campos experimental e computacional. Então, conte com a Revista BIOINFO para se manter em dia e estudando sobre as novidades da bioinformática. Afinal, quem sabe o que existe no futuro? Até a próxima!

#### Saiba mais 15.1

Este artigo está disponível em <https://bioinfo.com.br/la-e-de-volta-outra-vez-um-pouco-do-passado-presente-e-futuro-da-evolucao/>

## 15.1 Referências

[1] Lässig, M., Mustonen, V. Walczak, A. Predicting evolution. *Nat Ecol Evol* 1, 0077 (2017). <https://doi.org/10.1038/s41559-017-0077>

[2] Murray et al. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. In: *The Lancet*. 2022. [https://doi.org/10.1016/S0140-6736\(21\)02724-0](https://doi.org/10.1016/S0140-6736(21)02724-0)

[3] Xavier, Joicy. Inteligência Artificial aplicada à Bioinformática. In: BIOINFO – Revista Brasileira de Bioinformática. Vol. 1. Julho, 2021. DOI:10.51780/978-6-599-27532-6-09

# 16 DE ONDE VÊM AS PROTEÍNAS?

## Autores 16.1

Alisson Clementino da Silva , Bruno Rafael Pereira Nunes , Joicymara Xavier 

Revisão: Marcos Antonio Nobrega de Sousa , Wylerson Nogueira 

## Cite este artigo 16.1

Silva, AC; Nunes, BRP; Xavier, J. **De onde vêm as proteínas?** BIOINFO. ISSN: 2764-8273. Vol. 3. p.16 (2023). doi: 10.51780/bioinfo-03-16

### Resumo 16.1

Proteínas são as mais abundantes moléculas dos seres vivos. Correspondem a 50% da massa seca das células e são um dos componentes básicos da dieta dos organismos, juntamente com os lipídeos e carboidratos. Elas podem apresentar formas e tamanhos diferentes, enquanto algumas são longas e fibrosas, outras são finas e fibrosas ou globulares. Uma célula pode conter centenas de diversas proteínas trabalhando ao mesmo tempo em funções completamente distintas [1-2].

## 16.1 Afinal, o que são proteínas?

Proteínas são polímeros produzidos no interior das células dos seres vivos e que norteiam praticamente todas as suas reações fisiológicas. Atuam como catalisadores de reações químicas, compõem os tecidos e músculos, mediam transporte de substâncias e outras moléculas essenciais, participam ativamente da sinalização celular e ainda estão presentes no seu próprio processo de síntese [1]. Elas são formadas por unidades menores chamadas de aminoácidos (Figura 1), que são moléculas orgânicas que apresentam uma região comum denominada de esqueleto peptídico, formado por radicais orgânicos que se ligam ao mesmo átomo de carbono e diferem-se a partir da cadeia lateral que lhes confere características diferentes. Para que uma proteína seja formada, uma sequência de aminoácidos é conectada por uma interação química chamada ligação peptídica (Figura 16.1).

Na ligação peptídica, a carboxila ( $\text{COOH}^-$ ) de um aminoácido e o grupo amino ( $\text{NH}_3^+$ ) do aminoácido adjacente reagem, para formar um grupo amida que compõe as cadeias polipeptídicas. O processo tem como resultante a interação dos aminoácidos, que passam a ser resíduos por se apresentarem diferentes da sua forma livre, e moléculas de água [1-2].

O genoma humano codifica um total de 20 aminoácidos para construir proteínas, que podem ser classificados como essenciais, quando o próprio

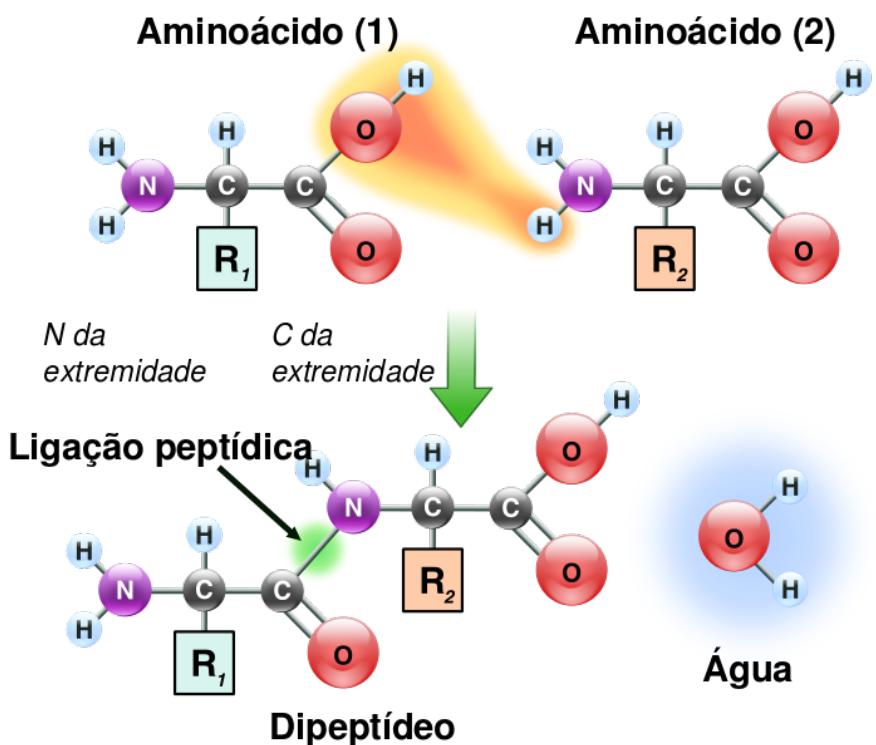


Figura 16.1: Esqueleto peptídico dos aminoácidos. O carbono alfa (centro), ligado à carboxila ( $\text{COOH}-$ ), a um grupo amino ( $\text{NH}_3^+$ ), a um Hidrogênio (H) e a um radical que varia para cada aminoácido. Fonte: Khemis (CC-BY 4.0) [3].

organismo tem a capacidade de produzi-los, e não essenciais, quando precisam ser obtidos via nutrição [2]. O entendimento sobre aminoácidos e ligações peptídicas são a base para desvendar a origem das proteínas de tamanhos e formas variadas. Nas sessões seguintes exploraremos a síntese dessas estruturas e as diferentes formas assumidas por elas.

## 16.2 Síntese de proteínas

A síntese é um processo de alta complexidade biológica que necessita de um maquinário especializado (proteínas específicas, ácidos nucleicos, ribossomos, aminoácidos, etc). O DNA, que contém a informação para a produção da proteína, tem uma fração sua (um gene) transcrita para RNA, que ao migrar do núcleo para o citoplasma é lido por um complexo ribossomal que traduz a informação genética [1-2].

O processamento da informação na síntese pode ser dividido em duas etapas: a primeira, chamada de transcrição (Figura 16.2), inicia-se no interior celular, no caso de organismos procariontes. Nos eucariontes, o processo começa no núcleo celular, com a ação da enzima RNA polimerase. Essa enzima se liga à região promotora, que antecede o gene de interesse, para a produção de uma fita de RNA mensageiro a partir da fita de DNA (sentido 3'-5'). O RNA mensageiro é sintetizado considerando a mesma regra de pareamento de bases que o DNA, mas adota um diferencial alterando as ligações Adenina – Timina por Adenina – Uracila. O processo ocorre até que um códon de terminação (U-A-A, U-A-G ou U-G-A) seja compreendido pela enzima, indicando o fim do gene e da etapa transcrecional [4].

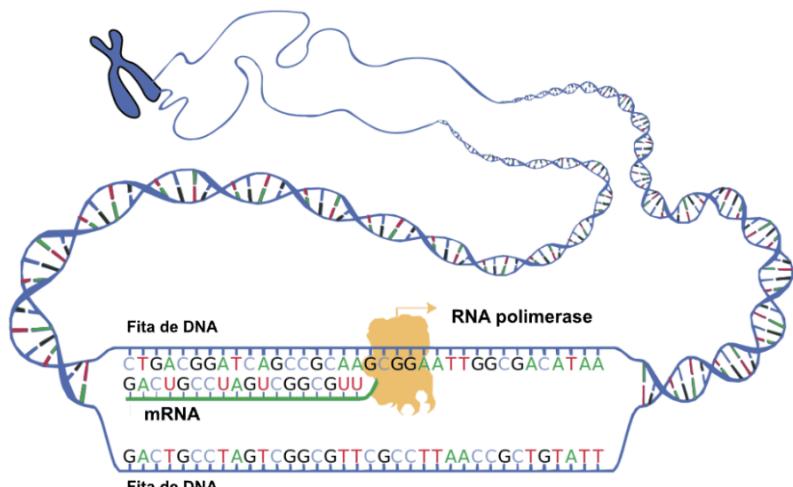


Figura 16.2: Etapa de transcrição do DNA. A dupla fita de DNA é desdoblada por ação da RNA polimerase, permitindo a transcrição da informação genética para RNA mensageiro. Fonte: Sulai (Domínio público) [5].

O pré-mRNA permanece no núcleo celular dos eucariontes para que seja submetido a um tipo de curadoria biológica que se chama splicing, onde porções transcritas e não codificantes são retiradas da sequência [4]. Duas modificações importantes ocorrem na nova fita pela adição de grupos químicos em suas extremidades. A cap 5', que é uma guanina modificada (G), é adicionada na extremidade 5' para impedir rompimentos por ação de fosfatases e nucleases e, mais tarde, mediar a ligação do complexo enzimático que codifica a proteína no

citoplasma, enquanto uma sequência de aproximadamente 200 adeninas (cauda poli-A) é adicionada na extremidade 3' para proteger o transcrito e conferir maior estabilidade à molécula [1-2].

Com a retirada dos íntrons, e demais modificações, o mRNA se torna legível para os ribossomos, que consideram que a cada três nucleotídeos (um códon), um aminoácido seja adicionado à sequência (Figura 16.3). Assim, a leitura é baseada nos 61 códons do DNA que podem ser transcritos para codificar os aminoácidos necessários à proteína. Essa relação é chamada de código genético [1].

				Segunda letra			
		U	C	A	G		
Primeira letra	U	UUU } Phe UUC } UUA } Leu UUG }	UCU } UCC } Ser UCA } UCG }	UAU } Tyr UAC } UAA Parada UAG Parada	UGU } Cys UGC } UGA Parada UGG Trp	U	C A G
	C	CUU } CUC } Leu CUA } CUG }	CCU } CCC } CCA } Pro CCG }	CAU } His CAC } CAA } Gln CAG }	CGU } CGC } Arg CGA } CGG }	U C A G	U C A G
	A	AUU } AUC } Ile AUA } <b>AUG Met</b>	ACU } ACC } ACA } Thr ACG }	AAU } Asn AAC } AAA } Lys AAG }	AGU } Ser AGC } AGA } Arg AGG }	U C A G	U C A G
	G	GUU } GUC } Val GUA } GUG }	GCU } GCC } GCA } Ala GCG }	GAU } Asp GAC } GAA } Glu GAG }	GGU } GGC } Gly GGA } GGG }	U C A G	U C A G

Figura 16.3: O código genético. Tabela do código genético que relaciona os códons aos respectivos aminoácidos. Fonte: traduzido de Nirenberg/Khorana [6].

A segunda etapa, conhecida como tradução, ocorre com a saída do RNA mensageiro do núcleo para o citoplasma, e conta com a ação dos outros dois tipos de RNA: ribossômico e transportador, para que a síntese seja possível.

Ao chegar no citoplasma, dos eucariontes, a fita de RNA mensageiro se junta ao ribossomo, que inicia a leitura a partir do códon A-U-G, no sentido 5'-3'. O complexo ribossomal, media a ligação entre as moléculas de RNA transportador,

que carregam novos aminoácidos, com a fita de RNA mensageiro, permitindo a interação dos códons com os anticódons (Figura 16.4) [1-4].

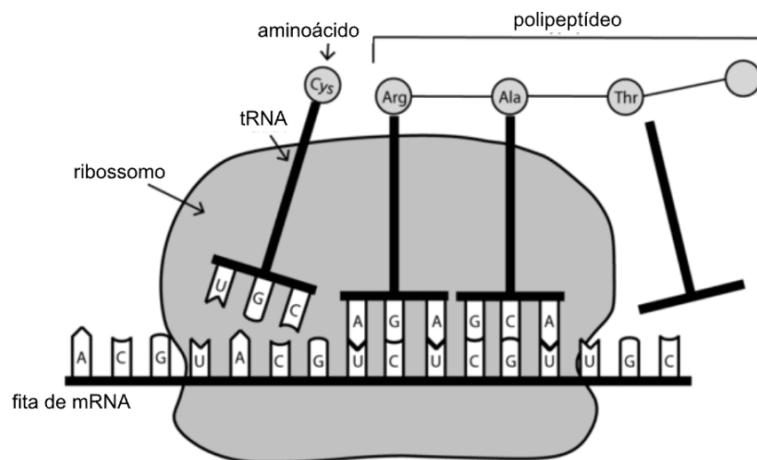


Figura 16.4: Etapa de tradução. O complexo ribossomal se junta a fita de RNA mensageiro e permite a ligação do RNA transportador, carregando aminoácidos, pelo pareamento dos anticódons a seus respectivos códons. Fonte: traduzido de Sarah Greenwood (CC-BY 4.0) [7].

Após a metionina, um segundo tRNA é recebido no sítio ribossomal, carregando o próximo aminoácido que será adicionado à sequência. A ligação peptídica é estabelecida entre os dois e o primeiro RNA transportador é liberado, fazendo o segundo ser transferido para outro sítio e o ribossomo se deslocar na fita do mRNA para receber o próximo. Dessa forma, a fita é percorrida mediante a leitura e o processo continua até que um códon de terminação seja reconhecido pelo complexo para que haja a liberação da proteína, a separação do ribossomo e o fim da etapa de tradução [2-4].

### 16.3 Estruturas de proteínas

Ao serem sintetizadas, as proteínas passam a ser classificadas quanto ao nível conformacional adquirido (Figura 16.5), uma vez que a conformação tridimensional, se refere às diversas formas que elas podem assumir devido suas ligações, variando em complexidade e funções específicas [8].

A estrutura primária, como pode ser visto na Figura 16.5, é o nível mais básico de organização estrutural das proteínas, caracterizada por uma sequência de aminoácidos residuais que é quimicamente estabilizada pela ligação peptídica. Essa sequência proteica é determinada pelo gene que a codifica. Uma modificação substitutiva ou deletéria nos resíduos pode gerar mudanças conformacionais, resultando em desdobramentos parciais e perda da função biológica [1-8].

Arranjos formados pelo dobramento da cadeia principal da estrutura primária, constituem a estrutura secundária (Figura 16.5). Dentre os dobramentos mais recorrentes estão as alfa-hélices, conformações e voltas beta, que apresentam-se nas formas helicoidal, pregueada e espiral randômica, respectivamente. Essa classificação se refere à disposição espacial dos resíduos de aminoácidos, e consequentemente, a qualquer segmento da cadeia polipeptídica, sem considerar a posição das cadeias laterais e relações com outras partes [8-9]. Existem ainda estruturas secundárias que fogem a regra, não apresentando um padrão definido, impossibilitando a descrição adequada dos segmentos por apresentarem aleatoriedade espacial [1].

A estrutura terciária (Figura 16.5), é conhecida como forma nativa da proteína, considera o arranjo tridimensional total dos átomos que a compõem e resulta principalmente das interações entre radicais dos aminoácidos [8]. Os resíduos que estavam distantes na sequência polipeptídica ou em diferentes estruturas secundárias, com o dobramento da proteína, podem interagir com base em seus grupos carregados, mantendo suas posições terciárias por diferentes tipos de interações. Com forças de repulsão e atração, os grupos radicais se estabilizam espacialmente por ligações não covalentes, como hidrogênio, iônicas, interações dipolo-dipolo e forças de dispersão de London. As interações hidrofóbicas são importantes para a estabilidade da proteína, bem como as ligações dissulfeto, que ocorrem entre cadeias de cisteína contendo enxofre, unindo segmentos distintos [1-9].

Arranjos formados por proteínas de nível terciário, constituem as estruturas quaternárias, que são duas ou várias proteínas aglomeradas, por forças eletrostáticas, para formação de um complexo funcional maior [8-9], como

pode ser visto ainda na Figura 16.5. Nesse quarto e último nível de organização, elas podem ser divididas em: proteínas fibrosas; proteínas globulares; proteínas intrinsecamente desordenadas e proteínas de membrana [10].

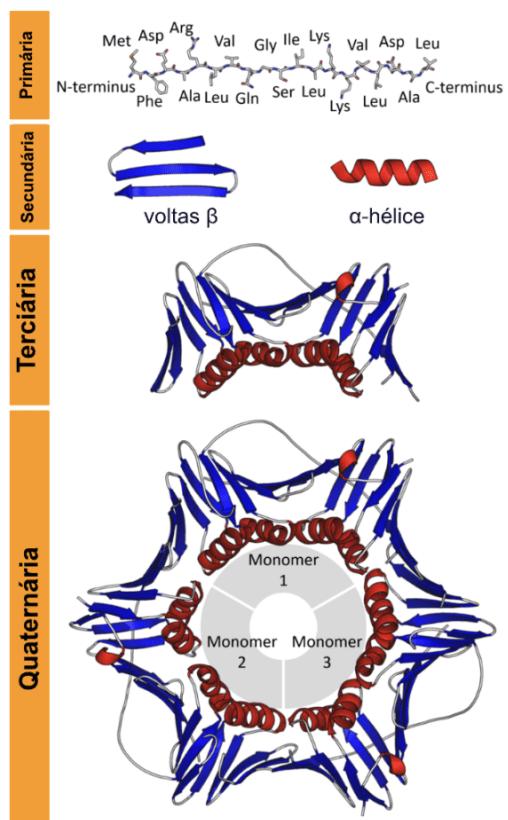


Figura 16.5: Níveis de organização das cadeias polipeptídicas. Estrutura primária representada por uma sequência de resíduos de aminoácido; Estrutura secundária representada pelos arranjos de alfa-hélice e folha-beta; Estrutura terciária: representada por um conjunto de arranjos secundários; Estrutura quaternária: formada por um conjunto de estruturas terciárias. Fonte: Thomas Shafee (CC-BY 4.0) [11].

Ao assumir sua estrutura nativa, as proteínas tornam-se funcionais e participam de diversos processos fisiológicos. Suas funções, estabilidade, expressão e solubilidade são propriedades amplamente estudadas a partir da identificação da sequência primária e da análise do enovelamento da cadeia polipeptídica [8-10]. Elas também apresentam padrões identificáveis no

enovelamento, que constituem os motivos e domínios, que são comumente utilizados para entender estruturas ainda não resolvidas.

## 16.4 Conclusão

Este artigo apresenta resumidamente uma resposta à pergunta sobre a origem das proteínas. A partir desse conhecimento, podemos explorar outros processos complexos, como o surgimento de modificações pós-traducionais e a dinâmica do próprio dobramento das cadeias proteicas. Mas isso é um assunto para os próximos artigos.

Saiba mais 16.1

Este artigo está disponível em <https://bioinfo.com.br/de-onde-vem-as-proteinas/>

## 16.5 Referências

- [1] Nelson, D. L., M. Cox. Princípios de Bioquímica. 2. ed. São Paulo: Sarvier, 2000.
- [2] Alberts, Bruce; et al. Biologia Molecular da Célula. 5.ed. Porto Alegre: Artmed, 2010.
- [3] Khemis. (CC-BY 4.0). Disponível em:  
[https://pt.m.wikipedia.org/wiki/Ficheiro:Peptidformationball\\_pt\\_BR.svg](https://pt.m.wikipedia.org/wiki/Ficheiro:Peptidformationball_pt_BR.svg). Acesso em 18 de Agosto de 2023.
- [4] Berg, J.M.; Tymoczko, J.L.; Stryer, L. Bioquímica. 6º ed. Rio de Janeiro: Guanabara Koogan, 2008.
- [5] Sulai. Domínio público. Dispovível em:  
[https://commons.wikimedia.org/wiki/File:DNA\\_transcription.svg](https://commons.wikimedia.org/wiki/File:DNA_transcription.svg). Acesso em 18 de Agosto de 2023.
- [6] Nirenberg/Khorana: Breaking the genetic code. (s.d.). Obtido em  
<http://www.mhhe.com/biosci/genbio/raven6b/graphics/raven06b/howscientiststhink/14-lab.pdf>. Acesso em 18 de Agosto de 2023.
- [7] Sarah Greenwood (CC-BY 4.0). Disponível em:  
[https://commons.wikimedia.org/wiki/File:Protein\\_Synthesis-Translation.png](https://commons.wikimedia.org/wiki/File:Protein_Synthesis-Translation.png). Acesso em 18 de Agosto de 2023.
- [8] Verli, H. et al. Bioinformática da Biologia à flexibilidade molecular. Porto Alegre, 2014.

[9] Zaha, A.; Ferreira, H. B.; Passaglia, L. M. P; Biologia Molecular Básica. 5<sup>a</sup> edição. Porto Alegre: Artmed, 2014.

[10] Lemos, R. P.; Santos, P. H.; Rocha, A. Introdução à Biologia Estrutural de Proteínas. In: Revista BIONFO. Disponível em:  
<https://bioinfo.com.br/introducao-a-biologia-estrutural-de-proteinas>. Vol. 3. 2023.  
Acesso em 14 de Agosto de 2023.

[11] Thomas Shafee (CC-BY 4.0). Disponível em:  
[https://commons.wikimedia.org/wiki/File:Protein\\_structure\\_\(full\).png](https://commons.wikimedia.org/wiki/File:Protein_structure_(full).png). Acesso em 18 de Agosto de 2023.

# 17

## VÍRUS ENDÓGENOS HUMANOS: COMO ANALISÁ-LOS *in silico*?

### Autores 17.1

Juan Diego Cipriano Ramalho Sampaio , João Gonçalves da Costa Neto , Isaac Farias Cansanção 

Revisão: Bruna Espiño dos Santos , Diego Lucas Neres Rodrigues 

### Cite este artigo 17.1

Sampaio, JDCR; Neto, JGC; Cansanção, IF. **Vírus endógenos humanos: como analisá-los *in silico*?** BIOINFO. ISSN: 2764-8273. Vol. 3. p.17 (2023). doi: 10.51780/bioinfo-03-17

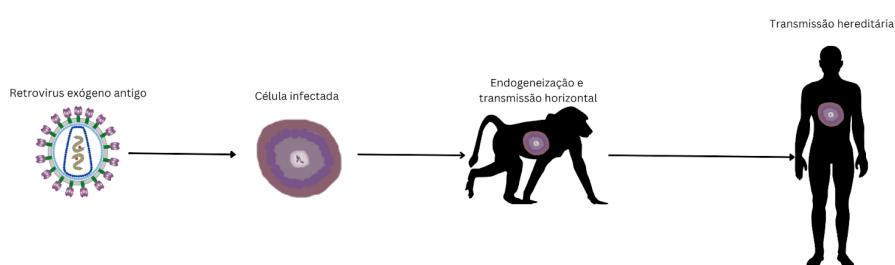
### Resumo 17.1

Os retrovírus endógenos humanos (HERVs) são importantes constituintes do DNA humano e estão associados a diversas características genotípicas e fenotípicas, desde a prevenção de doenças até a ancestralidade entre espécies. Com o tempo, esses vírus tornaram-se alvo de várias pesquisas pelo mundo, fazendo-se uso de métodos laboratoriais e de análises *in silico* computacionais. Neste estudo, fez-se uma pesquisa focada na utilização de ferramentas virtuais voltadas para a avaliação de vírus endógenos. Após a pesquisa das bases de dados existentes, com ênfase nas metodologias usadas pelos artigos para análise genômica, percebeu-se a relevância destes mecanismos computacionais na análise de fragmentos de ácidos nucleicos (primers) previamente sequenciados e disponíveis em bancos de dados, como o NCBI/BLAST, para a identificação de HERVs e sua relação com infecções virais atuais, como a COVID-19. Notou-se também: a utilização do tBLASTn para screening de sequências base de cromossomos encontradas no NCBI e posterior comparação com endovírus Mavericks; o uso de programas para análise e visualização de polimorfismo no vírus HERV-K; o uso na prevenção de rotas metabólicas com maior eficiência, possibilitando sua adesão na virologia. Os testes *in silico* demonstram grande potencial para agilizar processos e permitir análises de extensas sequências de HERVs.

## 17.1 Introdução

O processo de evolução das espécies envolve diversos fatores externos responsáveis por mudanças na expressão e conformação genética. Os vírus endógenos são um desses contribuintes de alterações, sendo caracterizados como sequências de genes pertencentes a vírus originalmente exógenos (vírus que infectam as células do hospedeiro por meios externos) que foram incorporados ao genoma da espécie, entrando em relativo equilíbrio com o hospedeiro [1].

Em seres humanos, os retrovírus endógenos (HERVs) são as formas mais comuns dessas sequências, totalizando 8% do genoma humano, e foram adquiridos através da inserção de DNA retroviral em células germinativas de espécies ancestrais, permitindo a hereditariedade desses trechos. Devido à pressão seletiva cumulativa diferenciada em relação aos retrovírus exógenos, os ERVs se distanciaram geneticamente dos vírus originais.



*Figura 17.1: Esquema do processo de endogeneização de vírus exógenos antigos, resultando em HERVs.*

**Fonte:** traduzido de RANGEL, S. et al. *Human endogenous retroviruses and the inflammatory response: a vicious circle associated with health and illness*. *Frontiers In Immunology*, [S.L.], v. 13, p. 1-14, 23 nov. 2022.

O interesse no estudo dos ERVs revelou grande potencial na expressão de características, especialmente no entendimento de resistência a certas doenças virais. Ao longo dos anos, a análise de ERVs mostrou-se cada vez mais eficiente com a utilização de ferramentas de sequenciamento capazes de avaliar características genotípicas e a história evolutiva viral, podendo ser realizada através de bancos de dados publicamente disponíveis, como o HERVd. Nesse artigo, busca-se elucidar os processos envolvidos na análise computacional (*in silico*) dos ERVs e quais são suas características de destaque.

## 17.2 Metodologia

O estudo trata-se de um artigo original sobre o uso de análises *in silico* em vírus endógenos humanos. Foram consultados artigos originais publicados nas línguas portuguesa e inglesa, abrangendo os períodos de publicação entre 1990 a abril de 2023 em 2 (duas) bases de dados: PubMed e Google Scholar. Foram excluídos capítulos de livros, artigos de revisão, textos jornalísticos e artigos pagos. Ao final do processo de busca, 12 artigos foram selecionados e utilizados no estudo.

### 17.3 Resultados e Discussão

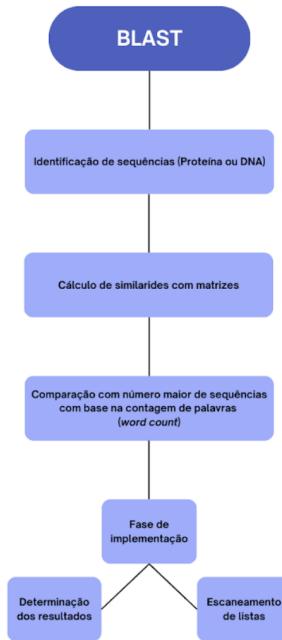
Várias das análises in silico são feitas utilizando programas feitos em linguagens de programação, principalmente R e Java, desenvolvendo scripts próprios para propósitos daquela pesquisa ou usando ferramentas online disponíveis gratuitamente, como o BLAST (sigla em inglês, *Basic Local Alignment Search Tool*). Por se tratar de estudos do genoma humano, muitas vezes as amostras são adquiridas por métodos labororiais. Porém, a existência de bancos de dados como o NCBI permite a realização de análises computacionais mais específicas. Os códigos de muitas dessas ferramentas estão disponíveis para download em sites como GitHub (<https://github.com/>), permitindo a simulação dos resultados por outros usuários e oferecendo maior dinamicidade à pesquisa.

Uma das possibilidades encontradas com o estudo de vírus endógenos in silico é a descoberta de ancestralidade comum a outras espécies. A DNA polimerase  $\beta$  (polB) de uma espécie de tartaruga-de-caixa foi utilizada como uma sonda para rastrear o genoma humano, tornando possível identificar seu parentesco com sequências de cromossomo 7 e 8 humanos, através da utilização do software tBLASTn, que realizou a leitura de genes codificadores de proteínas como marcos genéticos e comparou às sequências disponíveis no repositório genômico do NCBI de outras espécies de vertebrados. Há também o Galaxy, uma plataforma de análise utilizada na manipulação de arquivos FASTA das sequências encontradas, e o MAFFT, para agrupar os sub-alinhamentos dos cromossomos. Com isso, foram encontradas evidências, na grande maioria dos mamíferos placentários, de genes ortólogos Mavericks, ou seja, sequências genéticas provindas de integrações feitas por vírus endógenos (ou endovírus) Mavericks que apresentam semelhança com o genoma de espécies oriundos de um mesmo ancestral comum, baseado nas integrações feitas por vírus Mavericks que circulavam pelo DNA de mamíferos ancestrais há pelo menos 102 milhões de anos, fazendo, assim, parte do processo evolutivo desses animais. [2]

É válido ressaltar também a aplicação conjunta de vários programas com o intuito de aprimorar o processo de pesquisa e também de permitir o estudo mais aprofundado sobre o papel dos vírus endógenos, como é visto na análise

filogenética. Nesse sentido, ferramentas como o BLAST se mostram muito importantes para esse fim. Após a identificação de sequências de proteína ou de DNA, é realizado um cálculo de similaridades entre as regiões de sequências distintas que resulta em um valor para um par de segmentos, que pode ser feito através de uma matriz específica, como a PAM-120 e a BLOSUM-45. Em seguida, é realizada essa comparação com um número bem maior de sequências, sendo considerado aqueles segmentos associados com a proteína produzida, processo esse acelerado ao adotar um limite de contagem de palavras relacionadas com o segmento, eliminando sequências que não satisfazem o valor fixo. Por fim, na fase de implementação, são escaneadas as listas com todas as sequências e o resultado são daquelas que se mostraram mais próximas do valor de tolerância determinado, leituras essas que podem chegar a até 500.000 resíduos por segundo. O esquema abaixo (figura abaixo) mostra de forma simplificada o processo realizado pelo BLAST. Para acessar a integração de genomas de outras espécies, pode ser feita a busca por genes mais próximos, utilizando a ferramenta de buscador de genoma, como o Ensembl, que permite a visualização de dados genômicos, como predição de genes e escores de conservação de vertebrados, desde bases únicas até cromossomos completos, com uma vasta livraria providenciada pela “EMBL’s – European Bioinformatics Institute”. Com isso, foi possível determinar as regiões ortólogas dos vírus Maverick em relação a outros mamíferos, incluindo detalhes sobre os locais de início de leitura para produção proteica [2-4].

Além disso, os mecanismos computacionais de análise podem contribuir em pesquisas para identificação e estudo de doenças humanas, como a esclerose múltipla, sendo utilizados para facilitar a investigação aprofundada de famílias específicas de provírus, que se tratam da forma de DNA dos retrovírus incorporados ao material genético do hospedeiro [5, 6]. Um exemplo da aplicação destes mecanismos foi observado em estudo feito para o grupo de vírus endógenos HERV-K, o mais jovem entre os retrovírus endógenos humanos e o único polimórfico, no qual foi desenvolvida uma plataforma computacional capaz de analisar subsequências do genoma que continham o provírus HERV-K polimórfico e comparou com a base de dados do projeto de mil genomas KGP, que consistia em sequências genéticas obtidas de 26 populações de um grupo de 5 super populações representantes do mundo. As sequências de referência, obtidas



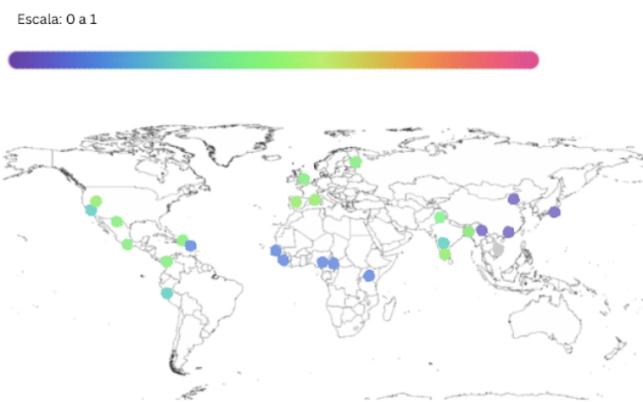
*Figura 17.2: Esquema demonstrando as etapas básicas do BLAST. Fonte: autoria própria.*

por sequenciamento de genoma completo, foram lidas utilizando conjuntos de subsequências de nucleotídeos de  $k$  elementos ( $k$ -mers) únicas para cada locus de HERV-K. Após o mapeamento, são comparados os  $k$ -mers do objeto de estudo ( $n$ ) e os  $k$ -mers de referência ( $T$ ), resultando em uma razão  $n/T$  que varia de 0 (alelo ausente) a 1 (alelo presente). Essa separação em  $k$ -mers permitiu uma análise mais eficiente dos dados, juntamente da utilização de um modelo de mistura de clusters baseado no valor de  $n/T$ . Foi também desenvolvida uma ferramenta de visualização feita em Java (D3.js), que permite a investigação dos alelos polimórficos do HERV-K em populações humanas, o que pode ser útil no estudo de condições clínicas. Os métodos utilizados permitiram a criação de um programa robusto que irá auxiliar na comparação de múltiplos sítios em populações diversificadas e facilitar pesquisas futuras sobre o provírus [5].

O uso de programas de visualização contribui para facilitar o entendimento sobre os vírus endógenos de forma mais concreta. Nesse sentido, o D3 é de grande ajuda, tratando-se de uma ferramenta de código aberto na qual

os usuários conseguem vincular valores a elementos arbitrários que permitem dinamizar e modificar o conteúdo a ser mostrado. Este utiliza como base o Modelo de Documento Objeto (DOM) e aceita a implementação de módulos que aumentam suas capacidades. Com a possibilidade de receber arquivos como JSON adaptados para pesquisa de bioinformática, o D3 transforma arquivos em uma interface gráfica visual com possibilidade de animações e interações diversas, além de boa compatibilidade com navegadores. O D3, que pode ser instalado como uma livraria do javascript, pode ser usado, por exemplo, para receber dados sobre sequências genéticas e representar sua prevalência em um mapa geográfico baseado na sua localização (figura a seguir), mostrando seu potencial de praticidade [5, 7].

### Prevalência na população



*Figura 17.3: Visualização da co-ocorrência de genes HERV-K polimórficos em 26 populações em uma representação baseada na localização geográfica, utilizando a ferramenta D3. A prevalência relativa entre os genes foi representada no programa com um gradiente de cores, que, então, preenche as bolhas dos diversos pontos do mapa. Fonte: Adaptado de LI, Weiling; LIN, Lin; MALHOTRA, Raunaq; YANG, Lei; ACHARYA, Raj; POSS, Mary. A computational framework to assess genome-wide distribution of polymorphic human endogenous retrovirus-K In human populations. Plos Computational Biology, [S.L.], v. 15, n. 3, p. 1-21, 28 mar. 2019.*

Além disso, a bioinformática pode ser utilizada para investigar a influência dos endovírus em doenças específicas. Para isso, é necessário quantificar sua presença em amostras, utilizando um processo que consiste em duas etapas: identificação e análise. Na etapa de identificação, é necessário determinar sequências genômicas correspondentes aos endovírus, para que estas sejam utilizadas como modelos

e comparadas com os resultados das amostras. Estes modelos estão presentes em bancos de dados, como o Gypsy 2.0 (pertencente à NCBI build 37.p13), no qual podem ser compiladas diversas sequências genômicas com fragmentos detectáveis de vírus endógenos [6, 8].

Durante a etapa de análise, os achados presentes nos bancos de dados são comparados com os obtidos nas amostras, a partir de programas computacionais. Em um processo denominado Alinhamento Múltiplo de Sequências, softwares como o Clustal Omega são capazes de alinhar e organizar até 190.000 sequências em poucas horas, utilizando os modelos prévios como base para detectar os endovírus presentes em cada amostra. Ao final do sequenciamento, os resultados são dispostos em árvores filogenéticas [6, 9].

Um exemplo promissor desse uso da bioinformática é o alinhamento múltiplo de sequências de proteínas GAG e ENV dos HERVs em amostras encefálicas. Esse mecanismo foi utilizado em um estudo que buscou entender a relação entre a presença de HERVs no tecido nervoso e a incidência de esclerose múltipla. Com a utilização de PCR em tempo real, 28 sequências de HERV GAG e 88 sequências de HERV ENV foram alinhadas a partir de 42 amostras retiradas de cérebros congelados (33 com esclerose múltipla e 9 do grupo controle). Em seguida, o alinhamento múltiplo de sequências foi realizado pelo programa Clustal Omega, que considerou para cada análise o maior trecho de DNA que pode ser traduzido em uma proteína. Posteriormente, as sequências foram organizadas em táxons e sua presença foi quantificada nas amostras. Os resultados indicaram que as amostras com a doença apresentaram maior expressão das proteínas supracitadas, especialmente em algumas famílias de endovírus, como a HERV-E e a HERV-K. Contudo, a diferença relativa entre o grupo de estudo e o grupo controle não foi grande o suficiente para determinar causalidade entre a presença de endovírus e a esclerose múltipla [6].

## 17.4 Conclusão

Logo, é possível observar que a bioinformática se mostra como um pool de ferramentas fundamentais para o estudo de vírus endógenos, por meio das

análises *in silico*, caracterizadas pelo uso de simulações computacionais na interpretação de amostras [10]. A relevância destes mecanismos no estudo com vírus endógenos se encontra na análise de fragmentos de ácidos nucleicos (primers) previamente sequenciados e disponíveis em bancos de dados como o NCBI/BLAST [6, 11], com diversas finalidades, como o estudo filogenético e a investigação da relação entre famílias de endovírus e diversas doenças [2, 5, 6, 12]. Esse processo ocorre em decorrência da integração entre a capacidade de cálculo fornecida pelas tecnologias atuais e os conhecimentos preexistentes em biotecnologia, que permite o estudo de grandes volumes de dados e a previsão de rotas metabólicas com maior eficiência, o que possibilita sua adesão na virologia, área que estuda os vírus e suas propriedades [12].

#### Saiba mais 17.1

Este artigo está disponível em <https://bioinfo.com.br/virus-endogenos-humanos-como-analisa-los-in-silico/>

## 17.5 Referências

- [1] JOHNSON, Welkin E. Endogenous Retroviruses in the Genomics Era. *Annual Review of Virology*, [S.L.], v. 2, n. 1, p. 135-159, 9 nov. 2015. Annual Reviews. <http://dx.doi.org/10.1146/annurev-virology-100114-054945>.
- [2] BARREAT, Jose Gabriel Nino; KATZOURAKIS, Aris. Evolutionary Analysis of Placental Orthologues Reveals Two Ancient DNA Virus Integrations. *Journal Of Virology*, [S.L.], v. 96, n. 22, p. 1-11, 23 nov. 2022. American Society for Microbiology. <http://dx.doi.org/10.1128/jvi.00933-22>.
- [3] ALTSCHUL, Stephen F; GISH, Warren; MILLER, Webb; MYERS, Eugene W; LIPMAN, David J. Basic local alignment search tool. *Journal Of Molecular Biology*, [S.L.], v. 215, n. 3, p. 403-410, out. 1990. Elsevier BV. [http://dx.doi.org/10.1016/s0022-2836\(05\)80360-2](http://dx.doi.org/10.1016/s0022-2836(05)80360-2).
- [4] CUNNINGHAM, Fiona; ALLEN, James e; ALLEN, Jamie; ALVAREZ-JARRETA, Jorge; AMODE, M Ridwan; ARMEAN, Irina M; AUSTINE-ORIMOLOYE, Olanrewaju; AZOV, Andrey G; BARNES, If; BENNETT, Ruth. Ensembl 2022. *Nucleic Acids Research*, [S.L.], v. 50, n. 1, p. 988-995, 17 nov. 2021. Oxford University Press (OUP). <http://dx.doi.org/10.1093/nar/gkab1049>.
- [5]: LI, Weiling; LIN, Lin; MALHOTRA, Raunaq; YANG, Lei; ACHARYA, Raj; POSS, Mary. A computational framework to assess genome-wide distribution of polymorphic human

endogenous retrovirus-K In human populations. Plos Computational Biology, [S.L.], v. 15, n. 3, p. 1-21, 28 mar. 2019. Public Library of Science (PLoS).  
<http://dx.doi.org/10.1371/journal.pcbi.1006564>.

[6] PJ, Bhetariya. Analysis of Human Endogenous Retrovirus Expression in Multiple Sclerosis Plaques. Journal Of Emerging Diseases and Virology, [S.L.], v. 3, n. 2, p. 1-17, 2017. Sci Forschen, Inc. <http://dx.doi.org/10.16966/2473-1846.133>.

[7] BOSTOCK, M.; OGIEVETSKY, V.; HEER, J. D3 Data-Driven Documents. IEEE Trans Vis Comput Graph. 2011; 17: 2301–2309.

[8] LLORENS, C.; FUTAMI, R.; COVELLI, L.; DOMINGUEZ-ESCRIBA, L.; VIU, J. M.; TAMARIT, D.; AGUILAR-RODRIGUEZ, J.; VICENTE-RIPOLLES, M.; FUSTER, G.; BERNET, G. P. The Gypsy Database (GyDB) of mobile genetic elements: release 2.0. Nucleic Acids Research, [S.L.], v. 39, n., p. 70-74, 29 out. 2010. Oxford University Press (OUP).  
<http://dx.doi.org/10.1093/nar/gkq1061>.

[9] SIEVERS, Fabian; WILM, Andreas; DINEEN, David; GIBSON, Toby J; KARPLUS, Kevin; LI, Weizhong; LOPEZ, Rodrigo; MCWILLIAM, Hamish; REMMERT, Michael; SÖDING, Johannes. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular Systems Biology, [S.L.], v. 7, n. 1, p. 1-6, jan. 2011. EMBO.  
<http://dx.doi.org/10.1038/msb.2011.75>.

[10] EKINS, S; MESTRES, J; TESTA, B. In silico pharmacology for drug discovery: methods for virtual ligand screening and profiling. British Journal of Pharmacology, [S.L.], v. 152, n. 1, p. 9-20, set. 2007. Wiley. <http://dx.doi.org/10.1038/sj.bjp.0707305>.

[11] TEMEROZO, Jairo R.; FINTELMAN-RODRIGUES, Natalia; SANTOS, Monique Cristina dos; HOTTZ, Eugenio D.; SACRAMENTO, Carolina Q.; SILVA, Aline de Paula Dias da; MANDACARU, Samuel Coelho; MORAES, Emilly Caroline dos Santos; TRUGILHO, Monique R. O.; GESTO, João S. M. Human endogenous retrovirus K in the respiratory tract is associated with COVID-19 physiopathology. Microbiome, [S.L.], v. 10, n. 1, p. 1-15, 22 abr. 2022. Springer Science and Business Media LLC.  
<http://dx.doi.org/10.1186/s40168-022-01260-9>.

[12] SANTOS, C. M.; VEIGA, F. C. C.; DA SILVA, S. L.; DOS REIS, S. P. ANÁLISE IN SILICO E PREDIÇÃO DE EPÍTOPOS DAS VARIANTES DE SARS-CoV-2 COM MAIOR POTENCIAL IMUNOGÊNICO. REVISTA FOCO, [S. l.], v. 16, n. 4, p. e1572, 2023. DOI: 10.54751/revistafoco.v16n4-039. Disponível em:  
<https://ojs.focopublicacoes.com.br/foco/article/view/1572>. Acesso em: 31 jul. 2023.

# 18 A BIOINFORMÁTICA E A COMPREENSÃO DA VIDA

## Autores 18.1

Thiago M. N. de Camargo 

Revisão: Wylerson Nogueira, , Ana Carolina Silva Bulla, 

## Cite este artigo 18.1

Camargo, TMN. **A Bioinformática e a Compreensão da Vida.** BIOINFO. ISSN: 2764-8273.

Vol. 3. p.18 (2023). doi: 10.51780/bioinfo-03-18

### Resumo 18.1

#### Opiniões & Perspectivas

**A** Bioinformática é um campo interdisciplinar que combina a Ciência da Biologia com a Ciência da Computação, a Estatística e outras disciplinas relacionadas [2, 3]. Através da análise de grandes quantidades de dados genômicos, a Bioinformática tem sido fundamental para entender a origem e evolução da vida na Terra [14]. Estudos têm mostrado que as moléculas de RNA podem ter desempenhado um papel crucial nos estágios iniciais da evolução [6], levando à hipótese do Mundo de RNA [4].

Utiliza-se a Bioinformática para diversos fins dentre eles analisar e interpretar informações biológicas, como genomas, proteomas, vias metabólicas e redes regulatórias [7], até simulações de evolução molecular cujo fim pode ser ampliar nossa compreensão das origens e evolução da vida [11, 14].

Considerando a Bioinformática uma ciência e não apenas um conjunto de técnicas, metodologias e afins, sua história recente é marcada não apenas pela descoberta da estrutura do DNA, mas também pelo desenvolvimento de tecnologias de sequenciamento molecular [6], o que possibilitou o seu uso para a leitura de sequências de nucleotídeos que formam o material genético de um organismo, mas também para a própria modelagem de complexos químicos fundamentais, como as proteínas [11].

As tecnologias de sequenciamento de nova geração (NGS) têm gerado uma grande quantidade de dados, sendo necessário que bioinformaticas desenvolvam ferramentas e métodos computacionais que lidem de maneira cada vez mais eficiente com esses dados [8].

A Bioinformática também é empregada em pesquisas médicas, contribuindo para a identificação de mutações genéticas em doenças hereditárias [8], descoberta de novos medicamentos ou, ainda, para melhor entender a biologia de doenças como Câncer, Parkinson e Alzheimer [2]. Pode ser também empregada

em estudos metagenômicos de ambientes especiais, como cavernas [1] e objetos artificiais, como Minas subterrâneas, abrindo, assim, um campo de possibilidades não apenas para a descoberta de novos fármacos e compostos de interesse industrial, mas a própria compreensão da vida uma vez que esses ambientes possuem baixa concentração de matéria orgânica sendo classificados como oligotróficos. [1, 10]

A partir da análise de grandes quantidades de dados genômicos e da identificação de novos microrganismos e compostos com atividade biotecnológica relevante, essas informações podem ser utilizadas para o desenvolvimento de novos produtos químicos, medicamentos e processos biotecnológicos [2], além de contribuir para a compreensão da ecologia dos microrganismos nesses ambientes [1].

A Bioinformática está em contínuo desenvolvimento, exercendo um importante papel na pesquisa biológica nos setores de economia, biotecnologia e indústria. Com o avanço das tecnologias de sequenciamento e a crescente disponibilidade de dados biológicos, a Bioinformática continuará a ser crucial para a compreensão da biologia dos organismos, a origem e evolução da vida e no desenvolvimento de novas terapias e medicamentos que auxiliarão a nossa persistência sobre o planeta [5].

#### Saiba mais 18.1

Este artigo está disponível em <https://bioinfo.com.br/a-bioinformatica-e-a-compreensao-da-vida/>

### 18.1 Referências

[1] ADETUTU, Eric M.; BALL, Andrew S. Microbial diversity and activity in caves. *Microbiology Australia*, v. 35, n. 4, p. 192-194, 2014.

[2] ATTWOOD, Teresa K. *Introduction to bioinformatics*. Addison-Wesley Longman Limited, 1999.

- [3] BAXEVANIS, Andreas D.; OUELLETTE, Francis. Bioinformatics: A practical guide to the analysis of genes and proteins. John Wiley & Sons. Inc. NY, USA, v. 518, 2001.
- [4] BENNER, Steven A.; ELLINGTON, Andrew D.; TAUER, Andreas. Modern metabolism as a palimpsest of the RNA world. *Proceedings of the National Academy of Sciences*, 1989, 86.18: 7054-7058.
- [5] GAUTHIER, Jeff et al. A brief history of bioinformatics. *Briefings in bioinformatics*, v. 20, n. 6, p. 1981-1996, 2019.
- [6] JOYCE, Gerald F. The antiquity of RNA-based evolution. *Nature*, 2002, 418.6894: 214-221.
- [7] LANDER, E. S. et al. 403 Doyle M, FitzHugh W et al: Initial sequencing and analysis of the human genome. 404. *Nature*, v. 409, n. 6822, p. 860-921, 2001.
- [8] LESK, Arthur M. Introduction to genomics. Oxford University Press, 2017.
- [9] MARRA, Marco et al. A physical map of the human genome. *Nature*, v. 409, n. 6822, p. 934-941, 2001.
- [10] MOLDOVAN, Oana Teodora; KOVÁČ, Ľubomír; HALSE, Stuart (ed.). Cave ecology. 2018.
- [11] NEW, Michael H.; POHORILLE, Andrew. An inherited efficiencies model of non-genomic evolution. *Simulation Practice and Theory*, 2000, 8.1-2: 99-108.
- [12] POHORILLE, Andrew; DEAMER, David. Artificial cells: prospects for biotechnology. *Trends in biotechnology*, 2002, 20.3: 123-128.
- [13] VALENCIA, Alfonso; PAZOS, Florencio. Computational methods for the prediction of protein interactions. *Current opinion in structural biology*, v. 12, n. 3, p. 368-373, 2002.
- [14] WOESE, Carl R.; KANDLER, Otto; WHEELIS, Mark L. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences*, v. 87, n. 12, p. 4576-4579, 1990.

# 19 VISÃO INTEGRATIVA DA BIOLOGIA DE SISTEMAS

Autores 19.1

Vitor Lima Coelho 

Revisão: Wylerson Nogueira, 

Cite este artigo 19.1

Coelho, VL. Visão Integrativa da Biologia de Sistemas. BIOINFO. ISSN: 2764-8273. Vol. 3. p.19 (2023). doi: 10.51780/bioinfo-03-19

### Resumo 19.1

#### Opiniões & Perspectivas

**T**ECNOLOGIAS de alto rendimento são técnicas ou equipamentos que permitem o processamento, geração e análise de grandes quantidades de dados de origem biológica com alto desempenho. Por exemplo, sequenciadores de nova geração de DNA e RNA permite em poucos dias a análise de genomas e transcriptomas completos de organismos procariotos e eucariotos [1]. Espectrômetros de massa de larga escala permitem a identificação e quantificação de proteínas e metabólitos [2]. Técnicas de screening de alto rendimento permitem a testagem de múltiplos alvos e compostos farmacêuticos [3]. Por outro lado, representa também um desafio para processar, integrar e analisar de forma detalhada criando conexões, informações e hipóteses a partir de grandes bases de dados.

O estudo dos componentes ou fenômenos biológicos de forma individualizada permite uma compreensão valiosa, porém isolada, dos componentes celulares, funções moleculares e processos biológicos. Técnicas de análise e mineração de dados, aprendizado de máquina e bioinformática podem ser aplicadas para criar ligações entre essas informações isoladas [4]. Dessa forma, os dados produzidos por tecnologias genômicas, transcriptômicas, proteômicas, lipidômicas e metabolômicas podem ser processados e compilados de modo a construir bases de dados que permitem seus usuários consultar informações biológicas numa visão integrada do sistema complexo que o compõe. Por exemplo, o UniProt (Universal Protein Resource) é uma base de dados de anotações funcionais e sequências de proteínas que integra informações moleculares e estruturais, além de informações genômicas, transcriptômicas, interações com outras moléculas, informações evolutivas e a relação da proteína em questão com doenças humanas [5]. Então com a utilização de técnicas de modelagem e simulação computacional pode-se testar hipóteses e prever comportamentos do sistema usando como parâmetros ou variáveis essas diferentes fontes de informação biológica, incluindo diversas condições celulares ou ambientais,

níveis de expressão de genes em diferentes estados biológicos ou estágio de desenvolvimento, entre outras [6]. Essa abordagem permite testar hipóteses antes de realizar experimentos reais e pode levar a novas descobertas que não seriam possíveis com abordagens mais tradicionais [7].

Além dos evidentes benefícios da visão integrativa da biologia de sistema na pesquisa básica para a compreensão das interações biológicas em sistemas complexos, essa abordagem tem consequências diretas em outros segmentos importantes da sociedade [8-10]. Na medicina, por exemplo, o surgimento da biologia de sistemas como um campo de pesquisa mudou a forma como olhamos para a função fisiológica normal humana e ajudou a descobrir a complexidade de fisiopatologias [11]. Agora os cientistas usam abordagens de biologia de sistemas para entender o quadro geral de como todas as peças interagem em um organismo [12]. Uma compilação com curadoria de fontes de alta qualidade de interações é considerada um recurso primordial no campo da Biologia de Sistemas e, assim, permite uma compreensão mais profunda do cenário mais amplo – seja no nível do organismo, órgão, tecido ou célula – colocando seus componentes juntos [13]. Adquirir tecidos relevantes e/ou fontes de fluidos corporais de coortes de estudo em humanos, por exemplo, pode certamente ser difícil, assim, a biologia de sistemas comparativa pode ajudar a identificar quais organismos podem ser semelhantes o suficiente em cada aspecto para serem usados como modelos.

Investigações baseadas em dados usando abordagens de biologia de sistemas, embora ofereçam visões completas sobre a função dos sistemas biológicos, são limitadas pelo estado de integridade das informações biológicas anteriores.

Saiba mais 19.1

Este artigo está disponível em <https://bioinfo.com.br/visao-integrativa-da-biologia-de-sistemas/>

## 19.1 Referências

- [1] "Overview of High Throughput Sequencing Technologies to ... – NCBI." Overview of High Throughput Sequencing Technologies to Elucidate Molecular Pathways in Cardiovascular Diseases – PMC. Accessed 3 Aug. 2023.
- [2] "Ultra high-throughput mass spec for life sciences research – Nature." Ultra high-throughput mass spec for life sciences research. Accessed 3 Aug. 2023.
- [3] "High Throughput Screening – an overview | ScienceDirect Topics." High Throughput Screening – an overview | ScienceDirect Topics. Accessed 3 Aug. 2023.
- [4] "Methods for biological data integration: perspectives and challenges." 6 Nov. 2015, Methods for biological data integration: perspectives and challenges | Journal of The Royal Society Interface. Accessed 3 Aug. 2023.
- [5] "UniProt." UniProt. Accessed 3 Aug. 2023.
- [6] "Integrative Systems Biology for Data Driven Knowledge Discovery." Integrative Systems Biology for Data Driven Knowledge Discovery – PMC. Accessed 3 Aug. 2023.
- [7] "Promises and Challenges of Systems Biology." 16 Oct. 2020, Promises and Challenges of Systems Biology. Accessed 3 Aug. 2023.
- [8] "Industrial systems biology – Wiley Online Library." 4 Nov. 2009, Industrial systems biology. Accessed 3 Aug. 2023.
- [9] "Systems Biology for Smart Crops and Agricultural Innovation – PubMed." Systems Biology for Smart Crops and Agricultural Innovation: Filling the Gaps between Genotype and Phenotype for Complex Traits Linked with Robust Agricultural Productivity and Sustainability. Accessed 3 Aug. 2023.
- [10] "Systems Biology Approaches for Food and Health | SpringerLink." 1 Sep. 2020, Systems Biology Approaches for Food and Health | SpringerLink. Accessed 3 Aug. 2023.
- [11] "Systems Biology Approaches to Understanding the Human Immune ...." 24 Jun. 2020, Systems Biology Approaches to Understanding the Human Immune System. Accessed 3 Aug. 2023.
- [12] "Systems Medicine: The Application of Systems Biology Approaches ...." 18 Aug. 2015, Systems Medicine: The Application of Systems Biology Approaches for Modern Medical Research and Drug Development – PMC. Accessed 3 Aug. 2023.
- [13] "Systems biology: current status and challenges | SpringerLink." 13 Jan. 2020, Systems biology: current status and challenges | SpringerLink. Accessed 3 Aug. 2023.

# 20

## A IMPORTÂNCIA DA DOCAGEM MOLECULAR NO COMBATE ÀS BACTÉRIAS MULTIRRESISTENTES

Autores 20.1

Aline Sampaio Cremonesi 

Revisão: Wylerson Nogueira 

Cite este artigo 20.1

Cremonesi, AL. **A importância da docagem molecular no combate às bactérias multirresistentes.** BIOINFO. ISSN: 2764-8273. Vol. 3, p.20 (2023). doi:10.51780/bioinfo-03-20

## Resumo 20.1

### Opiniões & Perspectivas

**Q**UE a resistência bacteriana é tida como um problema de saúde pública não é mais uma novidade! Bactérias das mais diversas espécies estão adquirindo resistência aos antibióticos mais potentes que existem. Isso interfere significativamente em diversos aspectos da prática clínica, como a diminuição da eficiência terapêutica de substâncias e a redução no controle de doenças infectocontagiosas, colocando em risco profissionais da área da saúde e pacientes.

Caracteriza-se por resistência bacteriana os mecanismos desenvolvidos por diferentes espécies que buscam reduzir ou eliminar o efeito de agentes antimicrobianos como os antibióticos. Estes mecanismos podem ser a produção de enzimas que degradam ou alteram a molécula antimicrobiana, eliminando seus efeitos nas células, ou ainda, proteínas capazes de formar uma bomba de efluxo, que seria um transportador de membrana capaz de expulsar os antimicrobianos da célula [1]. Além dos mecanismos de resistência, as bactérias têm adquirido cada vez mais habilidades de colonizarem e proliferarem em seus hospedeiros, por meio de melhora na captação de nutrientes e a utilização destes para a produção de biofilme e fatores de virulência [2].

Atualmente, diversas espécies bacterianas têm ampliado a lista de resistência, o que tem gerado o aumento na preocupação por parte dos sistemas de saúde de como controlar e contornar tal problema. Espécies como *Acinetobacter baumanii*, *Escherichia coli*, *Klebsiella pneumoniae*, *Staphylococcus aureus* e *Clostridium difficile* encabeçam esta lista por possuírem cepas capazes de resistir a antibióticos de segunda linha, ou seja, antibióticos específicos utilizados em infecções graves e que geralmente são mais tóxicos ou apresentam efeitos colaterais mais pronunciados do que os de primeira linha [3].

A infectividade e patogênese dessas bactérias têm sido relacionadas com um grupo de transportadores conhecidos como transportadores do tipo ABC (*ATP-Binding Cassete* – Conjunto de proteínas ligadoras de ATP) envolvidos

principalmente com mecanismos de evasão, resistência ao hospedeiro, fatores de superfície celular e excreção, auxílio na captação de nutrientes, entre outros [4; 5]. Transportadores do tipo ABC utilizam a energia liberada pela hidrólise de adenosina trifofato (ATP) para translocar diferentes compostos através das membranas celulares. Isso ocorre devido a estrutura complexa destes transportadores (Figura 20.1) que é formada por duas proteínas ligadoras de nucleotídeos (*Nucleotide Binding Domain* – NBD), também conhecidas como ATPases e duas proteínas transmembrana (Transmembrane Domain – TMD) que formam um canal de passagem através da membrana celular, também conhecidas como permeases. Os transportadores do tipo importadores, ainda contam com uma proteína a mais: a proteína ligadora de substrato (*Substrate Binding Protein* – SBP), que é responsável por identificar e se ligar de forma específica a molécula a ser internalizada. Por se encontrar no periplasma (espaço entre a membrana e a parede celular), esta proteína também é popularmente conhecida como periplasmática [6].

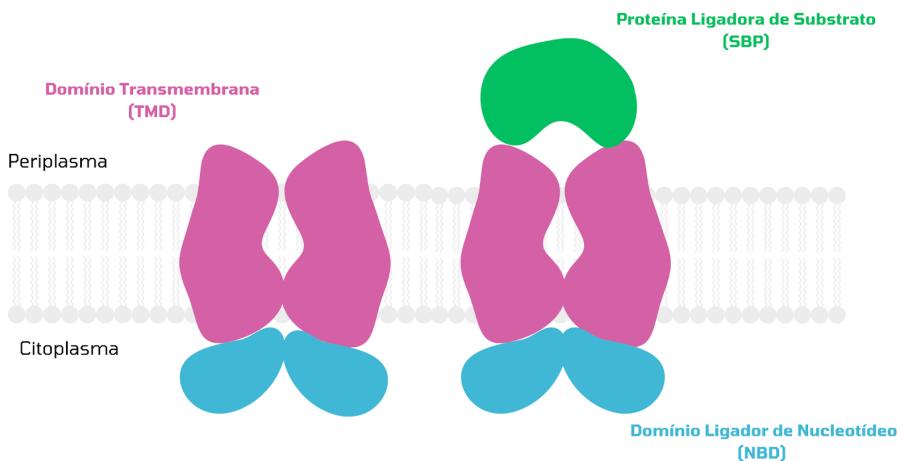


Figura 20.1: Esquema simplificado dos componentes dos transportadores ABC do tipo exportadores e importadores. Fonte: própria autora.

O conhecimento detalhado da estrutura e de etapas do transporte pode fazer dos transportadores ABC novos alvos para a terapêutica antimicrobiana [7] e o desenvolvimento de vacinas [8], principalmente porque os transportadores responsáveis pela internalização de nutrientes para as células são exclusivos de bactérias. Dessa maneira, os antimicrobianos desenvolvidos para inibir a

captação e transporte de nutrientes a partir da ação de transportadores ABC não influenciarão a atividade dos transportadores exportadores encontrados nos eucariotos.

Existem muitos gargalos para o estudo de estruturas e funções de proteínas e um deles é a complexidade da produção destas moléculas em laboratório de forma recombinante e sua manutenção para os ensaios de atividade e inibição. O custo para produzir grande quantidade de uma proteína para testar diferentes inibidores é alto e pode desmotivar muitos pesquisadores. Mas existe um caminho alternativo. As proteínas são formadas por aminoácidos, cuja sequência é determinada pela sequência de nucleotídeos de DNA do indivíduo, assim, as análises destas sequências nos permite saber muito sobre as proteínas, de forma que podemos predizer alguns comportamentos destas moléculas. A forma como as proteínas adquirem sua estrutura é determinada em parte pela sua sequência de aminoácidos, que obedece uma série de leis físicas e químicas para que a molécula fique estável [9]. Hoje já sabemos quais posições os aminoácidos podem adotar ou não em uma estrutura tridimensional protéica e, a partir de então, com o auxílio de algoritmos específicos, podemos determinar computacionalmente a estrutura de uma proteína apenas com a sua sequência linear de aminoácidos. É o que chamamos de análises *in silico*.

E qual a importância disso? Podemos estimar com uma relativa precisão a função de uma proteína a partir da sua estrutura, a partir de simulações de possíveis interações entre esta proteína e diversas moléculas. É possível, por exemplo, criar um modelo tridimensional para testar diversos compostos, a fim identificar aqueles que têm uma maior probabilidade de se ligar aos transportadores ABC de bactérias resistentes, possivelmente podendo causar a sua inibição. Este método é conhecido como docagem molecular. Aqueles compostos que apresentarem os maiores índices de interação serão validados experimentalmente, a fim de determinar sua eficácia como inibidores da proteína alvo. Esta metodologia é interessante pois seleciona quais compostos deverão ser testados experimentalmente, eliminando aqueles que certamente não apresentam características que favorecem a interação, reduzindo assim o número de ensaios que precisarão ser feitos, tempo e recursos.

A aplicação da bioinformática para o desenvolvimento de novos tratamentos contra infecções é muito ampla e crescente, dada a importância do tema na saúde pública [10-11]. Entretanto, muitas outras áreas têm usado destas análises computacionais, na busca por novos alvos terapêuticos para diferentes tipos de câncer, doenças genéticas, terapias gênicas e compreensão e monitoramento de epidemias e surtos de doenças infecciosas. Sendo assim, a bioinformática se mostra essencial para a melhoria da saúde humana, aprimoramento da prática clínica e desenvolvimento de novos fármacos.

Saiba mais 20.1

Este artigo está disponível em <https://bioinfo.com.br/a-importancia-da-docagem-molecular-no-combate-a-bacterias-multirresistentes/>

## 20.1 Referências

- [1] Browne, K.; Chakraborty, S.; Chen, R.; Willcox, M. D.; Black, D. S.; Walsh, W. R.; Kumar, N. A new era of antibiotics: the clinical potential of antimicrobial peptides. *Intern J. Mol. Sci.* 2020, 21(19), 7047.
- [2] SAMPAIO, Aline et al. The periplasmic binding protein NrtT affects xantham gum production and pathogenesis in *Xanthomonas citri*. *FEBS Open bio*, v. 7, n. 10, p. 1499-1514, 2017.
- [3] Prestinaci F, Pezzotti P, Pantosti A. Antimicrobial resistance: a global multifaceted phenomenon. *Pathog Glob Health*. 2015;109(7):309.
- [4] CREMONESI, Aline Sampaio et al. The citrus plant pathogen *Xanthomonas citri* has a dual polyamine-binding protein. *Biochemistry and Biophysics Reports*, v. 28, p. 101171, 2021.
- [5] TER BEEK, Josy; GUSKOV, Albert; SLOTBOOM, Dirk Jan. Structural diversity of ABC transporters. *Journal of General Physiology*, v. 143, n. 4, p. 419-435, 2014.
- [6] LOCHER, Kaspar P. Mechanistic diversity in ATP-binding cassette (ABC) transporters. *Nature structural & molecular biology*, v. 23, n. 6, p. 487-493, 2016.
- [7] COUNAGO, Rafael et al. Prokaryotic substrate-binding proteins as targets for antimicrobial therapies. *Current Drug Targets*, v. 13, n. 11, p. 1400-1410, 2012.

[8] CONVERSO, Thiago Rojas et al. A protein chimera including PspA in fusion with PotD is protective against invasive pneumococcal infection and reduces nasopharyngeal colonization in mice. *Vaccine*, v. 35, n. 38, p. 5140-5147, 2017.

[9] NELSON, David L.; COX, Michael M. *Princípios de bioquímica de Lehninger*. Artmed Editora, 2022.

[10] CHUKWUDOZIE, Onyeka S. et al. The relevance of bioinformatics applications in the discovery of vaccine candidates and potential drugs for COVID-19 treatment. *Bioinformatics and Biology Insights*, v. 15, p. 11779322211002168, 2021.

[11] SAEB, Amr TM. Current Bioinformatics resources in combating infectious diseases. *Bioinformation*, v. 14, n. 1, p. 31, 2018

# 21 RE-DOCKING MOLECULAR UTILIZANDO O PYMOL E AUTO DOCK VINA

Autores 21.1

Luana Luiza Bastos , Giovana Fiorini 

Revisão: Ana Carolina Silva Bulla , Lucianna Helene Santos , Thiago de Camargo 

Cite este artigo 21.1

Bastos, LL; Fiorini, G. **Re-docking Molecular Utilizando o PyMOL e AutoDock VINA.**

BIOINFO. ISSN: 2764-8273. Vol. 3. p.21 (2023). doi: 10.51780/bioinfo-03-21

## Resumo 21.1

O tutorial a seguir aborda a técnica de re-docking utilizada como etapa inicial em simulações de docking molecular para validar a ferramenta utilizada, bem como suas funções de pontuação. O re-docking consiste em separar um complexo proteína-ligante resolvido experimentalmente e buscar encontrar uma conformação parecida através do docking. Neste tutorial utilizaremos o PyMOL como uma ferramenta visual auxiliar para o re-docking utilizando o Vina.

**D**OCKING, também conhecido como ancoragem, atracamento ou, ainda, acoplamento molecular, é um processo computacional que consiste em prever a melhor posição e orientação de um ligante em comparação a outra molécula, resultando em um complexo estável [1]. À medida que mais estruturas de proteínas são determinadas experimentalmente, o docking passa a ser cada vez mais usado como uma ferramenta que auxilia o melhor entendimento das funções e interações das proteínas e sua utilização em diversos campos de pesquisa [2].

Para isso, técnicas de re-docking e cross-docking são utilizadas como uma forma de validar e verificar a precisão das ferramentas e funções de pontuação utilizadas nas simulações. O cross-docking, por exemplo, é executado ao “encaixar” a molécula a um receptor não nativo, mas com estrutura proteica notavelmente similar, avaliando o RMSD entre o ligante resultante e o cristalográfico. No entanto, é importante observar que essa metodologia não será abordada neste tutorial [7].

O Re-docking, também conhecido como Auto-docking, é uma abordagem amplamente utilizada para a avaliação da precisão de um processo de docking em um programa. Como método de validação, seu principal objetivo é recriar a posição original (conformação) do ligante e do receptor quando ligados comparando-a com a estrutura experimentalmente resolvida. Isso tem por objetivo verificar a capacidade do programa em encontrar uma pose (ou, posição, como acima) que se assemelhe ao máximo à conformação experimental [6].

O Re-docking molecular pode ser realizado em quatro etapas principais:

1. Obtenção das estruturas de receptor e ligantes
2. Preparação das estruturas
3. Realização do docking
4. Avaliação dos resultados

Neste artigo, você irá aprender a realizar o re-docking molecular utilizando os programas Vina e o PyMOL. Para isso vamos utilizar algumas ferramentas:

- **PyMOL** – necessário para a preparação da estrutura do receptor e selecionar resíduos de interface.
- **AUTODOCKTOOLS – ADT (MGLTools)** – necessário para a preparação da estrutura do receptor e criação do grid (caixa de execução, ou seja delimitação da região onde o docking será realizado) [3].
- **Vina** – necessário para a realização do docking [4].

Ressaltamos aqui que o tutorial a seguir se mostrou eficiente para os sistemas operacionais GNU/Linux e Windows, no entanto se mostrou ineficiente para MacOS.

## 21.1 Instalando os programas

### 21.1.1 AUTODOCKTOOLS – ADT

Para baixar a ferramenta acesse: <https://ccsb.scripps.edu/mgltools/downloads/>

No site, você irá encontrar versões para os diferentes sistemas operacionais Linux, Mac e Windows.

### Version 1.5.7

platform	installer
	mgltools_win32_1.5.7_Setup.exe (80Mb)
	mgltools-1.5.7-MacOS-X-Install.dmg (GUI installer 91Mb)
	mgltools_1.5.7_MacOS-X.tar.gz (tarball installer 85Mb)
	mgltools_Linux-x86_64_1.5.7_Install (Linux 64 GUI installer 109Mb)
	mgltools_x86_64Linux2_1.5.7.targz (Linux 64 tarball installer 108Mb)

Figura 21.1

Caso o seu sistema operacional seja Linux, após baixar o arquivo será preciso transformá-lo em um executável. Para isso, você deve clicar na tela com o botão direito e abrir o terminal. Após abrir o terminal você deve digitar o comando abaixo, para tornar o programa um executável. Note que essas etapas só devem ser realizadas se seu sistema operacional for Linux.

```
chmod +x ./mgltools_Linux-x86_64 *
```

Depois de digitar o comando, você deve voltar para a pasta onde seu executável se encontra e clicar nele. Clique em “próximo”, aceite os termos e escolha o local onde a pasta será instalada. Após a instalação, uma outra pasta será criada. Para executar o programa, basta abrir a pasta MGLTools-1.5.7, em seguida a pasta Bin, e clicar no ícone “adt”, assim o programa será executado (Figura 21.2).

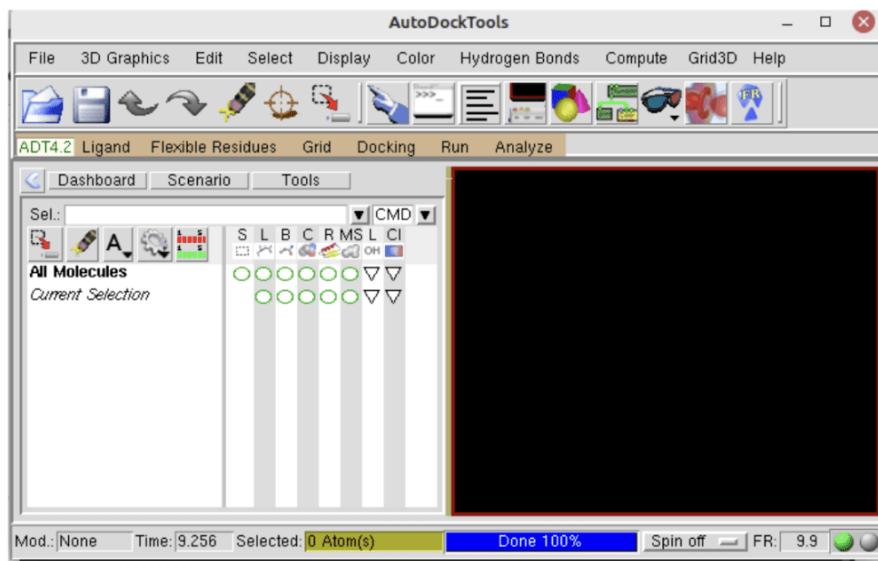


Figura 21.2: Página inicial do AutoDockTools.

**Atenção!** O AutoDockTools também está disponível na loja de aplicativos do Linux mint em sua versão 1.5.7-3 podendo ser instalado da própria loja apenas clicando em “instalar”, sem a necessidade de todos os passos descritos, porém nem sempre a versão da loja estará atualizada.

Para que não seja necessário estar sempre indicando o caminho dos arquivos executáveis no terminal, é possível criar variáveis de ambiente que serão reconhecidas por comandos em terminais abertos por qualquer usuário. Para isso devemos descrever no arquivo PATH quais diretórios serão exportados logo na inicialização do sistema e quais variáveis devem ser declaradas se assim estiverem disponíveis por comando no terminal. Para isso execute no Terminal Linux (Ctrl+T para abrir o Terminal).

```
sudo vim ~/.bashrc
```

Caso não possua o Vim, você pode instalá-lo utilizando o comando:

```
sudo apt-get install vim
```

Depois de instalar, digite o comando anterior para acessar o arquivo. Vá até o final do arquivo e acrescente um comando semelhante ao abaixo:

```
export PATH=\$PATH: ~/MGLTools-1.5.7/bin  
alias pmv=~/MGLTools-1.5.7/bin/pmv  
alias adt=~/MGLTools-1.5.7/bin/adt  
alias vision=~/MGLTools-1.5.7/bin/vision  
alias pythonsh=/*~/MGLTools-1.5.7/bin/pythonsh'
```

Para sair e salvar aperte “*Esc*”, digite “:wq” e aperte “ENTER” ao final – Atualize o arquivo PATH com o comando:

```
source ~/.bashrc
```

### 21.1.2 AUTODOCK VINA

O próximo passo é instalar o autodock Vina que pode ser baixado em: <https://vina.scripps.edu/downloads>.

Ou executando no terminal o comando para *Download*:

```
wget https://vina.scripps.edu/wp-content/uploads/sites/55/2020/12/autodock_vina_1_1_2_linux_x86.tgz
```

Com o gerenciador de arquivos, abra o arquivo compactado *autodock\_vina\_1\_1\_2\_linux\_x86.tgz*. Entre na pasta bin onde estão os executáveis: vina e vina\_split. Faça a extração dos programas para a pasta: /home/MGLTools-1.5.7/bin. É preciso selecionar os dois arquivos para extrair apenas eles e não os outros diretórios. Para saber se foi instalado corretamente execute no terminal:

```
vina --version
```

**Observação:** como esta pasta já está incluída nas variáveis de ambiente PATH, não é necessário realizar uma nova edição do PATH para inclusão do vina.

### 21.1.3 PyMOL

Para obter a versão educacional do PyMOL acesse: <http://pymol.org/edu/>.

Preencha o formulário e em seguida você receberá um email contendo login e senha, o link para baixar o PyMol e o arquivo de licença.

Para instalar a versão *open source* do PyMOL (mais recomendado) podemos acessar o link abaixo:

```
https://github.com/schrodinger/pymol-open-source
```

Após baixar o arquivo, para realizar a instalalção basta digitar o código abaixo:

```
python setup.py install --prefix=~/someplace
```

Você também pode usar o comando abaixo no terminal, se utilizar o Linux, para instalar a versão mais recente:

```
sudo apt install pymol
```

## 21.2 PLUGIN AUTODOCK VINA PYMOL

Para baixar o plugin no PyMOL acesse:

```
https://github.com/Pymol-Scripts/Pymol-script-repo/blob/master/plugins/autodock\_plugin.py
```

Vá em “*Plugin*” e em seguida em “*Install New Plugin*” (Figure 21.3), “*Choose file*” e insira o arquivo baixado na etapa anterior.

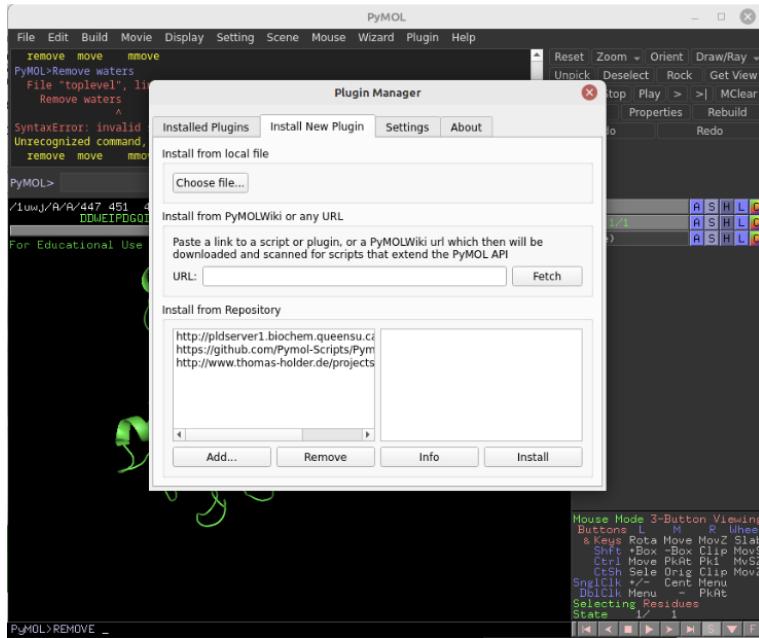


Figura 21.3: Página Install New Plugin no PyMOL

### 21.3 Re-docking com o PyMOL

No experimento de re-docking vamos selecionar uma estrutura complexo proteína-ligante resolvido experimentalmente e depositado no *Protein Data Bank* (PDB) (<https://www.rcsb.org/>). O ligante será removido e tentaremos chegar na mesma posição experimental através do docking. Portanto, obteremos uma validação do protocolo e um teste do algoritmo de atracamento para a proteína estudada. Ao final, podemos comparar as interações e o RMSD (*Root Mean Square Deviation* – Desvio Quadrático Médio da Raiz) entre pose predita e a posição cristalográfica do ligante. Nesse caso, para testar vamos utilizar a estrutura cujo **código de acesso no PDB é 1UWJ**.

O código corresponde a estrutura de uma proteína mutante, o complexo do mutante V599E B-Raf e BAY439006. A B-Raf é uma proteína codificada pelo gene BRAF e está envolvida na via RAS/MAPK, que regula o crescimento e a divisão celular. Mutações como a V599E de BRAF estão presentes em mais de 60% dos melanomas e foram encontradas em taxas mais baixas em carcinomas de pulmão,

côlon e ovário. Nessa estrutura a proteína está complexada com um inibidor BAX [5].

1. Abra o pymol e digite o comando no terminal do PyMOL:

```
fetch 1uwj
```

2. Vamos remover a segunda cadeia apenas para facilitar o nosso experimento, uma vez que ela não participará do redocking. Remova a cadeia B com o comando:

```
remove chain B
```

3. Caso as estruturas tenham moléculas de água, em geral nessa fase elas são removidas. Caso se tenha evidências na literatura que a interação do receptor com o ligante necessita da interação com as moléculas de água, as moléculas que são importantes para interação devem ser mantidas. Clique em “action” em seguida “remove waters” .

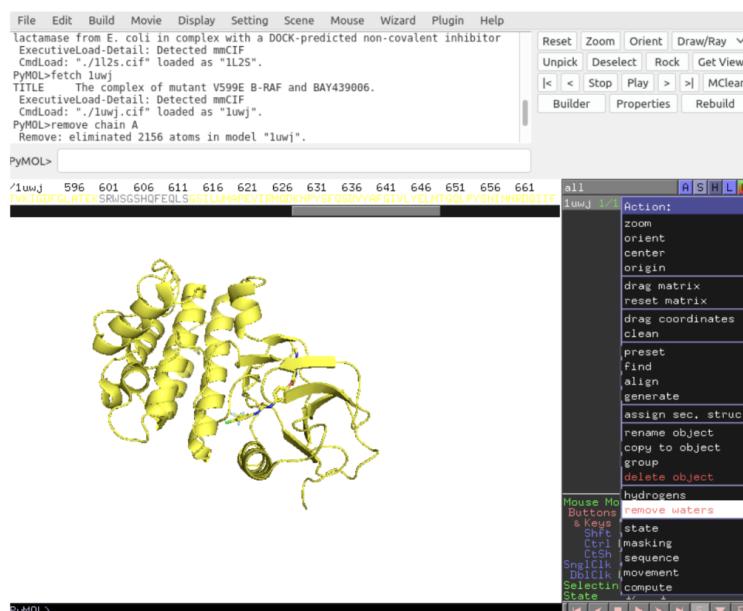


Figura 21.4: Removendo as moléculas de água da estrutura.

Ou digite no terminal de comando:

```
remove resn hoh
```

4. Agora vamos criar um objeto separado do ligante. Para isso vamos utilizar o seguinte comando.

```
extract ligante, resn BAX
```

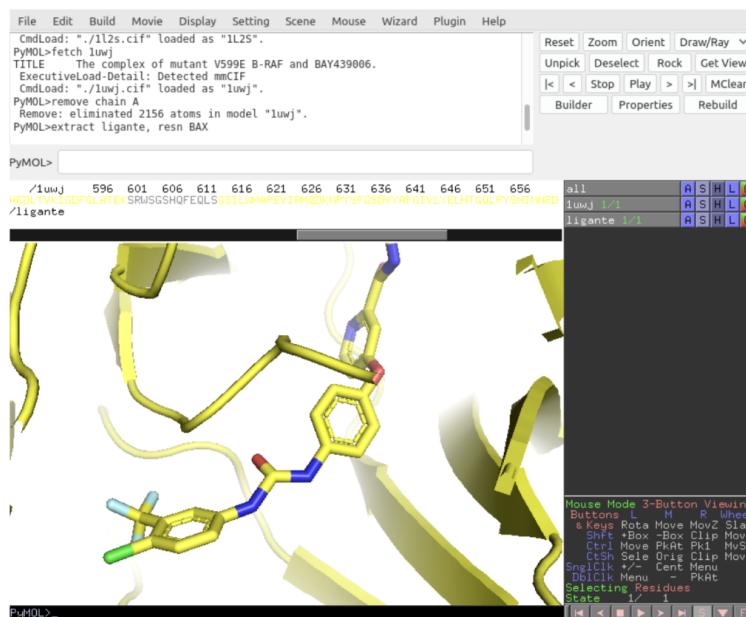


Figura 21.5: Extraiendo o ligante.

Em seguida, vamos criar um objeto contendo o inibidor e os resíduos de proteína a no máximo 5 Å de distância do composto, o objetivo aqui é selecionar os resíduos que fazem as interações moleculares mais próximas. Estamos utilizando 5 Å como exemplo, e esse valor pode ser alterado conforme o experimento realizado. Para isso vamos utilizar o comando abaixo:

```
create sitio_ligante, ligand around 5
```

Agora, para realizar o docking vamos acessar o plugin do autodock clicando em “Plugin” – “Legacy Plugins” – autodock/vina. Atente-se para inserir o local onde estão instalados o AutoDockTools e o executável do vina.

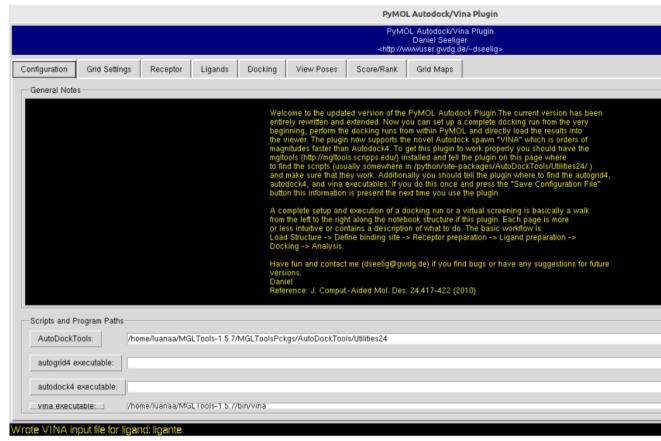


Figura 21.6: Inserindo caminho para os executáveis.

5. Na aba “Grid”, vamos estabelecer o valor da caixa e seu centro no ligante, usando o sítio que criamos anteriormente.



Figura 21.7: Delimitando a região onde será realizado o docking.

Basta marcar a opção “*Calculate grid center by selection*” e em seguida digite na aba de seleção o sítio do ligante que marcamos na etapa anterior. Clique no

botão “Show Box” e selecione a opção “Calculate Grid Center by Selection”. Digite sitio\_ligante na caixa de diálogo e clique no “enter”.

**Observação:** o Vina não faz um pré-cálculo de Grid. Porém, uma demarcação da área de docking (sítio ativo ou até mesmo a proteína toda) precisa ser estabelecida.

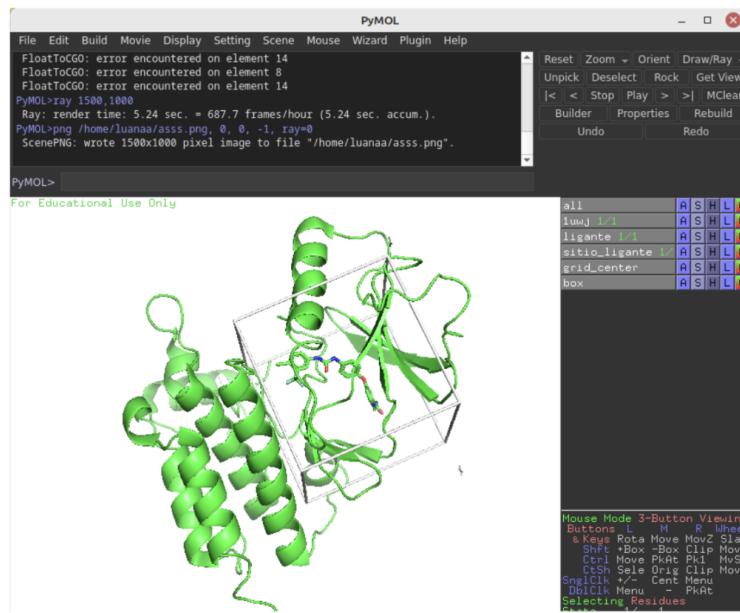


Figura 21.8: Visualizando a Grid Box.

6. A preparação do receptor é feita na aba “Receptor”. Em “PyMOL Selections” escolha 1UWJ. Clique em “Generate Receptor”, nessa etapa o arquivo será convertido no formato pdbqt que é exigido para o docking. Nesta fase é possível flexibilizar os resíduos de interesse do sítio do receptor, mas não utilizamos essa opção usualmente para o redocking.

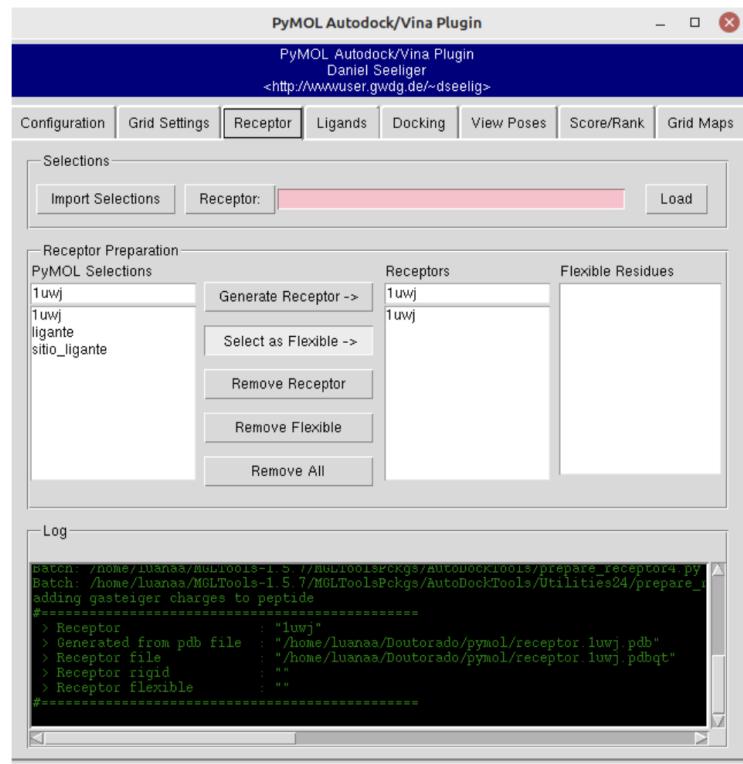


Figura 21.9: Preparando o receptor.

7. A preparação do ligante é feita na aba “Ligands”. Em “PyMOL Selections” escolha “ligante”. Clique em “Generate Ligand”, nessa etapa o arquivo será convertido no formato pdbqt que é exigido para o docking.

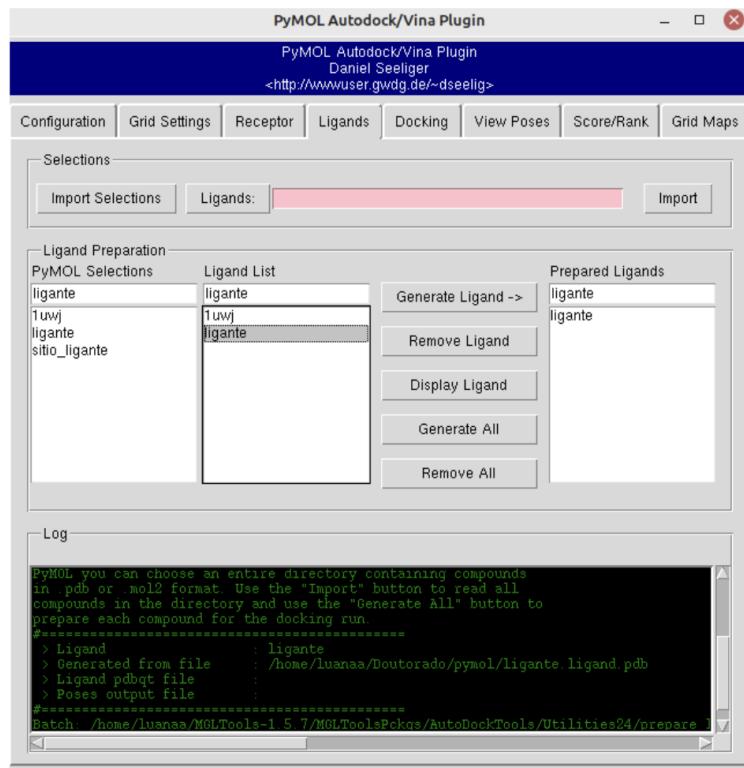


Figura 21.10: Preparando o ligante.

8. Com a área do grid selecionada, receptor e ligante preparados, podemos realizar o docking. Na aba “Docking”, clique no botão “Vina”. Aqui você pode selecionar o número de poses a serem geradas e se utilizaremos as cadeias flexibilizadas.

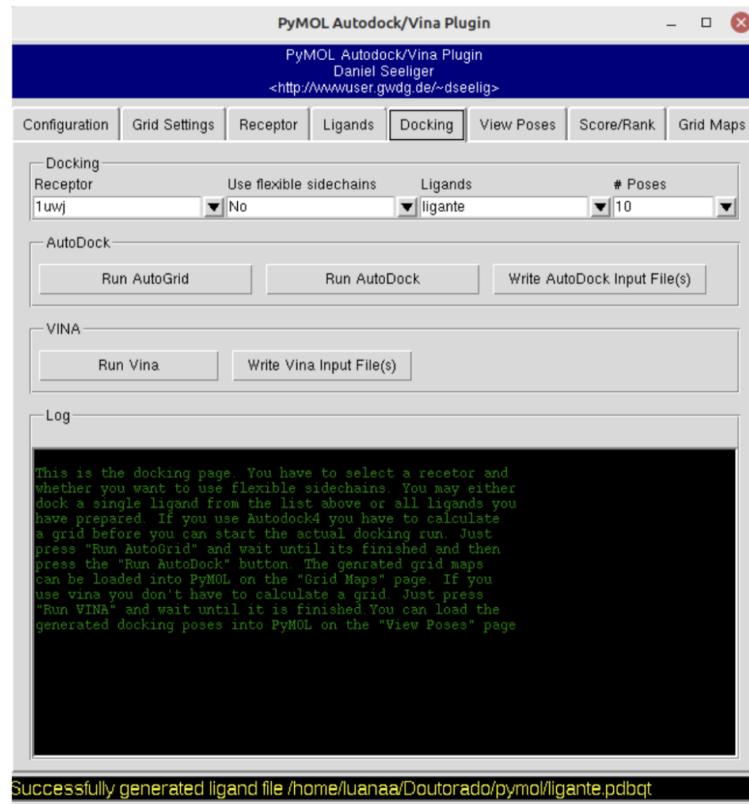


Figura 21.11: Executando o docking.

Depois de gerar as poses, é exibido no terminal de comando do computador a afinidade de interação. Lembre-se de que quanto mais baixa essa medida de energia, maior a afinidade predita entre o ligante e a proteína.

```

#
# O. Trott, A. J. Olson,
# AutoDock Vina: improving the speed and accuracy of docking
# with a new scoring function, efficient optimization and
# multithreading, Journal of Computational Chemistry 31 (2010)
# 455-461
#
# DOI 10.1002/jcc.21334
#
# Please see http://vina.scripps.edu for more information.
#####
Detected 4 CPUs
Reading input ... done.
Setting up the scoring function ... done.
Analyzing the binding site ... done.
Using random seed: -1235395869
Performing search ...
0% 10 20 30 40 50 60 70 80 90 100%
|---|---|---|---|---|---|---|---|---|
*****done.
Refining results ... done.

mode | affinity | dist from best mode
| (kcal/mol) | rmsd l.b.| rmsd u.b.
-----+-----+-----+
1    -12.3      0.000   0.000
2    -11.9      1.767   2.380
3    -11.8      5.956   12.150
4    -11.8      5.585   11.427
5    -11.7      5.381   11.745
6    -11.7      1.284   1.810
7    -11.5      1.964   2.680
8    -11.1      3.378   4.072
9    -10.9      3.133   3.930
10   -10.9      5.989   11.762
Writing output ... done.

```

Figura 21.12: Resultado do docking no terminal de comando.

9. Ao final você pode visualizar as poses geradas na aba “View Poses”. Clique no botão “Browse” e selecione o arquivo ligante.docked.pdbqt.

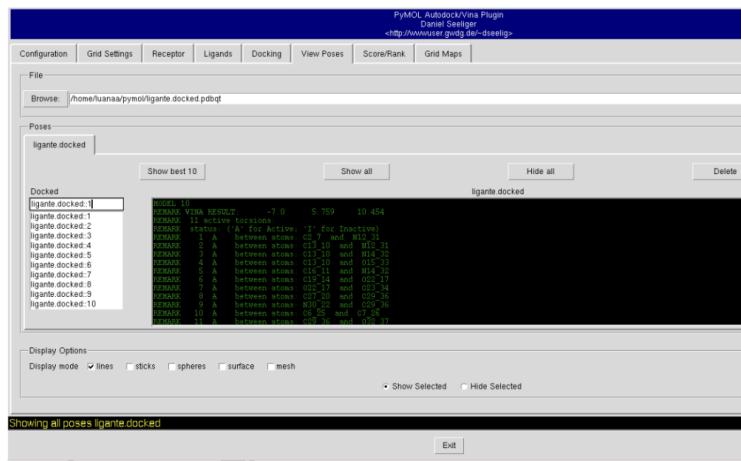


Figura 21.13: Abrindo os resultados do docking no PyMOL.

Para mostrar todas as poses obtidas na janela de visualização do PyMOL, clique no botão “Show all”.

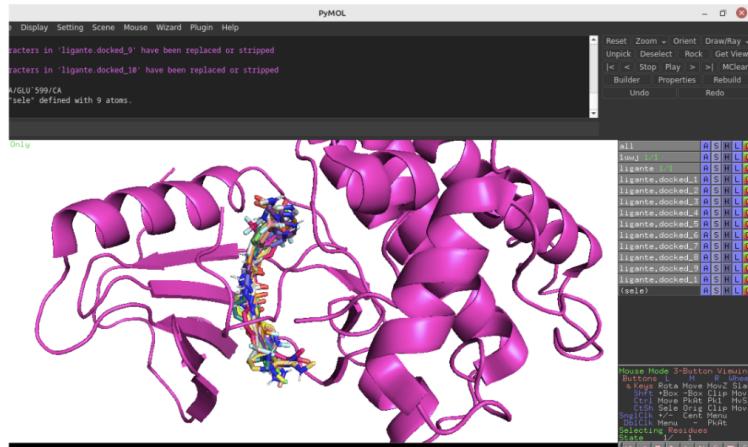


Figura 21.14: Visualizando os resultados do docking no PyMOL.

10. Por fim, vamos calcular o RMSD entre as poses obtidas no docking e estrutura experimental. Para isso podemos utilizar o comando:

```
rms_cur ligante.docked_1, ligante
rms_cur ligante.docked_2, ligante
rms_cur ligante.docked_3, ligante
rms_cur ligante.docked_4, ligante
rms_cur ligante.docked_5, ligante
rms_cur ligante.docked_6, ligante
rms_cur ligante.docked_7, ligante
rms_cur ligante.docked_9, ligante
rms_cur ligante.docked_10, ligante
```

Para ser considerado um bom resultado de re-docking você precisa ter obtido uma pose com RMSD < 2.0 Å. Caso isso não tenha ocorrido devemos ajustar os parâmetros, sendo eles tamanho, localização da caixa e resíduos que foram considerados com sítio.

## 21.4 Conclusões

No experimento acima podemos observar que o re-docking foi bem sucedido, observando os valores de RMSD. O que nos sugere que a ferramenta e os parâmetros utilizados estão dentro do esperado.

A realização do procedimento de re-docking nos ajuda a validar a ferramenta utilizada e sua utilização é altamente recomendada antes da realização do docking com seus ligantes de interesse.

### Nota de transparência 21.1

Este material foi originalmente produzido para um minicurso ministrado durante o Curso de Inverno em Bioinformática da UFMG, realizado em 4 de Julho de 2023, na Universidade Federal de Minas Gerais, Belo Horizonte, Brasil.

### Saiba mais 21.1

Este artigo está disponível em <https://bioinfo.com.br/re-docking-molecular-utilizando-o-pymol-e-autodock-vina/>

## 21.5 Referências

[1] MENG, Xuan-Yu et al. Molecular docking: a powerful approach for structure-based drug discovery. *Current computer-aided drug design*, v. 7, n. 2, p. 146-157, 2011.

[2] Lepšík, M., Řezáč, J., Kolář, M., Pecina, A., Hobza, P., & Fanfrlík, J. (2013). The Semiempirical Quantum Mechanical Scoring Function for In Silico Drug Design. *ChemPlusChem*, 78(9), 921–931. 3.

[3] Forli, S., Huey, R., Pique, M. E., Sanner, M. F., Goodsell, D. S., & Olson, A. J. (2016). Computational protein-ligand docking and virtual drug screening with the AutoDock suite. *Nature Protocols*, 11(5), 905–919. doi:10.1038/nprot.2016.051 4.Trott O, Olson A

[4] AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem*. 2010;31(2):455-461. doi:10.1002/jcc.21334.

[5] Wan PT, Garnett MJ, Roe SM, Lee S, Niculescu-Duvaz D, Good VM, Jones CM, Marshall CJ, Springer CJ, Barford D, Marais R; Cancer Genome Project. Mechanism of activation of the RAF-ERK signaling pathway by oncogenic mutations of B-RAF. *Cell*. 2004 Mar 19;116(6):855-67. doi: 10.1016/s0092-8674(04)00215-6. PMID: 15035987.

[6] Mateev, Emilio, et al. "Validation through re-docking, cross-docking and ligand enrichment in various well-resolved MAO-B receptors." *Int J Pharm Sci Res* 13 (2022): 1000-100

[7] Shamsara J. CrossDocker: a tool for performing cross-docking using Autodock Vina. Springerplus. 2016 Mar 17;5:344. doi: 10.1186/s40064-016-1972-4. PMID: 27652002; PMCID: PMC4797978.

# 22

## COLABFOLD: UMA FERRAMENTA WEB PARA MODELAGEM DE PROTEÍNAS

Autores 22.1

Giovana Fiorini , Luana Luiza Bastos , Rafael Pereira Lemos 

Revisão: Aline de Paula Dias da Silva , Bárbara Rebeca de Macedo Pinheiro 

Cite este artigo 22.1

Fiorini, G; Bastos, LL; Lemos, RP. **ColabFold: uma ferramenta web para modelagem de proteínas.** BIOINFO. ISSN: 2764-8273. Vol. 3. p.22 (2023). doi: 10.51780/bioinfo-03-22

### Resumo 22.1

A modelagem de proteínas é um desafio da biologia molecular que ficou em aberto por mais de 50 anos. Recentemente, estratégias computacionais obtiveram bastante sucesso em modelar tridimensionalmente a estrutura de macromoléculas. O AlphaFold é um software que utiliza aprendizado profundo para predizer estruturas de proteínas. Entretanto, sua instalação e uso ainda pode ser complexo para boa parte dos potenciais usuários. Em 2022, Milot Mirdita e colaboradores propuseram a ferramenta ColabFold: um programa rápido e fácil de usar para a previsão de estruturas de proteínas e complexos, que funciona por meio de um navegador de internet. Neste artigo, você irá conhecer um pouco das funcionalidades do ColabFold. A ferramenta está disponível em: <https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb>.

## 22.1 Modelagem de proteínas e o AlphaFold

As proteínas são biomoléculas essenciais para a vida, uma vez que estão relacionadas a funções estruturais na célula, de transporte, catálise, sinalização, defesa, dentre outras. Para desempenhar suas funções, é necessário que as proteínas apresentem estrutura e conformação adequadas (estrutura terciária ou quaternária). Dessa forma, há forte relação entre a sequência de resíduos de aminoácidos, estrutura e função de uma proteína [1, 4].

Para estudar as estruturas de proteínas podemos elucidá-las experimentalmente por meio de técnicas como cristalografia de raio-x ou ressonância magnética nuclear. Entretanto, essas são técnicas caras e que necessitam de mão de obra extremamente especializada. Assim, torna-se fundamental a criação de programas para auxiliar no estudo de estruturas de proteínas in silico, como a modelagem computacional de proteínas [1].

A modelagem computacional de proteínas pode ser realizada, por exemplo, utilizando uma estrutura resolvida como molde para o modelo. Caso haja estruturas experimentalmente resolvidas que apresentam sequências com alta identidade com a proteína que se deseja modelar ( $> 25\%$  de identidade), podemos utilizar a modelagem comparativa ou threading [8]. Caso não haja, é feita uma modelagem ab initio (também conhecida como modelagem de novo). Existem vários programas que realizam modelagem molecular de proteínas, cada um com suas próprias métricas específicas que ranqueiam modelos e estimam energia de ligação e interação entre os resíduos [1].

Dentre os vários programas disponíveis, destacamos aqui o AlphaFold. O AlphaFold é um software que utiliza aprendizado profundo para predizer estruturas de proteínas. Em 2018 e 2020, tanto sua primeira quanto segunda versão obtiveram resultados excepcionais no CASP (*Critical Assessment of Protein Structure Prediction*), uma competição que avalia métodos de predição de estruturas de proteínas. Com o AlphaFold é possível obter a estrutura terciária de uma proteína, usando como entrada apenas uma sequência de aminoácidos. Assim, é possível obter estruturas com alta acurácia, mesmo quando não existe nenhuma estrutura similar [2].

A segunda versão do AlphaFold funciona basicamente em três passos principais: uma etapa de preparo da sequência recebida, que envolve alinhamentos múltiplos de sequência (MSA); uma etapa de camadas repetidas de rede neural que leva como entrada a matriz de alinhamento de sequências gerada no passo anterior; e por fim, a geração da estrutura tridimensional da proteína a partir de mais blocos de redes neurais, realizando uma reciclagem iterativa e refinando a estrutura gerada [2].

A metodologia MSA expande o alinhamento de pares integrando sequências suplementares para revelar regiões conservadas e conexões evolutivas em uma infinidade de sequências. Os algoritmos utilizados podem ser amplamente classificados em duas categorias: métodos progressivos e métodos iterativos. ClustalW e T-Coffee são exemplos de métodos progressivos utilizados no alinhamento de sequências. Esses métodos constroem o alinhamento

progressivamente, inicialmente alinhando pares de sequências e posteriormente integrando sequências adicionais. Técnicas iterativas, exemplificadas por MUSCLE e MAFFT, aprimoram o alinhamento iterativamente, alinhando subconjuntos de sequências e, por fim, revisando o alinhamento com base nos resultados iniciais [1,2].

Embora o AlphaFold apresente resultados eficientes, a ferramenta apresenta algumas limitações computacionais. Uma delas é que, para realizar uma modelagem de alta qualidade, o AlphaFold precisa construir diversos MSAs, e para isso é necessário ter acesso a uma grande coleção de sequências de proteínas de referência em bancos de dados públicos. Esse acesso é realizado por métodos sensíveis de detecção de homologia, como HMMer e HHblits cujo a busca requer um tempo considerável para ser realizada. Em segundo lugar, para executar as redes neurais profundas, são necessárias unidades de processamento gráfico (GPUs) e uma grande quantidade de RAM (memória de acesso aleatório), mesmo para tamanhos de proteína relativamente comuns de  $\sim 1.000$  resíduos [3].

Para tentar solucionar a exigência de um grande suporte computacional, em 2022, Mirdita e colaboradores [3] propuseram o ColabFold: um programa rápido e fácil de usar para a previsão de estruturas de proteínas e complexos (homo e heteroméricos). O ColabFold pode ser usado por meio de um Jupyter Notebook dentro do Google Colaboratory. Isso torna possível para usuários comuns utilizar o método de modelagem do AlphaFold sem a necessidade de possuir uma grande capacidade computacional. Além disso, o Google Colab é acessível gratuitamente para usuários que possuem conta Google, incluindo acesso a GPUs poderosas, além de ser uma ferramenta em nuvem [3, 5].

## 22.2 ColabFold

O primeiro passo de funcionamento do ColabFold é receber as sequências FASTA de entrada, seja por meio do Google Colab, em servidor web (acessando o site: <https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb>) ou por meio de linha de comando (Figura 22.1a). O próximo passo, assim como o AlphaFold 2, é realizar um MSA (Figura 22.1b).

Entretanto, o método padrão utilizado pelo ColabFold é o MMseqs2, capaz de oferecer maior rapidez, captar bem as diversidades das sequências, além de ser pequeno e suficiente para rodar em máquinas com menos memória RAM. Para realizar o alinhamento, os autores do ColabFold usaram as bases UniRef100 e PDB70, além de criarem sua própria base de dados, a qual chamaram de ColabFoldDB. O ColabFoldDB não possui a redundância das bases de dados já utilizadas pelo AlphaFold2 (*Big Fantastic Database* e *MGNify database*), além de possuir outras contendo dados de eucariotos, por exemplo. No passo de MSA, são feitas três iterações de busca por sequências em cada um dos bancos de dados. Após essa etapa, usa-se uma biblioteca Python para preparar as entradas de acordo com o tipo da predição: cadeia simples (Figura 22.1c.i) ou complexo de proteínas (Figura 22.1c.ii). Então, uma nova matriz MSA é retornada para ser usada como entrada para os modelos do AlphaFold 2. A partir daí, o funcionamento das duas ferramentas é semelhante [3].

Para modelos de complexos de proteínas, é estabelecido o contato entre cadeias por meio de um alinhamento das sequências mais promissoras numa mesma espécie, e então, gera-se um MSA não pareado, para guiar a previsão das cadeias de proteína (Figura 22.1c.ii). Ao final, os cinco melhores modelos são gerados, juntamente aos gráficos para avaliação (Figura 22.1d). Resumidamente, o funcionamento do ColabFold se baseia em alinhar a sequência de input com outras sequências dos bancos de dados para a geração da matriz MSA. Essa matriz será utilizada como entrada para as redes neurais do AlphaFold, que por fim, geram cinco modelos de estruturas.

O ColabFold utiliza as mesmas métricas utilizadas pelo AlphaFold 2: lDDT predito, para cadeias simples, e PAE para complexos [3]. O Local difference distance test (LDDT, em português, Teste de Diferença de Distância Local) é uma métrica para comparar estruturas de proteínas que avalia a diferença entre as distâncias de cada par de átomos e independe da sobreposição estrutural. Já o *Predicted Aligned Error* (PAE, em português, Erro Predito de Alinhamento) avalia a posição relativa entre as cadeias da proteína utilizando a distância entre dois resíduos.

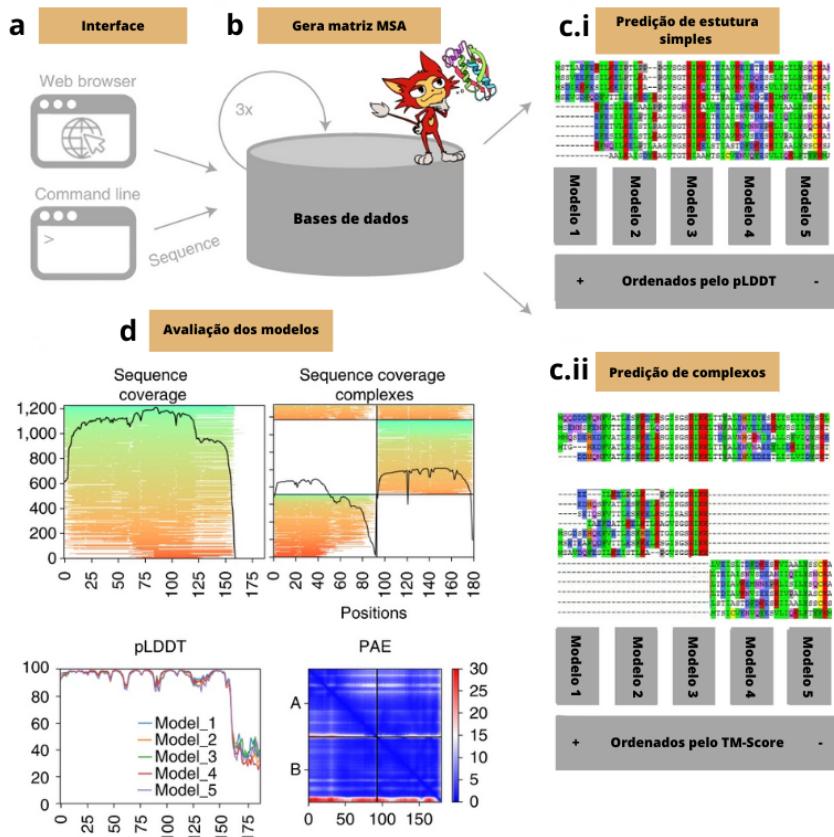


Figura 22.1: Workflow resumido do ColabFold. (a). Input da sequência FASTA de aminoácidos da proteína via servidor web ou linha de comando. (b). Alinhamentos múltiplos de sequências usando as bases de dados UniRef100, PDB70 e base de dados própria, usando três iterações para pesquisa de sequências para cada uma. (c). Predição de estruturas tridimensionais das proteínas a partir do MSA e algoritmos usados pelo AlphaFold 2. (d). Avaliação dos modelos gerados por meio de gráficos de cobertura das sequências, pLDDT e PAE. Fonte: adaptado de Mirdita et al., 2022 [3].

Comparando o ColabFold a outras ferramentas, ele se mostrou cinco vezes mais rápido do que o AlphaFold 2 para predição de estruturas únicas [3]. Além disso, obteve maior TM-score comparado com AlphaFold 2 e RoseTTAFold, quando modelados 65 alvos do CASP14. O TM-score é uma medida de similaridade entre duas estruturas de proteínas, que varia entre 0 e 1, sendo quanto mais próximo de 1, mais semelhantes. Ele é um importante parâmetro de qualidade para avaliar modelos, uma vez que seu cálculo não é dependente do tamanho das sequências das proteínas [7].

Por fim, em relação à modelagem de complexos, o ColabFold obteve maior precisão na previsão em conjuntos de dados do ClusPRO, comparado ao AlphaFold Multimer (versão do AlphaFold 2 criada para predizer complexos) [3].

### 22.2.1 Versões do ColabFold

O ColabFold está disponível em diversas versões, que possuem vantagens e desvantagens umas sobre as outras (Figura 22.2). Todas as versões disponíveis podem ser encontradas no GitHub (disponível em <https://github.com/sokrypton/ColabFold>).

A primeira lista de versões (Figura 2a) se refere às mais bem estabelecidas e funcionais, e são as mais recomendadas para serem usadas. Também é possível encontrar versões em desenvolvimento (“BETA”, Figura 2b), e versões que não são mais utilizadas, mas que podem ser úteis para casos específicos (Figura 2c).

As versões são divididas de acordo com a presença ou ausência de cinco funcionalidades principais identificadas pelas colunas das tabelas, sendo elas (Figura 22.2): (i) capacidade de modelagem de cadeias monoméricas de proteínas; (ii) capacidade de modelagem de cadeias complexas (multiméricas) de proteínas; (iii) utilização do algoritmo MMseqs2 de alinhamento; (iv) utilização do algoritmo jackhmmer de alinhamento (uma variação do algoritmo HMMER capaz de realizar múltiplas etapas de iteração); (v) capacidade de modelagem a partir de moldes de proteínas (templates). Sendo assim, não há uma única versão que possua as cinco funcionalidades, e a escolha deverá ser feita conforme a pergunta a ser resolvida.

**a)**

Versões Funcionais					
Versão	Monômeros	Complexos	MMseqs2	Jackhmmer	Templates
AlphaFold2_mmseqs2	Sim	Sim	Sim	Não	Sim
AlphaFold2_batch	Sim	Sim	Sim	Não	Sim
AlphaFold2 (versão da Deepmind)	Sim	Sim	Não	Sim	Não
Relax_amber (relaxamento da estrutura de entrada)					
ESMFold	Sim	Talvez	Não	Não	Não

**b)**

Versões em Desenvolvimento (Beta)					
Versão	Monômeros	Complexos	MMseqs2	Jackhmmer	Templates
RoseTTAFold2	Sim	Sim	Sim	Não	Em Progresso
OmegaFold	Sim	Talvez	Não	Não	Não

**c)**

Versões Antigas					
Versão	Monômeros	Complexos	MMseqs2	Jackhmmer	Templates
RoseTTAFold	Sim	Não	Sim	Não	Não
AlphaFold2_advanced	Sim	Sim	Sim	Sim	Não
AlphaFold2_complexes	Não	Sim	Não	Não	Não
AlphaFold2_jackhmmer	Sim	Não	Sim	Sim	Não
AlphaFold2_noTemplates_noMD					
AlphaFold2_noTemplates_yesMD					

Figura 22.2: Diferentes versões do ColabFold. São mostradas as versões atuais (a), em desenvolvimento (b), e que não são mais atualizadas (c). As colunas são detalhadas no texto. Versões com colunas em branco desempenham apenas uma função (Relax\_amber) ou são demonstrativas (AlphaFold2\_noTemplates). Dados obtidos em 02 de Agosto de 2023 e disponíveis em <https://github.com/sokrypton/ColabFold>.

Apesar das diferentes versões possuírem diferentes funcionalidades, todas seguem o mesmo padrão de utilização do Google Colab, além de possuírem um pequeno tutorial e seção de instruções em suas próprias páginas.

### 22.2.2 Conhecendo a interface do ColabFold

O objetivo desta seção é apresentar na prática o ColabFold, mostrando as entradas e saídas do programa, além de apresentar os parâmetros que podem ser alterados dependendo do objetivo do usuário.

Para esta seção, utilizaremos a versão do AlphaFold2 que foi implementada com o algoritmo de alinhamento MMseqs2.

Ao abrirmos o ColabFold no Google Colab, nos deparamos inicialmente com informações gerais sobre o programa, bem como a referência do artigo em que foi proposta a ferramenta (Figura 22.3).

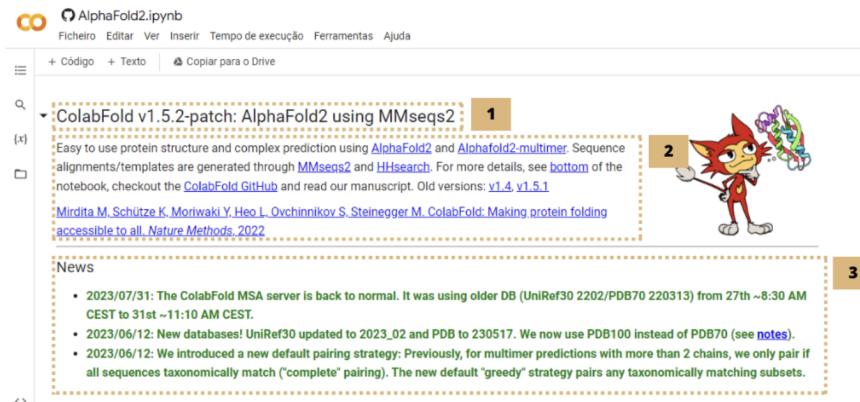


Figura 22.3: Página inicial do ColabFold, via GoogleColab. 1. Título e versão usada do ColabFold. 2. Descrição breve da ferramenta e referência. 3. Atualizações sobre o programa.

O primeiro bloco de código consiste na entrada da sequência de aminoácidos da proteína a ser modelada (“query\_sequence”). (Figura 22.4.1). Para a modelagem de complexos, a entrada deve ser a primeira sequência, seguida de dois pontos (:), e logo em seguida a segunda sequência, sejam as proteínas homo ou heterodímeros (Figura 22.5).



Figura 22.4: Primeiro bloco de código do ColabFold. 1. Informar sequência de aminoácidos de entrada. 2. Nome dado ao trabalho. 3. Número de estruturas a serem relaxadas por meio do campo de força AMBER (opções: 0, 1 ou 5). 4. Tipo de template a ser utilizado (opções: “none”, “pdb100” e “custom”).

```

1
query_sequence: "PIAQIHITEGRSDEQKETLIREVSEAIRSLDAPLTSVRVIITEMAKGHFGIGGELASK"

2
query_sequence: "PIAQIHITEGRSDEQKETLIREVSEAIRSLDAPLTSVRVIITEMAKGHFGIGGELASK:ATDEQKILKR"

```

Figura 22.5: Exemplo de entrada de sequências no ColabFold. 1. Entrada de sequência para modelagem de proteína com somente uma cadeia. 2. Entrada de sequência para modelagem de complexo proteico.

Os próximos passos são: dar um nome ao trabalho (Figura 22.4.2); escolher o número de estruturas para serem relaxadas usando o campo de força AMBER (nenhuma, apenas a melhor, ou todas), até que todas as violações e conflitos estruturais sejam resolvidos (Figura 22.4.3); e escolher se o programa utilizará estruturas com coordenadas atômicas já definidas como templates. É possível selecionar a opção “none”, na qual o programa não utiliza informação de *template*, “pdb100” em que o programa busca templates no banco de dados pdb100, e “custom”, onde o próprio usuário pode inserir o seu *template* (Figura 22.4.4).

O próximo bloco de código é dedicado à instalação das dependências do programa (não se preocupe, nada será instalado em seu computador, apenas na instância da nuvem que você estiver utilizando). Logo em seguida, encontramos parâmetros relacionados ao alinhamento múltiplo de sequências (MSA) (Figura 22.6). O parâmetro “*msa\_mode*” (Figura 22.6.1) nos permite escolher qual base de dados será usada para a criação da matriz de alinhamento múltiplo. As opções dadas são (lembrando que a versão que estamos utilizando possui exclusivamente o algoritmo MMseqs2 de alinhamento):

- “*mmseqs2\_uniref\_env*”, o programa realizará o alinhamento múltiplo utilizando o algoritmo padrão do ColabFold (MMseqs2), e buscará na base de dados UniRef100, além da base de dados criada pelo próprio ColabFold.
- “*mmseqs2\_uniref*”, em que o programa realizará o alinhamento utilizando dados somente da base UniRef100.
- “*single\_sequence*”, onde o programa não realizará alinhamento e partirá para as próximas etapas somente com sua sequência inicial (essa opção é interessante quando sua proteína é desenhada in silico, e você não espera encontrar sequências semelhantes a ela em nenhuma base de dados).
- “*custom*”, em que o usuário submete seu próprio alinhamento múltiplo de sequências.

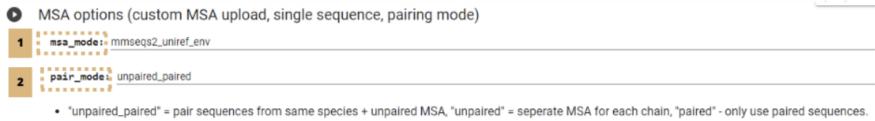


Figura 22.6: Parâmetros relacionados ao MSA. 1. Definir qual base de dados o algoritmo usará para o MSA. 2. Definir o pareamento das sequências do MSA.

O parâmetro “*pair\_mode*” (Figura 22.6.2) nos permite definir como será o pareamento das sequências no MSA. As opções dadas são:

- “*unpaired\_paired*”, em que se pareiam sequências da mesma espécie, e então há a adição de MSAs não pareados (interessante para modelagem de complexos).
- “*unpaired*”, são criados alinhamentos múltiplos separados para cada cadeia (Também interessante para modelagem de complexos).
- “*paired*”, em que são usadas somente sequências pareadas (Interessante para estruturas simples).

Em seguida, temos as configurações avançadas (Figura 22.7). Em “*model\_type*” é possível escolher qual modo do AlphaFold 2 será utilizado, seja ele o AlphaFold

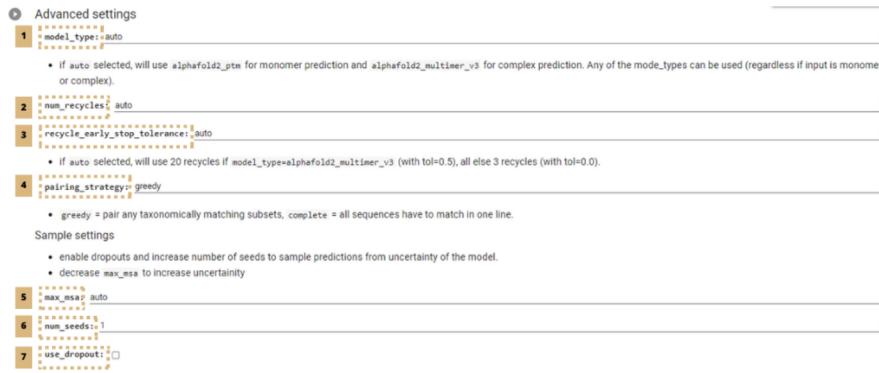


Figura 22.7: Configurações avançadas e configurações relacionadas à amostragem. 1. Definição da versão do AlphaFold a ser utilizada. 2. Número de ciclos de reciclagem da estrutura. 3. Tolerância para parada precoce da reciclagem. 4. Estratégia de pareamento das sequências. 5. Dimensão da matriz de MSA. 6. Número de sementes para a amostragem. 7. Habilitar desistência.

2 simples (“AlphaFold2\_ptm”) ou multimer (alphafold2\_multimer\_v1, v2 ou v3) (Figura 22.7.1). Em “*num\_recycle*” (Figura 22.7.2) é possível escolher quantos ciclos de reciclagem de estrutura o programa realizará (0, 1, 3, 6, 12, 24 ou 48). Essa seleção controla o número de vezes que a previsão é repetidamente alimentada através do modelo. Para alvos difíceis, como proteínas sem homólogos, iterações de reciclagem adicionais podem resultar em uma previsão de mais alta qualidade. Esse parâmetro está relacionado ao “*recycle\_early\_stop\_tolerance*” (Figura 22.7.3), em que o usuário decide quando parar antecipadamente os ciclos de reciclagem, dependendo da confiabilidade da estrutura gerada no final. O parâmetro “*paring\_strategy*” (Figura 22.7.4) está relacionado a estratégia de pareamento que será utilizada no alinhamento de sequências. Ele pode ser “*greedy*” (“gulosa”), em que o programa pareia qualquer conjunto de sequências que tenha correspondência taxonômica, ou “*complete*”, em que todas as sequências devem corresponder entre si, em pelo menos uma linha.

Nas configurações de amostragem temos “*max-msa*” (Figura 22.7.5), que corresponde a dimensão máxima da matriz MSA. Com a redução das dimensões da matriz, temos um aumento da incerteza em relação à previsão da estrutura, uma vez que a quantidade de dados será menor. Em “*num\_seeds*” (Figura 22.7.6), podemos ajustar o número de sementes iniciais. Cada semente representa

um ponto de partida diferente no início da predição, aumentando assim a variabilidade gerada, mas também aumentando consideravelmente o tempo de execução do programa. Por fim, há a opção de habilitar o interrompimento da execução, marcando a caixa de “*use\_dropout*”, a fim de parar o programa quando os modelos começam a convergir demais entre si, evitando *overfitting* na predição (Figura 22.7.7).

Após essas etapas, há opções para salvarmos as configurações usadas nesta análise para que possam ser utilizadas posteriormente: salvar todas as configurações, salvar somente a quantidade de ciclos de reciclagem, salvar as configurações no Google Drive, além de alterar a resolução das imagens que serão geradas como saída do programa (Figura 22.8).

```
Save settings
save_all: 
save_recycles: 
save_to_google_drive: 
  • if the save_to_google_drive option was selected, the result zip will be uploaded to your Google Drive
dpi: 200
  • set dpi for image resolution
```

Figura 22.8: Opções para salvar as configurações e definir resolução das imagens.

Já com todos os parâmetros ajustados, agora podemos rodar a predição clicando em “Run Prediction”. Ao final da predição (que deve demorar alguns minutos por estarmos utilizando o algoritmo MMseqs2, mais rápido que o jackhmmer) podemos visualizar os cinco melhores modelos em “Display 3D structure”. Os gráficos gerados (Figura 22.9) também podem ser visualizados para avaliação da qualidade da estrutura em “Plots”. Por fim, podemos baixar os resultados em forma de um arquivo .zip contendo:

- Arquivos .pdb correspondentes aos modelos gerados, classificados de acordo com o IDDT médio (monômeros) ou TM score (complexos).

- Gráficos que avaliam qualidade dos modelos (plDDT e PAE), bem como a cobertura e identidade do alinhamento de sequências.
- Histórico com os parâmetros definidos para aquela corrida.
- Matriz MSA usada como input para o algoritmo em formato A3M.
- Arquivo .json contendo listas com lDDT e PAE para cada estrutura.
- Arquivo .bibtex contendo referências de cada ferramenta e bases de dados utilizadas.

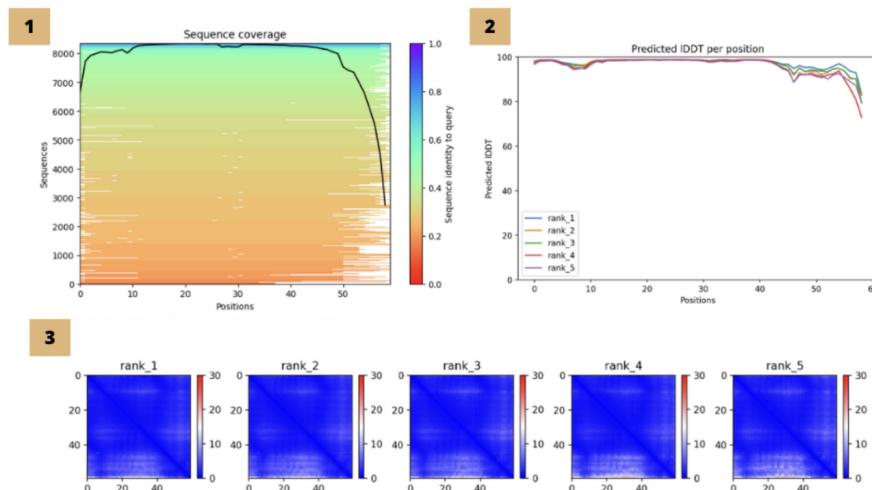


Figura 22.9: Gráficos gerados pelo ColabFold. 1. Cobertura e identidade de sequências encontradas no MSA. 2. Gráfico representando o lDDT predito por posição de cada resíduo em cada um dos cinco modelos gerados. 3. Gráficos representando o erro de alinhamento predito (PAE) em cada um dos cinco modelos gerados.

Na Figura 22.9.1, observamos a matriz de MSA. No eixo x, encontramos as posições dos resíduos de aminoácidos na sequência de entrada, e no eixo y o número de sequências que foram alinhadas à sequência de referência. Na sua escala de cores que vai de vermelho (0.0) a azul (1.0), o gráfico representa a identidade da sequência alinhada com a sequência de entrada. Quanto maior a identidade, mais próximo de azul, e quanto menor identidade, mais próximo de vermelho.

Na Figura 22.9.2, podemos observar o gráfico de IDDT predito por posição. No eixo x, temos a posição dos resíduos na sequência de referência, e no eixo y a escala de IDDT que vai de 0 a 100. No gráfico podemos visualizar 5 linhas, uma de cada cor, cada uma representando um dos 5 modelos gerados pelo ColabFold. Quanto mais próximo de 100 melhor a qualidade da modelagem do resíduo.

Por fim, na Figura 22.9.3 visualizamos o erro de alinhamento predito (PAE) em cada um dos cinco modelos gerados. Nos eixos x e y encontramos as posições dos resíduos na sequência de aminoácidos. Sua escala de cores vai de azul (0) até vermelho (30), onde quanto menor o PAE (mais próximo de 0), maior a qualidade da estrutura.

No fim da página do ColabFold, você pode encontrar instruções de uso disponibilizadas pelos próprios autores do programa, além de informações adicionais como limitações e agradecimentos.

## 22.3 Conclusão

Neste artigo, apresentamos uma breve introdução ao ColabFold, suas vantagens em relação a outros programas de predição de estruturas de proteínas, e suas principais diferenças em relação ao AlphaFold 2. Também evidenciamos as diferentes versões disponíveis do programa, e como cada uma possui pontos vantajosos e desvantajosos em relação às outras. Explicamos também como funciona o ColabFold no Google Colab e cada um dos parâmetros que podem ser ajustados. Com isso, é interessante que o leitor use este artigo como ponto de partida para buscar mais informações sobre cada ferramenta aqui apresentada, além de avaliar as possíveis utilizações dependendo do seu objetivo.

### Nota de transparência 22.1

Este material foi originalmente produzido para um minicurso ministrado durante o Curso de Inverno em Bioinformática da UFMG, realizado em 4 de Julho de 2023, na Universidade Federal de Minas Gerais, Belo Horizonte, Brasil.

Saiba mais 22.1

Este artigo está disponível em <https://bioinfo.com.br/colabfold-uma-ferramenta-web-para-modelagem-de-proteinas/>

## 22.4 Referências

- [1] SILVA, L., BASTOS, L., SANTOS, L. Modelagem computacional de proteínas. In: BIOINFO - Revista Brasileira de Bioinformática e Biologia Computacional (2021): 1-38. Vol. 1. Ed. 1. doi: 10.51780/978-6-599-275326-08.
- [2] JUMPER, J., Evans, R., Pritzel, A. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589 (2021).
- [3] MIRDITA, M. et al. ColabFold: making protein folding accessible to all. *Nature Methods*, v. 19, n. 6, p. 679–682 (2022).
- [4] NELSON, D.L.; COX, M.M. Princípios de Bioquímica de Lehninger. 6. ed. São Paulo: Artmed, 2014.
- [5] TUNYASUVUNAKOOL, K, Adler J, Wu Z. et al. Highly accurate protein structure prediction for the human proteome. *Nature*.(2021).
- [6] LEMOS, R; DOS SANTOS, P. H.; ROCHA, A. “Extração de Informações de Sequências e Estruturas de Proteínas” *Revista Brasileira de Bioinformática* (2023). Disponível em <https://bioinfo.com.br/extracao-de-informacoes-de-sequencias-e-estruturas-de-proteinas/>.
- [7] ZHANG, Y et al. Scoring function for automated assessment of protein structure template quality. *Proteins* 57, 702–710 (2004).
- [8] VERLI, H et al. Bioinformática: da Biologia à Flexibilidade Molecular. 1. edição. – São Paulo : SBBq, 2014. 282.

# 23

## ALPHAFOLD 2: REVOLUCIONANDO

### A MODELAGEM DE ESTRUTURAS

### 3D DE MACROMOLÉCULAS

#### Autores 23.1

Vivian Morais Paixão  , Angie Atoche Puelles  , Eduardo Utsch Madureira Moreira  , Luana Luiza Bastos  , Raquel Cardoso de Melo-Minardi 

Revisão: Ana Carolina Silva Bulla  , Ariany Rosa Gonçalves  , Filipe Augusto Teixeira 

#### Cite este artigo 23.1

Paixão, VM et al. **AlphaFold 2: revolucionando a modelagem de estruturas 3D de macromoléculas.** BIOINFO. ISSN: 2764-8273. Vol. 3. p.23 (2023). doi: 10.51780/bioinfo-03-23

### Resumo 23.1

Desenvolvido pela DeepMind em 2020, o **AlphaFold** é uma inovadora ferramenta de inteligência artificial que surgiu na CASP14, uma competição de predição de estruturas proteicas. À época, ele foi apresentado como uma solução para o desafiante problema do enovelamento de proteínas. Esse problema envolve a compreensão de como uma sequência de aminoácidos se converte em uma estrutura tridimensional. Uma proteína não enovelada vai mudando de conformação, diminuindo a entropia até chegar no estado de menor energia, em que ela estará em seu estado nativo. O “paradoxo de Levinthal” destaca a complexidade desse processo, sugerindo que, embora uma proteína possa se dobrar em milissegundos, o tempo necessário para calcular todas as estruturas possíveis é maior do que a idade do universo conhecido. Embora o AlphaFold não tenha resolvido completamente esse desafio, ele marcou um avanço significativo ao prever com precisão as estruturas proteicas a partir de sequências primárias, revolucionando a pesquisa em biologia.

Neste artigo, nossa intenção será elucidar o processo pelo qual esse sistema constrói as estruturas tridimensionais, abordando também aulas práticas sobre modelagem molecular, utilizando o AlphaFold como ferramenta central em nossos experimentos. Com isso, esperamos proporcionar uma compreensão mais aprofundada das capacidades dessa tecnologia e seu potencial impacto no avanço da pesquisa em bioinformática e biologia molecular.

## 23.1 Introdução

**A**LPHAFOld é uma ferramenta altamente sofisticada que prevê estruturas de proteínas através de outras já conhecidas, baseada em redes neurais profundas. Até então, já realizou mais de 200 milhões de predições [1], tendo sido treinada com estruturas experimentais das proteínas disponíveis no *Protein Data Bank* (PDB) [2]. Sua primeira versão, AlphaFold 1, foi construída em 2018,

com base no trabalho desenvolvido por várias equipes anteriores. Elas tentavam encontrar mudanças em diferentes resíduos que pareciam estar correlacionados, embora não fossem consecutivos na cadeia principal. Tais correlações sugeriam que os resíduos poderiam estar próximos fisicamente, embora não próximos na sequência, e isso permitiu que os cientistas estimassem um mapa de contatos baseado nessas informações [3, 4]. O AlphaFold 1 estendeu isso para estimar uma distribuição de probabilidade de quão próximos os resíduos poderiam estar, construindo um mapa de distâncias prováveis. Assim, o AlphaFold 1 é um preditor de mapas de distância implementado como redes neurais profundas. Juntamente com um mapa de distância na forma de um histograma, o AlphaFold prevê ângulos  $\phi$  e  $\psi$  (Figura 23.1) para cada resíduo, que são usados para criar a estrutura 3D inicial prevista.

Os ângulos descritos acima são de torção em torno das ligações peptídicas, de forma que o ângulo Phi ( $\phi$ ) é medido entre o átomo de nitrogênio (N) e o átomo de carbono-alfa ( $C\alpha$  ou CA), enquanto o ângulo Psi ( $\psi$ ) é medido entre o átomo de carbono- $\alpha$  ( $C\alpha$ ) e o átomo de carbono do grupo carbonila (C=O). Ambos possuem um papel importante na conformação proteica, uma vez que esses ângulos são restritos devido a limitações estéricas e interações eletrônicas. Essas variações nos ângulos contribuem para a diversidade estrutural de proteínas, levando a diferentes configurações tridimensionais e, consequentemente, influenciando suas propriedades funcionais.

O gráfico de Ramachandran, gráfico comumente utilizado para verificar a qualidade das estruturas, representa o espaço conformacional das proteínas, definindo as regiões permitidas e proibidas com base nas conformações estericamente aceitáveis das ligações peptídicas [5]. Dessa forma, a compreensão dos ângulos Phi e Psi é fundamental na predição da estrutura proteica e alterações conformacionais, permitindo, inclusive, a modificação racional de proteínas a fim de melhorar sua estabilidade, atividade catalítica e afinidade por ligantes [6].

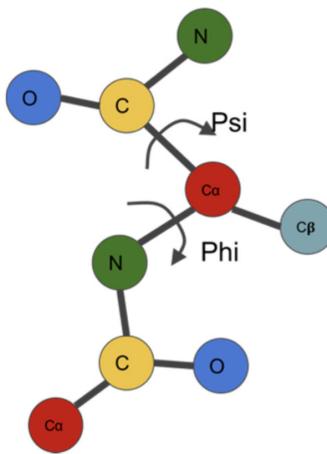


Figura 23.1: Ilustração dos ângulos  $\Phi$  ( $\phi$ ) e  $\Psi$  ( $\psi$ ). Fonte: Fang, C. et al. (2018) [7], disponível em: [10.1109/TCBB.2018.2814586](https://doi.org/10.1109/TCBB.2018.2814586). Acesso em 31/07/2023.

## 23.2 AlphaFold 2

Em 2021, foi criada a segunda versão do AlphaFold, uma vez que a equipe do **DeepMind** havia identificado que sua abordagem anterior tendia a superestimar as interações entre os resíduos que estavam próximos na sequência em comparação com as interações entre os resíduos mais distantes ao longo da cadeia. Como resultado, o AlphaFold 1 poderia preferir modelos com uma estrutura um pouco mais secundária (alfa-hélices e folhas-beta) do que na realidade [8, 9, 10]. Assim, o **AlphaFold 2** surgiu como uma inovação do primeiro, baseado no reconhecimento de padrões de estruturas e sequências. Vale ressaltar que toda a informação sobre o funcionamento da ferramenta pode ser acessada no artigo “Highly accurate protein structure prediction with AlphaFold” [11], referente à sua segunda versão

Para utilizar a ferramenta, deve-se realizar a instalação local usando o código open-source disponível no site da DeepMind através do link <https://www.deepmind.com/open-source/alphafold>, sendo necessário verificar o tipo de sistema e quantidade de memória necessários para uso (requer placa de vídeo).

Entretanto, pode-se ainda utilizar uma versão online disponível através do Google Colab, denominada ColabFold [12], através do seguinte link: <https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb>. Nela, utiliza-se apenas a sequência como input e pode-se alterar alguns parâmetros que julgar necessários.

### 23.2.1 Como funciona essa ferramenta?

A metodologia da ferramenta pode ser vista na Figura 23.2, e consiste de três partes:

1. Pré-processamento
  2. Evoformer
  3. Construção da estrutura
- A figura abaixo resume todo o processo:

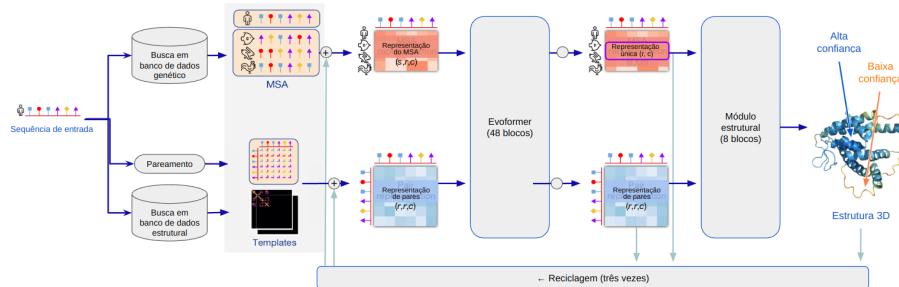
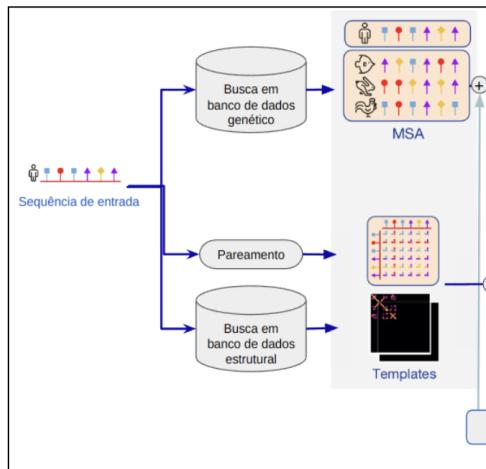


Figura 23.2: Esquema representando a metodologia utilizada pelo AlphaFold 2. Fonte: adaptado de Jumper, J. et al. (2021) [12], disponível em: <https://doi.org/10.1038/s41586-021-03819-2>. Acesso em 31/07/2023.

### 23.2.2 Pré-processamento

É a primeira etapa do AlphaFold e tem como objetivo preparar a sequência de aminoácidos da proteína para a predição da sua estrutura tridimensional. Seu esquema de funcionamento pode ser visto com maior clareza na Figura 23.3, abaixo.



*Figura 23.3: Esquema representando a fase de pré-processamento da metodologia utilizada pelo AlphaFold 2. Adaptado de: Jumper, J. et al. (2021) [12], disponível em: <https://doi.org/10.1038/s41586-021-03819-2>. Acesso em 31/07/2023.*

Essa etapa consiste em:

A ferramenta recebe como entrada a sequência da proteína, que é então dividida em segmentos estruturados e não estruturados. Os segmentos estruturados são aqueles que possuem uma estrutura 3D bem definida e estável, como alfa-hélices, folhas-beta e regiões de loop, ao passo que os não estruturados são mais flexíveis. Dessa forma, os segmentos não estruturados são removidos da sequência.

A sequência é utilizada para realizar uma busca em bancos de dados genéticos e de estruturas, a fim de encontrar sequências semelhantes para que possa ter uma base para a criação da estrutura. Em seguida, as sequências semelhantes são alinhadas com a sequência de interesse para gerar um alinhamento múltiplo de sequências (MSA), que permite ver a similaridade entre elas e determinar quais são mais semelhantes. O MSA é filtrado para remover sequências redundantes e de baixa qualidade, garantindo apenas sequências mais confiáveis na predição da estrutura.

O resultado final da primeira etapa é uma representação do MSA, uma matriz  $N_{seq} \times N_{res}$ , onde  $N_{seq}$  é o número de sequências na MSA e  $N_{res}$  é o número de resíduos na sequência de aminoácidos; além de uma representação de pares: uma matriz  $N_{res} \times N_{res}$ , onde cada elemento representa a relação entre dois resíduos, mostrando os prováveis aminoácidos que estarão em contato com os outros.

Uma observação importante é que muitas proteínas desempenham funções similares em diversas espécies por compartilharem um ancestral comum. Apesar das sequências sofrerem mutações ao longo do tempo, a sua estrutura tende a permanecer semelhante, uma vez que mudanças bruscas podem desestabilizar uma estrutura e possivelmente inviabilizar a função desempenhada. Essa informação ganha relevância no contexto do AlphaFold, já que a ferramenta utiliza sequências e estruturas de proteínas de espécies semelhantes. Essa escolha é estratégica, pois o AlphaFold busca por alinhamentos similares em bancos de dados genéticos para criar sua representação do MSA, de forma a projetar uma estrutura baseada nas sequências similares à sequência de entrada. Trazendo um exemplo atual, a proteína Spike do SARS-CoV-2, tão falada nos últimos anos, é muito semelhante à proteína Spike do vírus SARS-CoV, responsável por uma epidemia em 2002. Há pouca diferença entre os resíduos de aminoácidos entre as duas proteínas, e suas estruturas são muito semelhantes. Provavelmente antes de termos a estrutura experimental da proteína do SARS-CoV-2, o AlphaFold utilizaria a proteína do SARS-CoV como base para gerar sua estrutura teórica.

### 23.2.3 Evoformer

A segunda etapa é composta pelo processamento das entradas através de camadas repetidas de um bloco de redes neurais chamado Evoformer, um transformador, e pode ser vista com maior clareza na Figura 23.4, abaixo.

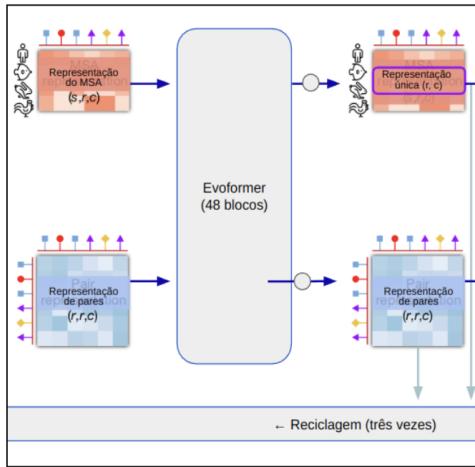


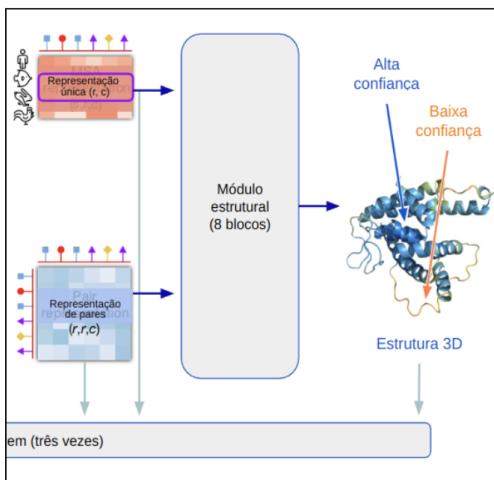
Figura 23.4: Esquema representando a fase do Evoformer da metodologia utilizada pelo AlphaFold 2.

Fonte: adaptado de: Jumper, J. et al. (2021) [12], disponível em: <https://doi.org/10.1038/s41586-021-03819-2>. Acesso em 31/07/2023.

As representações do MSA e de pares de resíduos, geradas na primeira etapa, são utilizadas como entrada no Evoformer. A representação do MSA gerada pelo Evoformer é uma versão refinada do MSA original, que leva em consideração as informações evolutivas e a relação entre as sequências no alinhamento, ou seja, ele filtra as informações mais relevantes das representações geradas na etapa anterior. A matriz gerada nesta etapa é utilizada para gerar uma representação tridimensional da proteína, levando em consideração a relação espacial entre os resíduos. Resumindo, como saída, ele tem representações melhoradas daquelas que foram utilizadas como entrada, que serão usadas na próxima etapa. Cada rede do AlphaFold possui 48 blocos do Evoformer e, dependendo da estrutura, pode passar diversas vezes até chegar em um resultado satisfatório, processo denominado reciclagem.

#### 23.2.4 Módulo de estrutura

É a terceira etapa e tem como objetivo gerar a representação tridimensional da proteína a partir das informações processadas pelo Evoformer. É possível ver seu funcionamento com clareza na Figura 23.5.



*Figura 23.5: Esquema representando a fase do módulo de estrutura da metodologia utilizada pelo AlphaFold 2. Fonte: adaptado de: Jumper, J. et al. (2021) [2], disponível em: <https://doi.org/10.1038/s41586-021-03819-2>. Acesso em 31/07/2023.*

O mapa de representações gerado na etapa anterior descreve as probabilidades de distância entre os pares de átomos presentes na proteína, ou seja, um mapa de distâncias. O módulo estrutural do AlphaFold utiliza um sistema de redes neurais, mais precisamente, oito blocos neurais, que são aplicados em série para traduzir o perfil de probabilidade de distância em uma estrutura tridimensional. Esta rede neural é treinada com um grande conjunto de dados de proteínas com estruturas tridimensionais conhecidas. Durante o treinamento, a rede aprende a mapear os perfis de probabilidade de distância para estruturas tridimensionais que são consistentes com esses perfis. Cada bloco neural recebe como entrada a representação 3D gerada pelo bloco anterior e gera uma nova, dessa vez refinada. O resultado final é uma estrutura tridimensional prevista para a proteína, com cores que representam a confiabilidade de cada região da estrutura. Além disso, essa etapa também pode passar pelo processo de reciclagem, melhorando cada vez mais a estrutura.

### 23.3 AlphaFold 2 x ColabFold

Até o momento, entendemos que o AlphaFold 2 trata-se de um *software* baseado em IA para realizar a predição da estrutura 3D de uma proteína. Existem três práticas experimentais confiáveis para saber como é a conformação da proteína: cristalografia e difração de raios X, ressonância magnética nuclear (NMR) e microscopia eletrônica criogênica. No entanto, essas técnicas são trabalhosas, custosas e demoradas, sem contar a necessidade de mão de obra especializada. Partindo deste princípio, os cientistas vêm trabalhando há tempos para realizar previsões sobre a estrutura 3D a partir de diversos métodos computacionais. Eles obtiveram sucesso apenas quando a DeepMind lançou o AlphaFold, alcançando até 90

Como disponibilizado anteriormente, o AlphaFold 2 possui seu código disponível no Github e é possível compilar no próprio computador, porém, é necessário utilizar equipamentos robustos devido ao custo computacional e operacional da IA. A título de conhecimento, é necessário que o computador tenha pelo menos 3 TB de armazenamento para baixar o banco de dados do AlphaFold, além de placas de vídeo da NVIDIA. Com isso, a DeepMind e o Instituto Europeu de Bioinformática (EMBL-EBI) se uniram para resolver este problema! Essa parceria resultou em um banco de dados denominado AlphaFold DB (<https://alphafold.ebi.ac.uk/>), que disponibiliza gratuitamente 200 milhões de previsões de estruturas do proteoma humano e de outros 47 organismos importantes na pesquisa da saúde global. É possível baixar este repositório acessando o UNIPROT (repositório padrão de sequências e anotações de proteínas). Assim, basta pesquisar pelo código UNIPROT da proteína de interesse e baixar diretamente pela plataforma. Entretanto, uma dificuldade comum na rotina de bioinformaticas é de, muitas vezes, não ter disponíveis informações sobre o nome da proteína/gene, ou até mesmo o código UNIPROT, mas apenas um trecho de uma sequência. Neste contexto, é possível procurar esta sequência em um banco de dados, em busca de sequências semelhantes.

Mas, se quisermos predizer e visualizar a estrutura e não tivermos como rodar o Alphafold em um sistema computacional comum, como proceder? Se você pensou

em ColabFold, acertou! Com ColabFold é possível predizer estruturas de maneira simplificada utilizando o notebook Colab, com pouca diferença da precisão do AlphaFold 2. A depender do tamanho da proteína, em poucos instantes a estrutura será prevista pelo próprio ColabFold e estará pronta para ser baixada em seu sistema.

Agora que você já tem uma noção das diferenças, pode aprender com a gente com duas práticas de modelagem, a seguir.

## 23.4 Prática de modelagem

Neste primeiro tópico prático, começaremos a utilizar esta ferramenta poderosíssima para aprimorar nossa compreensão sobre a modelagem de proteínas. Porém, antes de começarmos nossa primeira prática, é necessário possuir conceitos biológicos básicos sobre o processo de enovelamento de proteínas e as condições físico-químicas para a sua configuração final.

### 23.4.1 Etapa 1: Prepare seus dados

Para obter a sequência primária, acessaremos o NCBI (*National Center for Biotechnology Information*) através do link <https://www.ncbi.nlm.nih.gov/>. Selecionamos o campo “*Protein*” (passo 1) e iremos escrever no campo de busca o nome da proteína de interesse, neste caso, vamos trabalhar com a “*superoxide dismutase*” (passo 2). Em seguida, clicamos no botão “*Search*” (passo 3). Já na página do resultado da nossa busca, selecionamos o organismo de interesse (“*Plants*” – passo 4). Neste caso, vamos trabalhar com a proteína superóxido dismutase na soja (*Glycine max* – passo 5). Esta enzima tem o papel importante na resposta ao estresse oxidativo nas plantas diante da ação de um herbicida, por exemplo.

The screenshot shows the NCBI Protein search results for 'superoxide dismutase'. The search bar at the top has 'superoxide dismutase' entered. A yellow arrow labeled '1' points to the search term. To the right of the search bar is a 'Search' button. Below the search bar, there are filters and a summary section. A yellow arrow labeled '2' points to the 'Summary' link. Another yellow arrow labeled '3' points to the 'Filters: Manage Filters' link. On the left, there is a sidebar with various databases and filters. A yellow arrow labeled '4' points to the 'Plants (9,001)' filter. In the main results area, a yellow arrow labeled '5' points to the protein entry for 'superoxide dismutase [Glycine max]'. The results page shows items 1 to 20 of 9001.

Figura 23.6: Página do NCBI mostrando o resultado na nossa primeira busca para a proteína de interesse.

Utilizaremos o primeiro resultado da nossa busca e, ao clicar no nome da proteína (passo 5), abrirá uma segunda página. Nessa página, você pode encontrar informações importantes para esta proteína. Como nosso objetivo é obter a sequência primária, vamos clicar no botão “FASTA” (passo 6).

Pronto! Agora, é só copiar e colar a sequência em um bloco de notas.

The screenshot shows the detailed view of the protein 'superoxide dismutase [Glycine max]'. At the top, the protein name is displayed. Below it, there are tabs for 'Identical Proteins', 'FASTA', and 'Graphics'. A yellow arrow labeled '6' points to the 'FASTA' tab. The FASTA sequence is shown below, starting with the LOCUS information: NP\_001235936. The sequence is 152 amino acids long and is linear. The page also includes details like DEFINITION, ACCESSION, VERSION, DBSOURCE, KEYWORDS, SOURCE, and ORGANISM.

Figura 23.7: Nesta seção, você consegue encontrar informações gerais da proteína de interesse, por exemplo: seu locus gênico, tamanho, autores da descoberta, comentários gerais, etc.

### 23.4.2 Etapa 2: Acesse e execute o AlphaFold 2

Para esta etapa, é necessário que você possua uma conta registrada no Google para o uso do AlphaFold 2 no Google Colab.

Para começar, no campo de busca do navegador, acesse o link: <https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb> e clique em “Conectar” e “ok” para gerar o aviso do uso (passo 7).

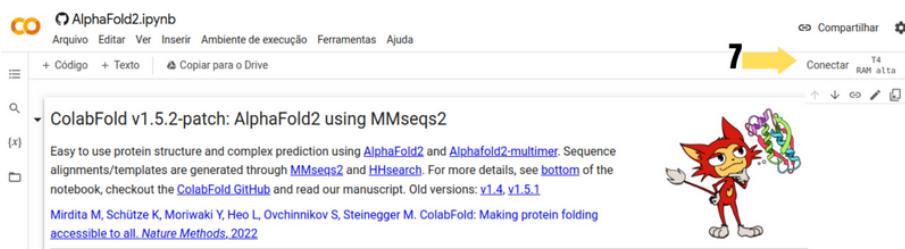


Figura 23.8: O Google Colaboratory, ou Colab, é um serviço disponibilizado pela própria Google. Esta ferramenta permite rodar códigos em Python em uma máquina google através da tecnologia Cloud Computing.

Selecionamos a sequência primária obtida na Etapa 1 e colamos no campo de query\_sequence. Por fim, selecionamos o botão Control + F9 para executar o código inteiro. Outra opção é executar cada célula de código individualmente, permitindo a visualização de cada etapa separadamente. É importante esperar que cada célula acabe de ser executada antes de iniciar a próxima.

Fique atento pois, para este tutorial, usaremos apenas a sessão de “query\_sequence” e “jobname”; os outros parâmetros da plataforma não serão utilizados e, portanto, não precisam ser modificados. No caso, “MSA options” diz respeito a parâmetros do alinhamento múltiplo de sequências e em “Advanced settings” o usuário pode modificar algumas configurações avançadas, como, por exemplo, o número de reciclagens a serem realizadas (explicado anteriormente no funcionamento do AlphaFold) ou o “dpi”, que é basicamente a qualidade da imagem a ser gerada.



Figura 23.9: Neste campo, há vários parâmetros de seleção. Para esta prática, utilizaremos apenas dois: `query_sequence` e `jobname`.

### 23.4.3 Etapa 3: Analisar e interpretar os resultados

Depois de pedir para executar, a etapa da modelagem pode levar algum tempo, dependendo do tamanho da sua sequência. No nosso caso, pode levar até 5 minutos, uma vez que a proteína possui 152 aminoácidos. Ao finalizar, o próprio sistema pedirá que salve o resultado no computador. No entanto, para nossa breve análise, não precisaremos baixar o resultado, já que ele pode ser facilmente visualizado no próprio Colab no final da página. Não se preocupe com os seguintes tópicos do Colab, como *Install dependencies*, *Run Prediction*, *Display 3D structure* e *Plots*. Eles se referem aos campos que o próprio Colab utilizará, ou seja, executarão de modo automático.

Pronto! No final da execução teremos nossa proteína de interesse com a sua modelagem predita. Para gerar o modelo tridimensional predito (Figura 23.11), basta visualizar a próxima etapa “*Display 3D structure*”, juntamente com o nível de confiança de cada região modelada.

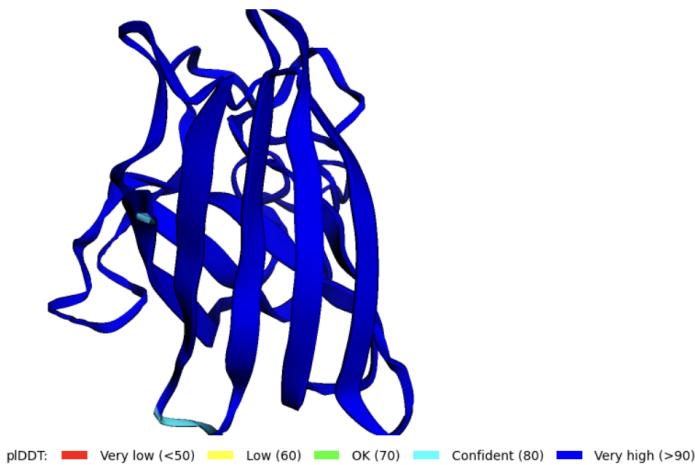


Figura 23.10: Resultado final da predição da proteína superóxido dismutase da soja *Glycine max L.*

No campo de “*Plots*”, podemos visualizar alguns gráficos que indicam a qualidade da estrutura. O primeiro parâmetro é constituído por quatro gráficos e é denominado erro de alinhamento previsto (PAE), que consiste na avaliação da confiança em relação à posição no enovelamento dos domínios proteicos. Esse dado será abordado em nossa segunda prática, a modelagem de complexos, pois não serve para proteínas com apenas um domínio. O gráfico abaixo, na esquerda, mostra o número de sequências por posição, como mostrado na Figura 23.10 (A). Esse gráfico mostra a cobertura da sequência tendo como base todas as sequências semelhantes encontradas, que são representadas no eixo Y, enquanto o eixo X mostra as posições dos resíduos nas sequências. Se houver um gap entre a sequência de entrada e as sequências encontradas, ou seja, regiões faltantes, o trecho faltante estará representado como uma região em branco, indicando uma baixa cobertura. Ao lado, há uma faixa vertical colorida indicando a identidade da sequência com a de entrada, ou seja, uma indicação de similaridade. Além disso, o AlphaFold produz uma estimativa de confiança por resíduo em uma escala de 0 a 100, que estará presente nos arquivos quando você baixar os resultados. Essa medida de confiança é chamada de *Local Distance Difference Test* (plDDT), (Figura 23.10 B), onde a confiabilidade é estimada por resíduo (azul para alta confiança, vermelho para baixa confiança). Espera-se que regiões com plDDT maior que 90 sejam modeladas com alta precisão. Por outro lado, regiões com plDDT entre 50 a

70 indicam baixa confiança e devem ser tratadas com cautela. Abaixo de 50, pode indicar regiões que não podem ser interpretadas [1]. Na legenda, são mostrados os “ranks” de 1 a 5, que são os cinco modelos estruturais da sequência que foram gerados pelo Alphafold. Os autores discutem melhor no artigo da ferramenta.

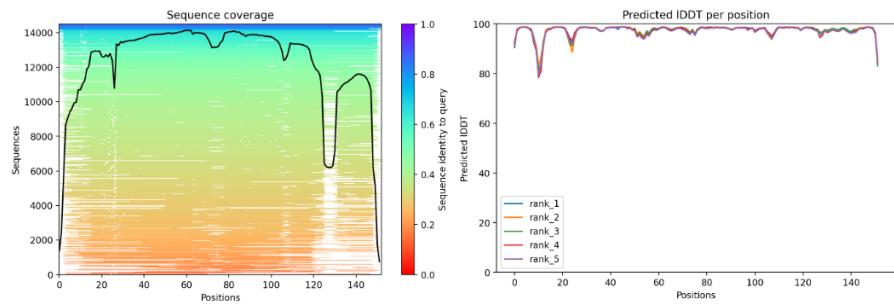


Figura 23.11: Gráficos mostrando o alinhamento múltiplo de sequências por resíduo (A - esquerda), ou seja, o número médio de leituras que se alinham ou “cobrem” bases da referência (sequência de input), e a estimativa confiabilidade por residuo (B - direita).

#### Arquivos gerados:

Ao final da prática, você também possui a opção de salvar os resultados da sua modelagem. Para isso, basta rodar a célula “*Package and download results*”, onde você pode salvar o arquivo em formato .zip onde desejar. Ao descompactá-lo, terão diversos arquivos dentro da pasta, dos quais você utilizará para sua análise:

- O arquivo “nomeDaEstrutura\_coverage.png”, contendo o gráfico de cobertura de sequências (sequence coverage);
- O arquivo “nomeDaEstrutura\_pae.png”, contendo os gráficos de erro de alinhamento predito (PAE);
- O arquivo “nomeDaEstrutura\_plddt.png”, contendo o gráfico de Local Distance Difference Test;

Os arquivos das estruturas geradas, em formato pdb. Eles estarão nomeados como “nomeDaEstrutura\_unrelaxed\_rank\_001\_alphaFold2\_ptm\_model\_5\_seed\_000.pdb”, sendo numerados de acordo com a classificação de melhor estrutura.

#### **23.4.4 Etapa 4: Validação das estruturas previstas (Bônus)**

Após obter o modelo predito da proteína, é necessário verificar suas semelhanças e scores (pontuações) de modelagem no PDB. No nosso caso, o modelo ainda não foi determinado pelos métodos convencionais (Cristalografia e difração de raio-X, Ressonância Magnética Nuclear ou Cryo-EM). Assim, é importante realizar métodos de alinhamento de sequências tridimensionais por proteínas homólogas a esta que já se encontram resolvidas.

Independente do nosso resultado, a etapa final de validação é crucial para o aprofundamento do nosso estudo e novos achados, pois compara o resultado obtido com dados experimentais, localização dos resíduos de sítios ativos, ligação de ligantes e outros detalhes. Ao validar este resultado através de comparações, e utilizando algumas ferramentas comumente utilizadas para este fim, garantimos maior qualidade e confiabilidade à pesquisa. Algumas dessas ferramentas incluem:

- MolProbity (<https://pubmed.ncbi.nlm.nih.gov/29067766/>), que identifica problemas de geometria e estereoquímica;
- Verify3D (<https://www.doe-mbi.ucla.edu/verify3d/>), que compara a proteína com outras já bem resolvidas;
- Prosa (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1933241/>), que identifica conformações atípicas na estrutura.

### **23.5 Prática de modelagem de complexos**

Iniciando a prática, é importante ressaltar que o AlphaFold não realiza o docking molecular propriamente dito. Diferentemente das ferramentas de docking comumente usadas, não calcula a afinidade de ligação entre as moléculas, como também não busca, no espaço conformacional, a melhor conformação por meio de algoritmos de busca. O que a ferramenta realiza é a modelagem de ambos os elementos, tanto receptor (proteína) quanto do ligante (proteína/peptídeo) em complexo, tentando modelar a região de ligação entre receptor-ligante.

Para exemplificar, realizaremos a modelagem do complexo TNF- $\alpha$ , uma citocina pró-inflamatória de fundamental importância para o processo de defesa do organismo, com um de seus receptores, TNR1 [14, 15].

### **23.5.1 Etapa 1: Adquirindo as sequências**

Para iniciar o tutorial, o primeiro passo é baixar as sequências das proteínas que usaremos. Nesta etapa vamos acessar o PDB (Protein Data Bank), no endereço <https://www.rcsb.org/>. Depois, vamos buscar pela estrutura de TNF-, utilizando o identificador 1TNF (Figura 23.12).



*Figura 23.12: Buscando pela estrutura de TNF- $\alpha$ .*

Após encontrarmos a estrutura no banco de dados, vamos baixar a sequência de aminoácidos. Para isso, clique na primeira estrutura encontrada. Depois de abrir, você deve clicar no botão lateral Download Files. Em seguida, clique em Fasta Sequence para baixar o arquivo com a sequência de aminoácidos (Figura 23.13). Após baixarmos a sequência de TNF-, vamos realizar o mesmo procedimento para o receptor TNFR1, nesse caso vamos buscar ar pelo ID 1TNR no PDB, e baixar a sequência selecionando a sequência correspondente a cadeia B.



Figura 23.13: Baixando a sequência de TNF- $\alpha$ .

### 23.5.2 Etapa 2: Modelando o complexo

Após baixar as sequências, vamos começar a modelar o complexo. Para isso, abriremos o ColabFold. Em seguida, vamos colar as sequências na aba query\_sequence. Como a TNF- $\alpha$  é um trímero, vamos colar a sequência da citocina três vezes, separando cada sequência utilizando “:” (dois pontos). A seguir, vamos colar a sequência do receptor 1TNR na aba query\_sequence, após a sequência de TNF- $\alpha$ . Após colar a sequência, podemos inserir um nome no trabalho na aba “jobname”. Clique em “Ambiente de execução” e depois clique em “Executar tudo”. Vamos aguardar que todo o script seja executado para analisar os resultados.

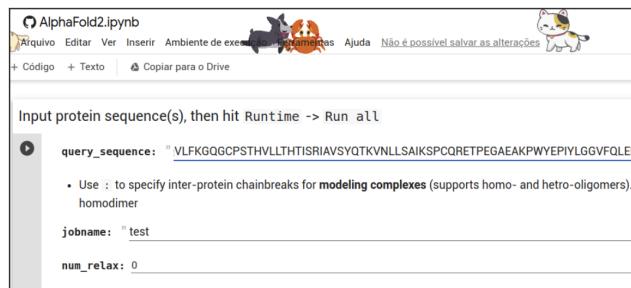


Figura 23.14: Iniciando a modelagem do complexo.

### 23.5.3 Etapa 3: Avaliando os resultados

O primeiro gráfico a ser observado é o da Figura 23.15. Nele, podemos observar a cobertura das sequências encontradas como template para modelagem. Algo importante a se observar é que, diferentemente da modelagem de uma proteína de cadeia única, o gráfico é dividido entre as cadeias. No caso do nosso gráfico, ele está dividido em quatro cadeias: as três primeiras correspondem a cadeias da proteína TNF- e a última cadeia corresponde à TNF- $\alpha$ . Na parte superior, encontramos mais um bloco de cobertura, relacionado às regiões de ligação entre as cadeias das proteínas. Nota-se que o ColabFold encontrou sequências com boa cobertura, o que nos leva a crer que a modelagem será bem executada.

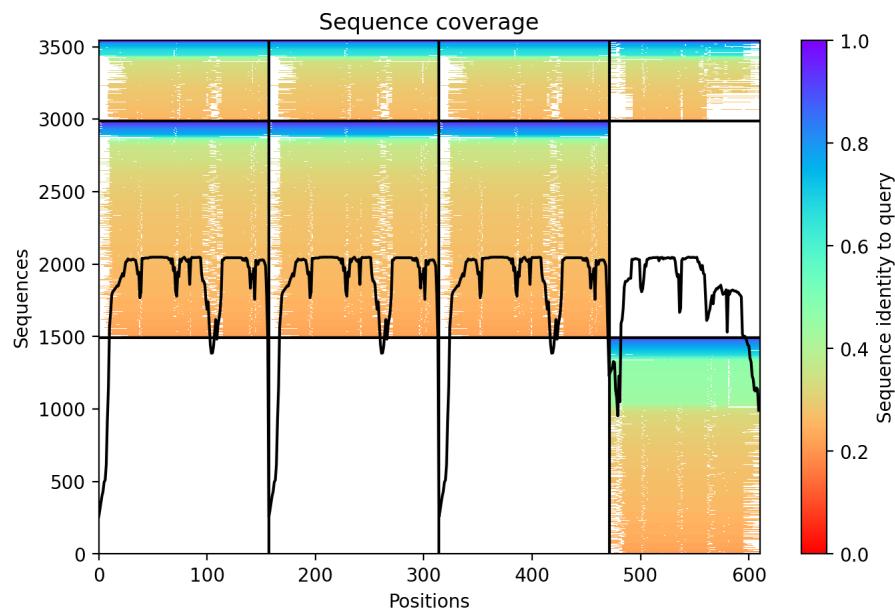


Figura 23.15: Gráfico de cobertura das sequências de entrada.

Em seguida, vamos avaliar a distribuição pLDDT (Figura 23.16) ao longo da estrutura. Nota-se que grande parte da estrutura possui um pLDDT > 90, o que demonstra que a estrutura foi modelada com alta confiança.

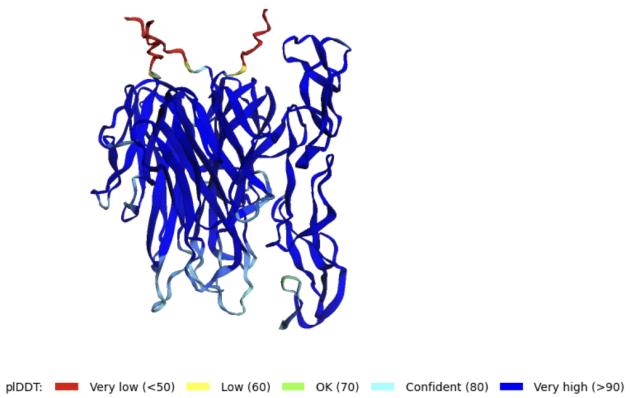


Figura 23.16: Visualizando o pLDDT ao longo da estrutura.

A ferramenta gera cinco modelos e, na Figura 23.17, observamos o erro predition ou PAE (*Predicted aligned error*). Nos eixos x e y encontramos a sequência e, na parte horizontal, temos a divisão de cadeias A, B, C e D. Essa métrica varia de 0 a 30 e, quanto menor o resultado, com mais confiança o complexo foi modelado. Como podemos observar na figura abaixo, a estrutura foi bem modelada, uma vez que o erro predition é baixo em todas as estruturas, tanto da sequência das cadeias quanto das regiões de ligação entre elas.

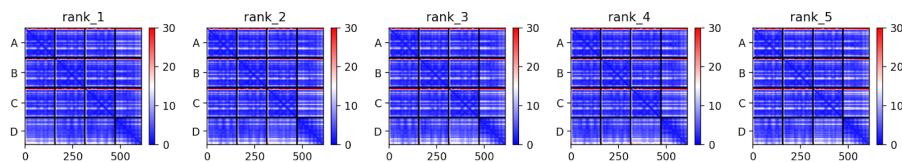


Figura 23.17: Visualizando o erro predition.

Por fim, podemos visualizar o gráfico de IDDT (Figura 23.18) por resíduo para os cinco modelos gerados. Nele, observa-se que os cinco modelos tiveram seus resíduos avaliados com IDDT acima de 70, com algumas regiões com IDDT acima de 80. Nota-se que o complexo foi bem modelado, o que já era esperado, uma vez que foram encontrados moldes com alta cobertura.

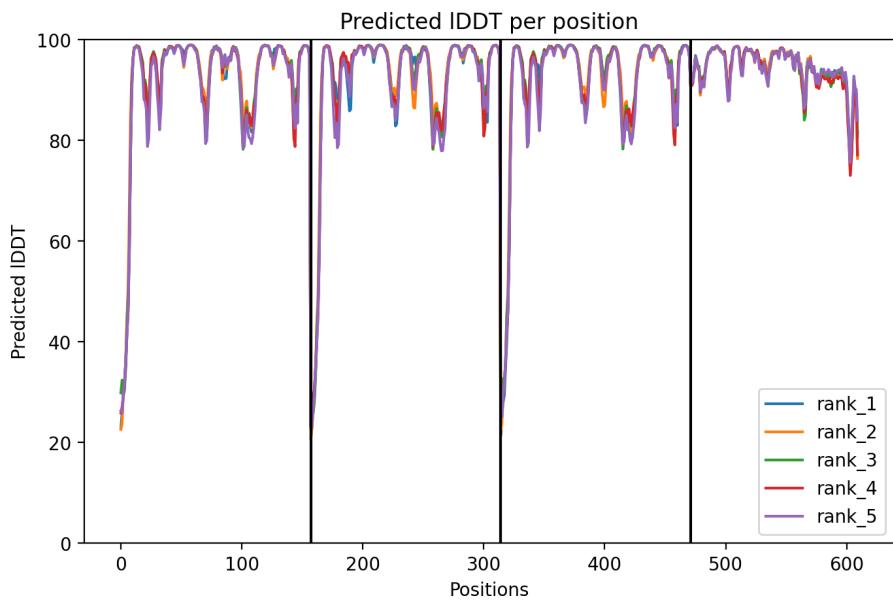


Figura 23.18: IDDT predito para os cinco modelos gerados.

## 23.6 Conclusão

Por fim, é importante observar que a ferramenta apresenta algumas limitações:

O AlphaFold não é recomendado para análise mutacional, uma vez que utiliza dados já existentes para a construção dos modelos estruturais e, portanto, sua previsão terá uma baixa confiança.

Os modelos construídos não consideram os ligantes: a ferramenta prevê apenas a cadeia peptídica principal, não as estruturas de cofatores, metais e modificações co- e pós-traducionais. Por outro lado, como o modelo é treinado a partir de modelos PDB, muitas vezes com essas modificações anexadas, a estrutura prevista é frequentemente consistente com a estrutura esperada na presença de íons ou cofatores” [16].

As proteínas flexíveis não são modeladas com alta qualidade. Pedaços flexíveis, como regiões C e N-terminais, possuem baixa qualidade.

Apesar das limitações citadas, como vimos neste artigo, o AlphaFold é uma ferramenta revolucionária no que diz respeito à previsão de estruturas de proteínas e tem potencial para melhorar ainda mais à medida que for treinada com mais estruturas. Além disso, é importante ressaltar que essa ferramenta pode ser utilizada para diversos estudos de alto impacto, uma vez que prevê estruturas até então desconhecidas. Dentre eles: desenvolvimento de fármacos, estudo de variantes patogênicas, auxiliar no entendimento de algumas doenças e no desenvolvimento de vacinas [17].

Saiba mais 23.1

Este artigo está disponível em <https://bioinfo.com.br/alphafold-2-revolucionando-a-modelagem-de-estruturas-tridimensionais-de-macromoleculas/>

## 23.7 Referências

[1] AlphaFold. Disponível em: <https://alphafold.ebi.ac.uk/>. Acesso em: 31 de julho de 2023.

[2] Protein Data Bank. Disponível em: <https://www.rcsb.org/>. Acesso em: 31 de julho de 2023.

[3] AlQuraishi, M. AlphaFold at CASP13, Bioinformatics, Volume 35, Issue 22, November 2019, Pages 4862–4865, <https://doi.org/10.1093/bioinformatics/btz422>.

[4] AlQuraishi, M. AlphaFold @ CASP13: “What just happened?. Some Thoughts on a Mysterious Universe. Disponível em: <https://moalquraishi.wordpress.com/2018/12/09/alphafold-casp13-what-just-happened/>. Acesso em: 31 de julho de 2023.

[5] Ramachandran, G. N., Ramakrishnan, C., Sasisekharan, V. (1963). Stereochemistry of polypeptide chain configurations. Journal of Molecular Biology, 7, 95–99. [https://doi.org/10.1016/s0022-2836\(63\)80023-6](https://doi.org/10.1016/s0022-2836(63)80023-6).

[6] Branden, C. I., Tooze, J. (1999). Introduction to Protein Structure (2nd ed.). Garland Science. Disponível em: <https://www.routledge.com/Introduction-to-Protein-Structure/Branden-Tooze/p/book/9780815323051>.

[7] Fang, C., Shang, Y., Xu, D. (2018). Prediction of Protein Backbone Torsion Angles Using Deep Residual Inception Neural Networks. IEEE/ACM Transactions on Computational

Biology and Bioinformatics, 10.1109/TCBB.2018.2814586.

<https://doi.org/10.1109/TCBB.2018.2814586>.

[8] John Jumper et al., “AlphaFold 2”. Apresentação na CASP 14. Dez 2020. Disponível em: [https://predictioncenter.org/casp14/doc/presentations/2020\\_201T\\_S\\_predictor\\_AlphaFold2.pdf](https://predictioncenter.org/casp14/doc/presentations/2020_201T_S_predictor_AlphaFold2.pdf).

[9] Jumper, J, Evans, R, Pritzel, A, et al. Applying and improving AlphaFold at CASP14. Proteins. 2021; 89(12): 1711- 1721. doi:10.1002/prot.26257.

[10] Jumper, J. et al., conference abstract (December 2020).

[11] Jumper, J., Evans, R., Pritzel, A. et al. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589 (2021). <https://doi.org/10.1038/s41586-021-03819-2>.

[12] Mirdita, M., Schütze, K., Moriwaki, Y. et al. ColabFold: making protein folding accessible to all. Nat Methods 19, 679–682 (2022). <https://doi.org/10.1038/s41592-022-01488-1>.

[13] Nature. ‘It will change everything’: DeepMind’s AI makes gigantic leap in solving protein structures. Disponível em: <https://www.nature.com/articles/d41586-020-03348-4>. Acesso em: 29 ago. 2023.

[14] Eck, M. J., and Sprang, S. R. “The structure of tumor necrosis factor- at 2.6 Å resolution: Implications for receptor binding.” Journal of Biological Chemistry.

[15] Banner D. W. et al. Crystal structure of the soluble human 55 kd TNF receptor-human TNF beta complex: implications for TNF receptor activation. Cell. 1993 May 7;73(3):431-45. doi: 10.1016/0092-8674(93)90132-a. PMID: 8387891.

[16] O AlphaFold e o desenvolvimento de vacinas. OnlineBioinfo – Comunicação científica em Bioinformática. Disponível em: <https://onlinebioinfo.com/2022/02/15/o-alphafold-e-o-desenvolvimento-de-vacinas/>. Acesso em: 31 de julho de 2023.

[17] Thornton, J.M., Laskowski, R.A. Borkakoti, N. AlphaFold heralds a data-driven revolution in biology and medicine. Nat Med 27, 1666–1669 (2021). <https://doi.org/10.1038/s41591-021-01533-0>.

[18] Mariano, D. AlphaFold e a busca pelo Santo Graal da Biologia Molecular. In: BIOINFO #02 - Revista Brasileira de Bioinformática e Biologia Computacional. Vol. 2, 10, 162-167 (2022). Disponível em: <https://bioinfo.com.br/alphafold-e-a-busca-pelo-santo-graal-da-biologia-molecular>. doi: 10.51780/978-65-992753-5-7-10

[19] Finkelstein, A.; Finkelstein, A.V. Protein Folding: Enigma and Solution. Encyclopedia. Available online: <https://encyclopedia.pub/entry/8524> (accessed on 15 November 2022); <https://medium.com/turing-talks/alphafold-2-entenda-seu-funcionamento-e->

implica%C3%A7%C3%B5es-para-biologia-computacional-2ace80b1b70b. Acesso em: 15 de Novembro de 2022.

[20] Thornton, J.M., Laskowski, R.A. Borkakoti, N. AlphaFold heralds a data-driven revolution in biology and medicine. *Nat Med* 27, 1666–1669 (2021).  
<https://doi.org/10.1038/s41591-021-01533-0>.

# 24 TERMODINÂMICA DE PROTEÍNAS: COMO AS PROTEÍNAS SE ENOVELAM?

## Autores 24.1

Alisson Clementino da Silva , Bruno Rafael Pereira Nunes , Joicymara Xavier 

Revisão: Aline Sampaio Cremonesi , Rafael Lemos 

## Cite este artigo 24.1

Silva, AC; Nunes, BRP; Xavier, J. *Termodinâmica de Proteínas: como as proteínas se enovelam?* BIOINFO. ISSN: 2764-8273. Vol. 3. p.24 (2023). doi: 10.51780/bioinfo-03-24

### Resumo 24.1

Neste artigo, você irá aprender sobre como o que são proteínas e como elas se enovelam.

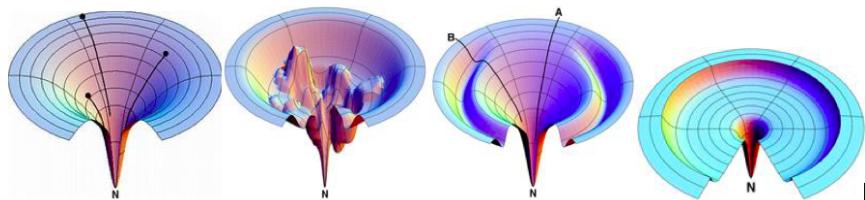
**A**s proteínas são as biomoléculas mais abundantes dos seres vivos. São compostos orgânicos formados por carbono, oxigênio, nitrogênio, hidrogênio e enxofre, constantemente sintetizados no interior das células [1]. Desempenham diversas funções necessárias à manutenção biológica dos organismos, como transporte e armazenamento de moléculas, regulação da expressão gênica, defesa, regulação do metabolismo, estruturação, dentre outras. Formadas por combinações sequenciais de resíduos de 20 aminoácidos diferentes, as proteínas são polímeros que, naturalmente, apresentam características estruturais, físicas e químicas distintas. Dessa forma, a determinação de suas estruturas tridimensionais é imprescindível para entender suas funcionalidades nos inúmeros processos biológicos que participam [1-2].

Ao serem sintetizadas, as proteínas assumem uma estrutura termodinamicamente estável (estrutura nativa) por meio do enovelamento (ou dobramento) de suas cadeias, que constitui um importante processo iniciado na etapa de tradução, que ocorre no ribossomo. A hierarquização das fases do enovelamento possibilita o entendimento das subestruturas conservadas evolutivamente e das estruturas proteicas em quatro níveis distintos [1-2].

O enovelamento das proteínas é um problema rapidamente solucionado na natureza [1]. Em geral, quanto maior a cadeia polipeptídica, maior será o tempo para a proteína enovelar-se. Entender como o estado nativo é encontrado em pouco tempo instigou Levinthal (1968) a postular que as inúmeras conformações possíveis não são testadas aleatoriamente até que se ache a mais estável. O paradoxo de Levinthal considera a sequência temporal dos eventos intermediários que ocorrem entre os estados desenovelado e nativo [3]. O pressuposto de que aleatoriamente as possíveis conformações vão sendo testadas à medida que a proteína emerge do ribossomo, foi contraposto com o tempo necessário para validar quase todas como não nativas [4].

Considerando que se a cada ~10-13 segundos (tempo de uma vibração molecular) uma conformação fosse assumida, para uma proteína formada por 100 resíduos de aminoácidos, seria necessário um período de tempo maior que a idade do universo para que todas as posições fossem testadas [1-3]. Levinthal então sugeriu que o enovelamento das estruturas devia ser acelerado e guiado pela rápida formação de interações locais de curto e longo alcance (ligações iônicas, ligações hidrogênio, interações hidrofóbicas e interações de van der Waals) [1]. Essas interações, com destaque para o efeito das interações hidrofóbicas, submetem a estrutura a vias de enovelamento que limitam o acesso da proteína a determinadas conformações não favoráveis, direcionando assim a cadeia para as posições mais favoráveis e de menor energia livre [4].

A dinâmica do enovelamento pode ser visualizada como um funil (Figura 24.1) que descreve a tendência termodinâmica da estrutura em assumir uma conformação de menor variação energética [1]. Essa variação energética é expressa em kcal/mol (variação da energia livre de Gibbs,  $\Delta\Delta G$ ). O início do processo é representado pela área superior do funil, onde a energia livre se encontra em maior grau. A partir da formação dos estados intermediários semi estáveis, depressões ao longo das paredes do funil representam as variações conformacionais e energéticas que ocorrem durante todo o processo. Convergindo para o fundo do funil, o ponto de menor variação de energia livre conformacional é encontrado, e o conjunto de intermediários é reduzido a uma conformação nativa.



*Figura 24.1: Paisagem termodinâmica de energia livre em forma de funil. As estruturas nativas em seu mínimo global guiam cada molécula de um conjunto de cadeias polipeptídicas desenoveladas de alta energia, por meio de diferentes vias de enovelamento até o fundo do funil, onde a estrutura estará condensada em uma forma que apresente a menor variação de energia livre de Gibbs. Fonte: [5].*

O ambiente circundante e a sequência de aminoácidos são fatores cruciais para que a proteína assuma uma determinada conformação [1]. Por meio dos experimentos de Anfinser (1973), foi demonstrada a desnaturação e renaturação *in vitro* de uma proteína de cadeia simples, a Ribonuclease (RNase) bovina [3]. O estado enovelado da enzima RNase pode ser atestado pela mensuração de sua atividade enzimática. Dessa forma, ele preparou amostras enzimáticas usando combinações de dois reagentes diferentes: ureia 8M ( $\text{CO}(\text{NH}_2)_2$ ), como agente desnaturante, e 0.2 M beta-mercaptoetanol ( $\beta\text{ME}$ ), como agente redutor [6].

Atestando a perda total de atividade da enzima, Anfinser removeu o ME por diálise, e a ureia em seguida. Quando a ureia e o ME foram removidos, a RNase desnaturada voltou à estrutura nativa correta, de modo que as ligações dissulfeto se restabeleceram nos mesmos lugares [3-6]. Após a renaturação da estrutura, Anfinser atestou que 1% da atividade catalítica da proteína nativa estava retida na proteína renaturada, a partir da comparação das atividades em ambos os estados. No entanto, ao adicionar quantidades catalíticas do ME para que a atividade fosse restaurada, a atividade enzimática da RNase bovina foi restaurada. Dessa forma, ele concluiu que a informação necessária para enovelar uma proteína está contida em sua sequência de aminoácidos. Estudos posteriores demonstraram que o princípio de renaturação não se aplica a diversas proteínas; algumas apresentam seu processo de enovelamento assistido por outras proteínas especializadas, conhecidas como chaperonas, se diferindo então do processo da RNase [1].

Tratando da renaturação de moléculas pequenas, como no experimento de Anfinsen, as alfa-hélices e conformações beta, são as estruturas secundárias formadas primeiramente, devido a uma série de restrições que norteiam seus surgimentos [2]. Em seguida, interações iônicas de grupos carregados e ligações de longo alcance são estabelecidas, além das interações hidrofóbicas que promovem a agregação das partes apolares dos resíduos, conferindo uma estabilidade entrópica a formas enoveladas intermediárias. Por fim, a dinâmica de forças termodinâmicas faz a estrutura assumir uma conformação de menor variação energética, fundindo os estados intermediários em um estado de maior estabilidade [1].

A estabilidade, como tendência de manter a conformação nativa, depende das forças atuantes nas proteínas para induzir o processo de enovelamento [1]. O efeito hidrofóbico é considerado a principal força indutora, a partir da interação das porções apolares, formadas por aminoácidos hidrofóbicos. Outro tipo de interação observada durante o enovelamento são as ligações dissulfeto, resultantes de processo oxidativo, que permitem a ligação de cisteínas não adjacentes [7-8].

Diante da complexidade das estruturas proteicas, todo conhecimento acerca dos componentes básicos de seus arranjos mais frequentes não oferece compreensão suficiente acerca dos mecanismos que levam uma cadeia polipeptídica estruturada aleatoriamente no espaço a assumir sua conformação nativa. Então, a determinação das estruturas tridimensionais por metodologias experimentais e *in silico* tornou-se fundamental para entender como as proteínas se enovelam [8-9].

## **24.1 Técnicas de determinação de estruturas 3D**

À medida que o conhecimento avança, novas metodologias são desenvolvidas e atualizadas, visando a predição de estruturas tridimensionais de proteínas a partir de métodos experimentais ou computacionais [1]. A necessidade de ter estruturas proteicas resolvidas (ou elucidadas) deriva da importância que a sua conformação tem em caráter essencial na funcionalidade [8].

### **24.1.1 Técnicas experimentais**

Dentre as técnicas experimentais que vêm sendo aplicadas para a obtenção de informações sobre as estruturas proteicas, destacam-se a difração de raios-X, a Calorimetria de Varredura Diferencial (do inglês Differential Scanning Calorimetry, DSC), a Ressonância Magnética Nuclear (RMN) e o Dicroísmo Circular (CD). O alto custo é a principal desvantagem da aplicação dessas técnicas [1,7,8].

### **24.1.2 Difração de raios-X**

A difração de raios-X permite que a estrutura de proteínas seja determinada em uma escala quase atômica. O ensaio é baseado na geração de cristais contendo

proteínas que recebem radiação incidente de um comprimento de onda específico. O raio-X é difratado pelos elétrons que estão distribuídos no cristal, e assim, é possível inferir da posição dos próprios núcleos, que também podem ser determinados por difração com feixe de nêutrons [1-7].

#### **24.1.3 Calorimetria de Varredura Diferencial**

A Calorimetria de varredura diferencial (DSC) é uma técnica utilizada para caracterizar a estabilidade de proteínas ou outras biomoléculas diretamente em sua forma nativa. Elas são aquecidas a uma taxa de varredura constante, que causa a absorção de calor a partir do desenovelamento da estrutura, resultando em um gradiente térmico ( $T$ ) entre as células. Ainda, os modelos termodinâmicos podem ser ajustados aos dados para obter a energia livre de Gibbs ( $G$ ), a entalpia calorimétrica ( $H_{cal}$ ), a entalpia de van't Hoff ( $H_{vH}$ ), a entropia ( $S$ ) e a mudança da capacidade de calor ( $C_p$ ) associada à transição [7].

#### **24.1.4 Ressonância Magnética Nuclear**

Ressonância magnética nuclear (RMN) é um ensaio realizado com macromoléculas em solução sob influência de um campo eletromagnético estático potente e baseia-se na liberação de um pulso de energia eletromagnética, em diferentes ângulos na solução. Parte da energia é absorvida à medida que o núcleo das moléculas muda do estado de menor energia, que corresponde à orientação paralela do dipolo magnético gerado pelo momento angular do spin nuclear, para o estado de maior energia, com orientação antiparalela ao campo. O espectro resultante apresenta informações sobre a identidade do núcleo e o ambiente químico das imediações. A RMN torna-se vantajosa por também esclarecer mudanças conformacionais no enovelamento e interações com outras moléculas [7].

#### **24.1.5 Dicroísmo Circular**

O Dicroísmo Circular (CD) é utilizado para mensurar a quantidade e tipo de estruturas secundárias presentes em solução, pois tem se mostrado um ensaio sensível, principalmente à estruturas alfa-hélice, folhas-beta e desordenadas, em

virtude dos diferentes espectros gerados em uma faixa de comprimentos de onda. O CD é causado por diferenças na absorção de luz entre os componentes em sentido horário e anti-horário de um feixe de luz polarizada que atravessa uma solução opticamente ativa [7-8].

## 24.2 Métodos computacionais

Os métodos computacionais para predição de estruturas tridimensionais de proteínas podem ser separados em quatro categorias: Modelagem comparativa por homologia, métodos de reconhecimento de padrões de enovelamento, métodos ab initio ou de novo [9-10].

### 24.2.1 Métodos de modelagem comparativa

A modelagem comparativa parte do pressuposto que há relação evolutiva entre duas sequências, que resultam em estruturas tridimensionais similares. Essa abordagem garante alta precisão para os modelos gerados quando se há semelhanças. Como a comparação depende de estruturas já determinadas experimentalmente, diferentes padrões de enovelamento não podem ser determinados por estes métodos (Tabela 24.1) [8-10].

*Tabela 24.1: Exemplos de Preditores de estrutura por modelagem comparativa. Fonte: próprio autor.*

PREDITOR	DESCRIÇÃO	LINK
Modeller [11]	Modelagem comparativa de estrutura de proteína por satisfação de restrições espaciais. O usuário fornece um alinhamento de uma sequência a ser modelada com estruturas relacionadas conhecidas e o MODELLER calcula automaticamente um modelo contendo todos os átomos	<a href="https://salilab.org/modeller/">https://salilab.org/modeller/</a>
SWISS-MODEL [12]	É um serviço web dedicado à modelagem de homologia de estruturas de proteínas. O servidor constrói modelos a partir da (1) identificação do(s) modelo(s) estrutural(is), do (2) alinhamento da sequência alvo e estrutura(s) do modelo, da (3) construção do modelo e avaliação da qualidade.	<a href="https://swissmodel.expasy.org/">https://swissmodel.expasy.org/</a>

É um serviço web dedicado à modelagem de homologia de estruturas de proteínas. O servidor constrói modelos a partir da (1) identificação do(s) modelo(s)

estrutural(is), do (2) alinhamento da sequência alvo e estrutura(s) do modelo, da (3) construção do modelo e avaliação da qualidade.

### 24.3 Métodos de reconhecimento de padrões de enovelamento

Também conhecidos como de fold recognition, estes métodos (Tabela 2) consideram a estrutura tridimensional da proteína evolutivamente mais conservada do que sua sequência correspondente [9-10]. Assim, ainda que não haja alta similaridade entre sequências, as estruturas tridimensionais conhecidas podem apresentar semelhanças que permitam inferências na estrutura estudada.

*Tabela 24.2: Exemplos de preditores de estrutura por reconhecimento de padrões. Fonte: próprio autor.*

PREDITOR	DESCRIÇÃO	LINK
HHpred [13]	Utiliza sequências ou alinhamento de sequência múltipla como entrada e procura por homólogos remotos em uma variedade de bancos de dados, como PDB, SMART e Pfam.	<a href="https://toolkit.tuebingen.mpg.de/tools/hhpred">https://toolkit.tuebingen.mpg.de/tools/hhpred</a>
SPARKS-X [14]	é um aprimoramento do reconhecimento do enovelamento de proteínas, empregando correspondência baseada em probabilística entre propriedades estruturais nativas e dos modelos gerados.	<a href="https://sparks-lab.org/server/sparks-x/">https://sparks-lab.org/server/sparks-x/</a>
Phyre e Phyre2 [15]	Protein Homology/analogY Recognition Engine	<a href="http://www.sbg.bio.ic.ac.uk/phyre2/">http://www.sbg.bio.ic.ac.uk/phyre2/</a>

### 24.4 Métodos ab initio ou de novo

Os métodos ab initio ou de novo (Tabela 3) baseiam-se em informações extraídas das estruturas tridimensionais depositadas em bases de dados, utilizando princípios físicos para gerar modelos de estruturas tridimensionais. A utilização dessas abordagens permite que padrões não recorrentes sejam preditos a partir de comparações realizadas por fragmentos [9-10]. Além disso, os princípios da hipótese de Anfinsen são utilizados na construção dos modelos.

#### 24.4.1 AlphaFold

Quando, em 2018, os resultados da Avaliação Crítica de Predição de Estruturas (CASP) foram anunciados, o DeepMind, um grupo de pesquisa de aprendizado de

*Tabela 24.3: Exemplos de predtores de estrutura de novo. Fonte: próprio autor*

PREDITOR	DESCRIÇÃO	LINK
I-TASSER [16]	É uma abordagem hierárquica para previsão da estrutura de proteínas e anotação de função baseada em estrutura. Ele identifica modelos estruturais do PDB pela abordagem de segmentação múltipla LOMETS, com modelos atômicos completos construídos por simulações interativas de montagem de fragmentos baseadas em modelos.	<a href="https://zhanggroup.org/I-TASSER/">https://zhanggroup.org/I-TASSER/</a>
ROSETTA [17]	É um software utilizado para prever e projetar estruturas de proteínas, mecanismos de dobramento de proteínas e interações proteína-proteína, a partir de estruturas depositadas em bases de dados.	<a href="https://www.rosettacommons.org/software">https://www.rosettacommons.org/software</a>
AMBER [18]	é um conjunto de programas de simulação, que utiliza um conjunto de campos de força mecânico molecular, de domínio público, para a simulação de biomoléculas.	<a href="https://ambermd.org/">https://ambermd.org/</a>
CHARMM [19]	É um programa de simulação molecular que visa, principalmente, sistemas biológicos. Disponibiliza um conjunto abrangente de funções de energia, uma variedade de métodos de amostragem aprimorados, para análise e construção de modelos.	<a href="https://www.charmm.org/">https://www.charmm.org/</a>
GROMOS [20]	É um pacote desenvolvido para a modelagem dinâmica de biomoléculas usando os métodos de dinâmica molecular, dinâmica estocástica e minimização de energia.	<a href="https://www.gromos.net/">https://www.gromos.net/</a>

máquina do Google, recebeu o primeiro lugar pelo desenvolvimento do AlphaFold. O AlphaFold é um algoritmo baseado em redes neurais profundas, no qual modelos estruturais de proteínas foram gerados por meio do uso de previsões de distância ou contato entre pares de resíduos de aminoácidos [9].

O AlphaFold foi treinado a partir de um conjunto de dados públicos de proteínas com estruturas tridimensionais conhecidas e um banco de dados de sequências sem estruturas conhecidas. Os modelos gerados apresentaram alta precisão, sendo considerado um grande passo para solucionar o problema de entendimento sobre o enovelamento de proteínas [9]. Além de contribuir com pesquisas em andamento, estudos ligados a doenças também foram beneficiados com a publicação dos autores, que disponibilizaram uma base de dados com mais de 350 mil estruturas referentes ao proteoma humano e a outros organismos [21].

Se quiser saber mais detalhes sobre o AlphaFold e como utilizá-lo, leia os artigos sobre o assunto que já estão disponíveis na revista BIOINFO:

- AlphaFold e a busca pelo Santo Graal da Biologia Molecular
- AlphaFold 2: revolucionando a modelagem de estruturas 3D de macromoléculas

## 24.5 Conclusão

As proteínas são as biomoléculas mais atuantes nos seres vivos e dependem da sua estrutura nativa para desempenhar suas funções. O enovelamento, sendo o processo pelo qual cadeias polipeptídicas assumem a respectiva conformação, é um fenômeno já explorado e que gradualmente está sendo elucidado. As metodologias, experimentais e computacionais, tornaram-se fundamentais para o estudo das estruturas e mapeamento de diferentes padrões de enovelamento. A exploração das subestruturas secundárias permitiu contribuições de uma perspectiva termodinâmica, que está sendo aplicada para responder questões de estabilidade das estruturas proteicas no ambiente celular, contribuindo para um maior entendimento das forças que atuam na proteína.

O objetivo deste artigo foi apresentar uma resposta introdutória e panorâmica sobre o enovelamento de proteínas. Nos próximos artigos, discutiremos ainda mais sobre a estrutura tridimensional, estabilidade termodinâmica, mutações e bases de dados associadas a esses conteúdos.

Saiba mais 24.1

Este artigo está disponível em <https://bioinfo.com.br/termodinamica-de-proteinas-como-as-proteinas-se-enovelam/>

## 24.6 Referências

- [1] Nelson, D. L.; Cox, M. M. Princípios de bioquímica de Lehninger. 6. Ed. Porto Alegre: Artmed, 2014.

- [2] Lemos, R. P.; Santos, P. H.; Rocha, A. Introdução à Biologia Estrutural de Proteínas. In: BIOINFO - #3 - Revista BIONFO. Disponível em: <https://bioinfo.com.br/introducao-a-biologia-estrutural-de-proteinas>. Vol. 3. Acesso em 14 de Agosto de 2023.
- [3] Tanouye, F. T. Enovelamento de proteínas e ligações de hidrogênio – estudo de modelos mínimos. 22 Sep. 2017.
- [4] Levinthal, C. Are there pathways for protein folding? *Journal of Chimie Physique*, vol. 65, pp. 44–45.v. 1968.
- [5] Ken A. Dill. (CC-BY 4.0) Obtido em: [https://commons.wikimedia.org/wiki/File:Funnel-shaped\\_energy\\_landscape.png](https://commons.wikimedia.org/wiki/File:Funnel-shaped_energy_landscape.png). Acesso em: 04 de Setembro de 2023.
- [6] Anfinsen C. B., Haber E., Sela M., White F. H. Jr. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proceedings of the National Academy of Sciences of the United States of America*, 47:1309–1314, 1961.
- [7] Devlin, T.M. Manual de Bioquímica com Correlações Clínicas, 7<sup>a</sup> ed., Ed. Blucher, 2011.
- [8] Verli, H. (Org). Bioinformática: da biologia à flexibilidade molecular. Sociedade Brasileira de Bioquímica e Biologia Molecular. 2014.
- [9] Marques, F. B. Predição da estrutura tridimensional de proteínas utilizando o método CReF com informações de contato. Dissertação (Mestrado) – Programa de Pós-Graduação em Ciências da Computação, PUCRS. Porto Alegre. 2021.
- [10] Dorn, Márcio. Uma proposta para a predição computacional da estrutura 3D aproximada de polipeptídeos com redução do espaço conformacional utilizando análise de intervalos. Dissertação (Mestrado em Ciência da Computação) – Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, 2008.
- [11] Sali, A. e Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology*, vol. 234, pp. 779–815. 1993.
- [12] Webb, B. e Sali, A. Comparative Protein Structure Modeling Using MODELLER. *Current Protocols in Bioinformatics*, vol. 54, pp. 5.6.1–5.6.37. 2016.
- [13] Arnold, K., Bordoli, L., Kopp, J. e Schwede, T. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics*, vol. 22, pp. 195–201. 2006.
- [14] Söding, J. Protein homology detection by hmm–hmm comparison. *Bioinformatics*, vol. 21, pp. 951–960. 2005.

- [15] Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. e Sternberg, M. J. (Mai, 2015). The phyre2 web portal for protein modeling, prediction and analysis. *Nature Protocols*, vol. 10, pp. 845–858. 2015.
- [16] Roy, A., Kucukural, A. e Zhang, Y. I-tasser: a unified platform for automated protein structure and function prediction. *Nature Protocols*, vol. 5, pp. 725–738. 2010.
- [17] Rohl, C. A., Strauss, C. E., Misura, K. M. e Baker, D. Protein structure prediction using rosetta. In: *Numerical Computer Methods, Part D*, vol. 383, pp. 66–93. Academic Press, 1 ed. 2004.
- [18] Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W. e Kollman, P. A. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *Journal of the American Chemical Society*, vol. 117, pp. 5179–5197. Maio de 1995.
- [19] Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. e Karplus, M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry*, vol. 4, pp. 187–217. jan de 1983.
- [20] Christen, M., Hünenberger, P. H., Bakowies, D., Baron, R., Bürgi, R., Geerke, D. P., Heinz, T. N., Kastenholz, M. A., Kräutler, V., Oostenbrink, C., Peter, C., Trzesniak, D. e Gunsteren, W. F. v. The GROMOS software for biomolecular simulation: GROMOS05. *Journal of Computational Chemistry*, vol. 26, pp. 1719–1751. 2005.
- [21] Mariano, D. AlphaFold e a busca pelo Santo Graal da Biologia Molecular. In: BIOINFO #02 - Revista Brasileira de Bioinformática e Biologia Computacional. Vol. 2. Ed. 2. Disponível em:  
<https://bioinfo.com.br/alphafold-e-a-busca-pelo-santo-graal-da-biologia-molecular/>. Alfahelix (2022). doi: 10.51780/978-65-992753-5-7-10

# 25 AVALIAÇÃO ADMET DE SUBSTÂNCIAS

Autores 25.1

Artur Gomes Barros 

Revisão: Bruna Espiño dos Santos , Diego Mariano 

Cite este artigo 25.1

Barros, AG. **Avaliação ADMET de substâncias.** BIOINFO. ISSN: 2764-8273. Vol. 3. p.25  
(2023). doi:10.51780/bioinfo-03-25

### Resumo 25.1

Avaliar se moléculas sintéticas são perigosas ou não aos seres vivos é uma preocupação presente há muito tempo na comunidade científica. Entretanto, esses testes dependiam de avaliações *in vitro*, o que aumentava consideravelmente os custos para realização de experimentos. Recentemente, novas estratégias computacionais, como as propostas pelas ferramentas SwissADMET e pkCSM, permitiram análises complementares *in silico*, trazendo assim uma redução de custos. Desse modo, a avaliação das moléculas através de fatores essenciais que determinam se um composto pode ser nocivo aos seres vivos tem se tornado cada vez mais popular devido a adoção das avaliações ADMET (sigla para Absorção, Distribuição, Metabolismo, Excreção e Toxicidade) por meio de técnicas computacionais antes da realização de testes *in vitro*. Neste artigo, você verá uma breve introdução de como essas técnicas podem ser utilizadas por meio de duas ferramentas: SwissADMET e pkCSM.

## 25.1 Introdução

**O**s testes **ADMET (Absorção, Distribuição, Metabolismo, Excreção e Toxicidade)** são formas utilizadas para avaliar se uma substância, por mais que ela seja útil para um determinado fator, será tóxica para seres humanos ou outras espécies. Esses testes se baseiam, principalmente, em questões experimentais de compostos com alta similaridade com o qual se deseja analisar ou até mesmo sobre testes já feitos com determinada substância *in vitro* ou os testes se baseiam exclusivamente em questões de química orgânica da molécula a qual deseja obter informações farmacológicas.

Avaliações de moléculas por meio do método ADMET antigamente eram feitos apenas de forma *in vitro*, o que gerava um maior custo às empresas e aos laboratórios que trabalhavam na descoberta de potenciais substâncias que muitas vezes nem eram úteis para determinada pesquisa [4]. Entretanto, assim como

existem substâncias potencialmente tóxicas, também existem substâncias que podem ser de extrema importância. Desse modo, foram criados métodos para possibilitar a avaliação dessas substâncias, reduzindo a quantidade de testes *in vitro* realizados como sendo a primeira opção para verificar se uma substância é ou não potencial para determinada pesquisa. Assim, foram propostos métodos para reduzir a quantidade de testes em seres vivos, através das análises computacionais que recebem o nome de testes de ADMET *in silico* [3].

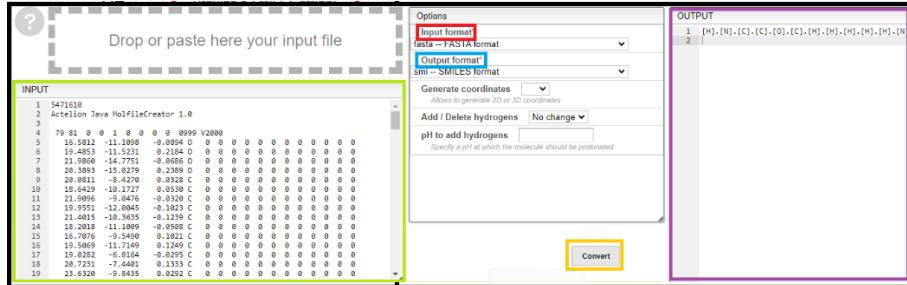
Esses testes não permitem, necessariamente, a descoberta de novas substâncias. Entretanto, é possível avaliar substâncias descobertas de outras maneiras e avaliar se essas têm ou não potencial de eficácia nos seres vivos em geral. Através do conhecimento em bioinformática de softwares de análises ADMET é possível avaliar desde a provável inibição de enzimas importantes para o organismo até se uma determinada substância tem um potencial carcinogênico. Por exemplo, pode-se citar o caso da Guttiferona-A, que de acordo com testes ADMET *in silico* pode ser uma ótima substância contra *Candida albicans* [1], porém foi descoberta a partir de análise não *in silico*.

## 25.2 Como são realizados os testes ADMET *in silico*

Primeiramente, todos os softwares apresentados usam o formato SMILE como forma de entrada para iniciar a análise. Assim, é necessário compreender como converter quase todo tipo de formato para o formato SMILE.

Vamos usar um software chamado Openbabel, que pode ser encontrado em: <https://www.cheminfo.org/Chemistry/Ceminformatics/FormatConverter/index.html> [6]. Nessa ferramenta, clique em “*input format*” para selecionar o formato de entrada. A seguir, insira em “*output format*” o formato de saída desejado (queremos converter nossa entrada para o formato SMILES).

Insira seu arquivo de entrada no campo esquerdo. A seguir, clique em “*convert*”. O programa irá converter seu texto de entrada em uma saída no formato SMILES, que aparecerá na lado direito da tela (conforme pode ser visto na Figura 25.1).



*Figura 25.1: Interface Openbabel. Para melhor exemplificação foi inserido um exemplo de uma molécula em formato *fasta* (lado esquerdo) o qual foi convertido para o formato *SMILES* (lado direito).*

Finalizado isso, é só copiar o texto em formato SMILES e partimos para o processo de se utilizar o software para análise ADMET. A seguir, será descrito apenas dois dos principais softwares utilizados. Vamos começar com o SwissADMET.

## 25.3 Software para avaliações ADMET

### 25.3.1 SwissADME

Segundo seu site oficial, SwissADME é uma ferramenta web que permite “calcular descritores físico-químicos, bem como prever parâmetros ADME, propriedades farmacocinéticas, natureza medicamentosa e compatibilidade química medicinal de uma ou múltiplas moléculas pequenas para apoiar a descoberta de medicamentos”. A ferramenta SwissADME pode ser encontrada em: <http://www.swissadme.ch> [2].

Assim, a página principal da ferramenta possui uma interface interativa que permite desenhar moléculas. No lado direito da interface, há um campo para inserção de texto do formato SMILES da molécula. Após inserir o texto correspondente ao código SMILES da molécula desejada, basta pressionar o botão “Run!” para executar a ferramenta.

Os principais parâmetros usados nesta ferramenta são:

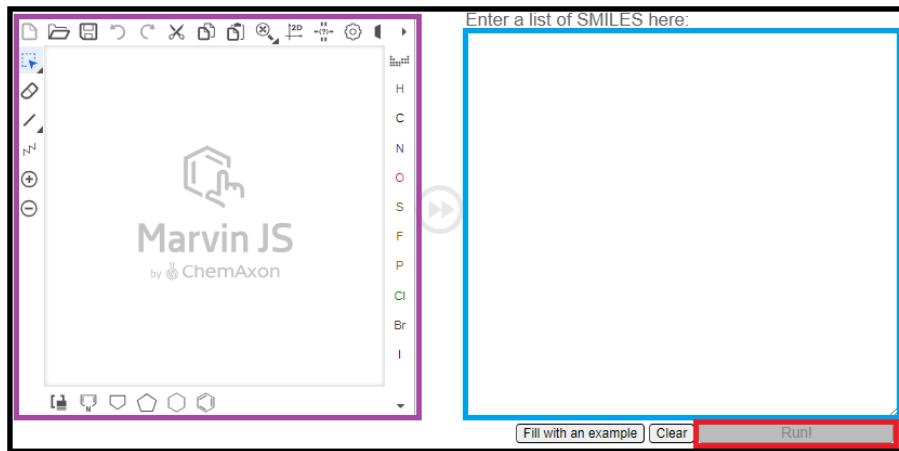


Figura 25.2: Interface SwissADME. À esquerda pode-se visualizar o local onde pode-se desenhar a molécula. À direita, pode-se visualizar o local onde deve ser inserido o formato SMILES da molécula. No lado inferior direito, podemos encontrar o botão “Run!”, que deve ser clicado para fazer uma análise geral da molécula.

- **Solubilidade em água:** há três possíveis opções para esse parâmetro: solúvel, pouco solúvel ou insolúvel. Com isso, um fármaco que é solúvel em água, terá uma maior biodisponibilidade corporéa, tanto para alcançar os mais diversos locais como também para uma excreção mais facilitada [3].
- **Regras de Lipinski (regra dos 5):** aqui, quanto menor a quantidade de violações melhor será esse parâmetro, mas normalmente acima de duas violações já começam a descartar as moléculas julgando-as ser ineficientes ou inadequadas [1]. São parâmetros avaliados nessa regra:
  - **Peso molecular:** o peso molecular do composto não deve exceder 500 Da.
  - **Aceptores de hidrogênio:** o número de átomos aceptores de ligação de hidrogênio (como grupos -OH e -NH) não deve exceder 10.
  - **Doadores de hidrogênio:** o número de átomos de hidrogênio doadores de ligação (como grupos -OH e -NH) não deve exceder 5.

- **logP <= 5:** O logP do composto deve ser menor que 5. Esta medida avalia a lipofilicidade, ou seja, a afinidade do composto por gordura em relação à água.

- **Farmacocinética**

- **Absorção GI (absorção gastrointestinal):** se trata de um parâmetro para avaliar se a molécula analisada tem alta ou baixa absorção gastrointestinal (sistema digestivo). Assim, esse resultado vai de acordo com a pesquisa a qual está sendo avaliada, se existe ou não a pretensão que tal substância seja absorvida pelo trato intestinal.
- **BBB permeabilidade (permeabilidade na barreira hematoencefálica):** avalia se determinado medicamento tem a capacidade ou não (sim ou não) de passar pela barreira hematoencefálica entrando assim no sistema nervoso do ser vivo. Esse parâmetro é essencial para fármacos desenvolvidos para tratamento de doenças no encéfalo. Porém, caso não seja de interesse que o fármaco atinja o cérebro então deve-se esperar um resultado negativo para esse parâmetro.
- **Inibição de proteínas:** aqui são avaliadas possíveis interações com proteínas as quais estão envolvidas diretamente no metabolismo de fármacos (podendo essas gerar efeitos colaterais). Sendo elas:
  - \* CYP1A2 é uma monooxigenase do citocromo P450 envolvida no metabolismo de vários substratos endógenos, incluindo ácidos graxos, hormônios esteróides e vitaminas
  - \* CYP2C19 é uma monooxigenase do citocromo P450 envolvida no metabolismo de ácidos graxos poliinsaturados
  - \* CYP2C9 é uma monooxigenase do citocromo P450 envolvida no metabolismo de vários substratos endógenos, incluindo ácidos graxos e esteróides

- \* CYP2D6 é uma monooxigenase do citocromo P450 envolvida no metabolismo de ácidos graxos, esteróides e retinóides
- \* CYP3A4 é uma monooxigenase do citocromo P450 envolvida no metabolismo de esteróis, hormônios esteróides, retinóides e ácidos graxos.

É possível observar que todas essas enzimas têm funções semelhantes, mas cada uma desempenha a formação de ácidos graxos e catálise de hormônios diferentes os quais são essenciais para o armazenamento de energia, sendo úteis para formação de estruturas celulares. Sua inibição pode causar problemas nessas vias,

- **Log K<sub>p</sub>**: nesse parâmetro é avaliado o coeficiente de permeabilidade na pele humana, assim através dele podemos verificar de tal composto caso seja eficiente no contato com a pele humana se ele irá ultrapassar a barreira da pele ou se ficará apenas superficialmente, ou caso estiver trabalhando com um composto tóxico se determinado composto caso entre em contato com a pele exista chances de atingir as vias internas do corpo.

### 25.3.2 pkCSM

Assim, tendo avaliado esses parâmetros nesse software, agora analisaremos fatores mais relacionados à toxicidade da molécula com o pkCSM. Segundo sua documentação, pkCSM é uma plataforma web que utiliza técnicas de aprendizado de máquina para prever propriedades farmacocinéticas de moléculas pequenas, se baseando em padrões de distância/farmacóforo codificados como assinaturas baseadas em grafos. A ferramenta foi treinada com diversos conjuntos de dados experimentais de descritores ADMET. pkCSM pode ser acessado em <https://biosig.lab.uq.edu.au/pkcsmprediction> [4].

Desse modo, o uso do pkCSM requer um processo muito parecido com o feito para o SwissADMET, uma vez que a entrada também requer o formato SMILES da molécula. A figura a seguir apresenta a interface do pkCSM:

**Upload your SMILES file:**

Nenhum ficheiro selecionado

Files are expected to have headers identifying the columns [File limits](#)

**OR**

**Provide a SMILES string:**

Example:  
CC(=O)OC1=CC=CC=C1C(=O)O

**Step 2: Please choose the prediction mode**

[Description](#)

**Prediction of pharmacokinetic properties**

Figura 25.3: Interface pkCSM. Em vermelho, pode-se ver o local onde deve ser inserido o formato SMILES da molécula.

Após inserir a molécula no formato SMILES no local indicado, basta clicar em “ADMET” localizado no fim da página (Figura 25.3).

A seguir, serão apresentados alguns os parâmetros que podem ser visualizados no resultado do pkCSM:

- **AMES toxicity**: esse parâmetro trata da toxicidade da molécula a qual tem ou não a capacidade mutagênica (gerar mutações genéticas) no ser vivo. Pode ser “sim ou não”.
- **Máx. dose tolerada (humano)**: revela a quantidade máxima em log mg/Kg/dia para um ser humano de modo que essa dose não cause problemas (valor dado em logaritmo).
- **Toxicidade Aguda e Crônica Oral em Rato (LD50)**: parâmetro referente à quantidade numérica de quanto mol/Kg seria a dose a qual mataria 50% da população de ratos podendo assim ser considerada ou não tóxica com base no experimento realizado.
- **Hepatotoxicidade**: categoria referente aos problemas que tal molécula tem potencial ou não causar problemas hepáticos (fígado) nos seres vivos.
- **Sensibilização da pele**: parâmetro similar ao Log Kp (apresentado na descrição do SwissADME) que tem avaliações referentes à pele. Todavia, esse parâmetro avalia a questão se tal fármaco gera ou não sensibilidade na pele.
- **Toxicidade de *T. pyriformis***: parâmetro parecido com o da dose tolerada para humano. Entretanto, refere-se ao protozoário *T. pyriformis* (o uso desse parâmetro depende da linha de pesquisa).
- **Toxicidade em peixes**: avalia a toxicidade em peixes.

## 25.4 Conclusão

Neste artigo, foram apresentados dois softwares que realizam análises ADMET *in silico*. Essas análises permitem verificar se determinada molécula tem potenciais

usos ou não aos seres vivos, mas também se essas substâncias podem causar problemas aos indivíduos.

Testes ADMET *in silico* têm um menor custo e, quando combinados com testes *in vitro*, podem ter maior eficiência. Ademais, por conta da eficiência dos testes *in silico* recomenda-se aos grupos de pesquisa possuírem em sua equipe bioinformaticas com esse conhecimento para auxiliar nas pesquisas que envolvam potenciais substâncias. Tanto *in silico* como *in vitro* são necessários, já que um complementa o outro. Todavia, cabe ressaltar que, como esses testes são feitos de forma preditiva, devem vir embasados de provas *in vivo*.

Saiba mais 25.1

Este artigo está disponível em <https://bioinfo.com.br/avaliacao-admet-de-substancias/>

## 25.5 Referências

[1] BARROS, A. G.; DE ARAÚJO, L. P.; FREITAS, N. J. Natural Resources for Human Health. 2023.

[2] DAINA, A.; MICHELIN, O.; ZOETE, V. SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. *Scientific reports*, v. 7, n. 1, p. 42717, 2017. ISSN 2045-2322.

[3] DÍAZ-CERVANTES, E.; ROBLES, J.; AGUILERA-GRANJA, F. Understanding the structure, electronic properties, solubility in water, and protein interactions of three novel nano-devices against ovarian cancer: a computational study. *Journal of Nanoparticle Research*, v. 20, p. 1-11, 2018. ISSN 1388-0764.

[4] DOS SANTOS FREIRE, C. M. A. et al. Proposta pedagógica em prática no ensino de bioquímica: Aproveitamento de softwares livres como facilitador do processo de ensino e de aprendizagem. *Revista Thema*, v. 15, n. 4, p. 1442-1455, 2018. ISSN 2177-2894.

[5] NISHA, C. M. et al. Molecular docking and *in silico* ADMET study reveals acylguanidine 7a as a potential inhibitor of  $\beta$ -secretase. *Advances in bioinformatics*, v. 2016, 2016. ISSN 1687-8027.

[6] PIRES, D. E. V.; BLUNDELL, T. L.; ASCHER, D. B. pkCSM: predicting small-molecule pharmacokinetic and toxicity properties using graph-based signatures. *Journal of medicinal chemistry*, v. 58, n. 9, p. 4066-4072, 2015. ISSN 0022-2623.

[7] YOSHIKAWA, N.; HUTCHISON, G. R. Fast, efficient fragment-based coordinate generation for Open Babel. *Journal of Cheminformatics*, v. 11, n. 1, p. 1-9, 2019. ISSN 1758-2946.

# 26 PERSPECTIVA HISTÓRICA DE GRANDES EVENTOS DA CIÊNCIA DA COMPUTAÇÃO E NA BIOLOGIA MOLECULAR: SEGUNDA GUERRA E GUERRA FRIA

Autores 26.1

Monique Cristina dos Santos , Aline de Paula Dias da Silva 

Revisão: Izabela M. C. A. Conceição 

Cite este artigo 26.1

Santos, MC; Silva, APD. Perspectiva histórica de grandes eventos da ciência da computação e na biologia molecular: segunda guerra e guerra fria. BIOINFO. ISSN: 2764-8273. Vol. 3. p.26 (2023). doi: 10.51780/bioinfo-03-26

### Resumo 26.1

**M**UITAS vezes, quando falamos sobre a história da genética ou da computação, negligenciamos o contexto histórico em que os principais eventos ocorreram, concentrando apenas nas evoluções científicas e no progresso que ocorreram até o momento presente.

Raramente lemos sobre a influência da conjuntura política nesses avanços. Essa abordagem pode nos levar a acreditar que as ciências naturais e matemáticas são objetivas, imparciais e desconectadas da política e economia.

Neste artigo, nosso objetivo é realizar uma revisão de alguns eventos importantes, com o propósito de refletir sobre como uma porção da ciência da computação e da biologia molecular interagiu com as demandas políticas e sociais do nosso planeta. Vamos explorar como esses avanços científicos foram moldados e influenciados pelo contexto político e social em que ocorreram. Dessa forma, poderemos entender melhor a interseção entre a ciência e as questões políticas do mundo.

## 26.1 Quando tudo ainda era mato!

Essa expressão é frequentemente usada na internet para se referir a momentos muito distantes do presente, quando os avanços tecnológicos eram escassos, ou seja, uma época histórica! Estudos indicam que o crescimento econômico passa por fases de expansão e declínio que podem durar décadas, esse ciclo de expansão e declínio é chamado de ondas longas. Na teoria das ondas longas ou Teoria Kondratieff, o momento da expansão é marcado pelo surgimento de inovações tecnológicas que levam ao crescimento industrial e o desenvolvimento de novos mercados que duram por décadas, no entanto, as forças do mercado levam à saturação industrial que culminam na contração. Dentre as forças do mercado, uma das mais conhecidas é a lei da oferta e demanda , que é basicamente o aumento do valor do produto quando a demanda do consumidor é alta frente à baixa oferta do mercado, entretanto, quando há um aumento na oferta e não há demanda do consumidor, os valores tendem a cair [1] . Nesse contexto, alguns

autores acreditam que a primeira recessão econômica do século passado começou no final da Segunda Guerra Mundial (1945) e durou até 1973, onde se iniciou a expansão que terminou em 1992. Esse período é marcado pela instabilidade da Segunda Guerra Mundial e da Guerra Fria que conflagrou um ambiente bélico a nível mundial interferindo diretamente na ciência e na tecnologia [2]. Por isso, vamos contar uma breve história que mescla disputas sócio-políticas, com os avanços na tecnologia e ciência molecular, como um elemento para se discutir a relação entre ciência e sociedade (Figura 1).

A máquina de Turing é considerada o conceito básico da tecnologia computacional moderna. Criada por Alan Mathison Turing entre os anos de 1936 e 1945, a máquina de Turing teve contribuições significativas para a história da computação. Em primeiro lugar, proporcionou um modelo matemático simples para uma máquina de computação universal. Em segundo lugar, participou da criação dos primeiros computadores digitais programáveis. E, por último, sua definição filosófica operacional influenciou o campo da inteligência artificial, uma vez que os programas são formas de dados que podem ser manipulados por outros programas [3].

Em 1936, o cenário sociopolítico europeu já estava marcado por avanços nazistas, como a invasão italiana da Etiópia, conhecida como a invasão da Abissínia, que ocorreu em 3 de outubro de 1935. Alan Turing, após sua primeira publicação, partiu para Princeton para realizar pesquisas mais avançadas em lógica, e retornou em 1938, apenas um ano antes do início da Segunda Guerra Mundial. Com o início do conflito global em 1939, ele integrou o corpo de Inteligência como líder da sessão Hut-8, responsável por decifrar mensagens dos navios alemães, contribuindo diretamente para o esforço de guerra dos Aliados (o grupo de países que combateu o nazismo). Seu trabalho nas máquinas de cifragem criptográficas persistiu até 1945, tornando-o uma figura científica proeminente e com exposição às tecnologias mais avançadas da época [4]. É relevante ressaltar que nesse período histórico muitos avanços tecnológicos da indústria, das Universidades e do Governo federal foram impulsionados por financiamentos militares, baseados no plano de sistema de inovação nacional [5].

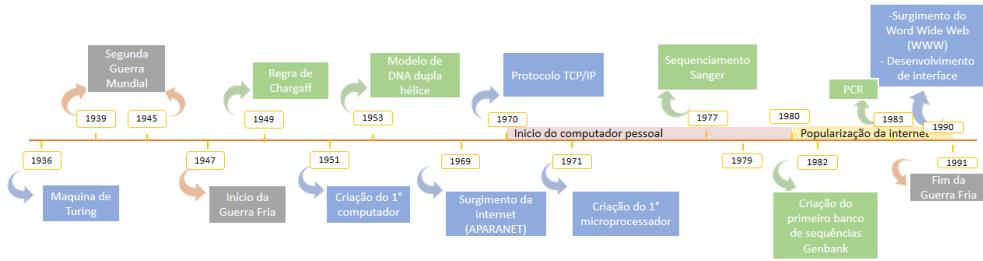


Figura 26.1: Linha do tempo comparativa entre avanços na bioinformática e o contexto político dos eventos do último século. Marcados em cinza são os eventos políticos, em azul os surgimento dos avanços nas pesquisas computacionais e em verde os principais eventos na história da biologia molecular. Fonte: Próprio autor

## 26.2 Chegou a Guerra Fria...

Os anos seguintes à Segunda Guerra foram marcados por tensões entre os Estados Unidos da América e a União Soviética, em um período conhecido como Guerra Fria, que se estendeu de 1947 a 1991. Esse evento sócio-político é descrito por especialistas como uma época de grandes avanços tecnológicos e científicos, uma vez que o conflito entre comunismo e capitalismo estimulou investimentos em ciência para fortalecer o poder bélico, resultando em ameaças nucleares significativas [6].

Na linha do tempo dos eventos de descobertas nas ciências biológicas, destacamos os avanços no entendimento do DNA. A Regra de Chargaff, descoberta pelo bioquímico austríaco emigrado para os Estados Unidos, Erwin Chargaff, em 1949, revelou a composição e a ligação das bases nitrogenadas. As bases nitrogenadas Timina se ligam à Adenina, assim como Citosina se liga à Guanina, por meio de pontes de hidrogênio. Notavelmente, as quantidades de pares de bases são equivalentes entre si, em condições estáveis [7]. A segunda descoberta, publicada na Nature, refere-se à pesquisa inglesa sobre a estrutura tridimensional do DNA, descrevendo a dupla fita da molécula com base em cálculos a partir das imagens de raios-x obtidas por Rosalind Franklin [8, 9]. Watson e Crick receberam o Prêmio Nobel em Medicina por sua descoberta do modelo de dupla hélice mais estável em solução da molécula de DNA. Durante os anos 50, esses cientistas continuaram a fazer importantes descobertas moleculares. Francis

Crick, descreveu o RNA transportador (RNAt), uma molécula responsável por transportar informações contidas no RNA mensageiro para serem traduzidas, e em 1958 escreveu o que hoje conhecemos como Dogma Central da Biologia Molecular [10]. Alguns pesquisadores desvinculam as descobertas de Watson e Crick do contexto bélico da Guerra Fria, afirmando que essas descobertas científicas seriam possíveis em uma Inglaterra pacífica [11]. No entanto, é fundamental reconhecer o papel da Inglaterra na Guerra Fria, sua aliança com os Estados Unidos, sua presença como um dos países fundadores da OTAN e a primeira nação, após os Estados Unidos e a União Soviética, a produzir armas nucleares independentes [12]. Assim, não estamos relacionando levianamente Watson e Crick ao cenário da Guerra Fria, como se estivessem trabalhando para a inteligência inglesa. Nossa abordagem é mostrar que a Inglaterra estava inserida no contexto da Guerra Fria e que, naquele momento, as descobertas científicas estavam em disputa entre os dois blocos polarizados. Nesse primeiro momento, a pesquisa científica molecular tinha um foco maior na compreensão da estrutura do DNA.

No final da década de 1960, surgiu a ARPANET (Advanced Research Projects Agency Network), considerada a precursora da internet, como a primeira rede descentralizada de computadores. Financiada pelo Departamento de Defesa dos Estados Unidos, seu propósito era o de fomentar a pesquisa básica e garantir a troca de informações entre instituições, mesmo em casos de guerra ou desastres naturais [13]. A ARPANET foi uma resposta ao lançamento do Sputnik pela União Soviética [14]. Atualmente, o TCP/IP (Protocolo de Controle de Transmissão/Protocolo de Internet) é a tecnologia base da internet, que comprehende um conjunto de regras para a comunicação na rede. A discussão sobre a relação da criação da internet com a Guerra Fria abrange diferentes perspectivas, desde sua origem até sua disseminação e estrutura, que alguns acreditam ter sido possível graças à hegemonia norte-americana. Além disso, há quem considere a maior contribuição da internet para a Guerra Fria relacionada aos dados que puderam ser obtidos através dela [15]. Entretanto, nosso propósito aqui é demonstrar como a ciência e a política se entrelaçam, embora nem sempre essa relação seja tão clara ou direta quanto se espera. A condução de um país pode influenciar direta ou indiretamente as linhas de pesquisa do momento, tornando-as mais possíveis de acontecer do que em qualquer outro período histórico.

Com base nessa linha de raciocínio, avançaremos um pouco na cronologia e chegamos ao sequenciamento Sanger. Nesse ponto, a tecnologia já havia avançado, com redução do tamanho dos processadores e o início da disseminação do uso do computador pessoal, em 1977. Vale ressaltar que a técnica de sequenciamento Sanger é anterior à descoberta do PCR. Em 1972, foram realizadas técnicas de extração, amplificação e sequenciamento em um vírus bacteriófago X174, que infecta a bactéria *Escherichia coli* [16]. Seu genoma é composto por aproximadamente 5.386 nucleotídeos. Naquela época, a extração era realizada através da infecção da colônia bacteriana para aumentar a quantidade viral, seguida da extração do DNA por meio de lise celular e purificação. A amplificação era feita através de vetores clonais de bactérias, e o sequenciamento no tipo Sanger era a técnica utilizada [17]. Essa metodologia permitiu avanços significativos no campo da genética e abriu caminho para pesquisas futuras no sequenciamento de DNA. É importante mencionar que a primeira sequência completa de 200 pares de bases (pb) de DNA também foi obtida pelo bioquímico britânico Frederick Sanger. Em 1972, ele já havia desenvolvido um protótipo do sequenciamento Sanger, que envolvia cadeias complementares de diferentes tamanhos ligadas ao fragmento de DNA alvo. A sequência era obtida através da análise das cores dos ddNTPs marcados com uma fluorescência, após a separação das cadeias resultantes por tamanho e radioatividade através de um gel de poliacrilamida [17]. Por suas contribuições, Sanger foi agraciado com o Prêmio Nobel de Química novamente em 1980, após ter recebido o mesmo prêmio em 1958 por desvendar a estrutura da proteína Insulina.

Chegamos a 1979, e o cenário científico desse momento era marcado pela descoberta de muitas sequências e proteínas, todas elas disponíveis de forma particular. Diante disso, 30 biólogos computacionais da Universidade Rockefeller em Nova York decidiram que era o momento ideal para criar um banco de sequências, com o objetivo de agrupar e catalogar todas as informações obtidas [18]. Porém, com os Estados Unidos envolvidos na Guerra Fria e enfrentando uma grande recessão, os pesquisadores buscaram apoio junto ao National Institutes of Health (NIH). No entanto, a resposta do instituto demorou, e foi então que o European Molecular Biology Laboratory (EMBL), na Europa, lançou o primeiro banco de sequências em 1980. Três anos após a ideia inicial, surgiram duas

candidaturas para financiar esse banco de dados. A primeira candidatura foi apresentada por Dayhoff, que possuía um pioneiro banco de dados de sequências proteicas, colecionadas desde o início dos anos 1960. Entretanto, tal qual os Museus de História Natural, Dayhoff recebia dos pesquisadores diretamente antes de serem publicizados. Esse modelo de projeto não obteve tanto sucesso, pois a publicação dos dados em seu atlas não garantia a autoria ou a prioridade [19].

A segunda candidatura foi proposta por Los Alamos, que tinha um projeto muito parecido com o de Dayhoff, mas com processos de autoria e prioridade melhor definidos. Walter Goad, defensor de Los Alamos, percebeu que o laboratório era um bom lugar para a criação do banco de dados, dada sua capacidade computacional. Ele também se reuniu com alguns laboratórios europeus que contribuíram com informações para o banco de sequências. Um fator político também influenciou na escolha do laboratório. A Suprema Corte dos EUA havia declarado que “qualquer coisa feita pelo homem, incluindo organismos geneticamente modificados, poderia ser patenteada [19]”. Goad sustentava que os dados deviam ser públicos e disponíveis na ARPANET, a precursora da internet. Enquanto Dayhoff só poderia disponibilizar os dados por meio de modems telefônicos, com acesso online limitado. Goad venceu a disputa, e o Los Alamos se tornou o GenBank. O sucesso do GenBank foi atribuído à sua aliança com o EMBL e posteriormente com o banco japonês DNA Databank of Japan (DDBJ), criando a União Internacional de Bancos de Dados de Sequências (INSDC) [20].

Finalmente, chegamos ao último evento antes de falarmos sobre o DNA barcoding: o surgimento do PCR em 1983. Kary B. Mullis conta que teve a ideia da Polymerase Chain Reaction (PCR) enquanto dirigia um carro à noite, na região das sequóias do norte da Califórnia [21]. É interessante notar que os anos do surgimento do GenBank são próximos ao surgimento do PCR, e, consequentemente, houve um aumento significativo no número de sequências armazenadas no GenBank, facilitado pelo acesso a novas ferramentas de sequenciamento.

Podemos afirmar que no pós-Guerra Fria, a genética humana obteve muitos incentivos, impulsionados principalmente por três pontos principais: (I) projetos

relacionados à raça e identidade nacional, pois as características hereditárias das populações eram vistas como relevantes para a saúde da nação, sua história e futuro; (II) as ameaças atômicas incentivaram projetos sobre hereditariedade, uma vez que se acreditava que a radiação afetava diretamente a herança genética; (III) a busca pela reconstrução do Genoma humano, permitindo conhecer marcadores importantes no código genético, o que poderia ser muito útil para fins comerciais e econômicos [22].

### **26.3 Nova Era...**

Após o término da Guerra Fria, a preocupação com o futuro era latente, uma vez que um novo século estava prestes a começar e a degradação ambiental se mostrava um desafio crescente. Essa necessidade de conhecer melhor a biodiversidade ficou evidente com a realização, um ano após o fim da Guerra Fria, da Conferência das Nações Unidas sobre o Meio Ambiente e o Desenvolvimento (ECO-92). Seu principal objetivo era debater o cenário ambiental global e propor medidas para a proteção do meio ambiente.

Nesse contexto, a metodologia de análise genética estava em ampla utilização, com pesquisas envolvidas recebendo considerável apoio financeiro. Porém, havia uma demanda significativa para conhecer melhor o meio ambiente, e a escassez de taxonomistas, especialistas que demandam muitos anos de formação, apresentava-se como um desafio adicional. Surgia, portanto, um novo conflito: como cuidar da biodiversidade se ainda a conhecemos de forma limitada? A resposta veio com o avanço da bioinformática e o aprimoramento tecnológico no sequenciamento ao longo dos anos [23].

### **26.4 Conclusão**

Em vez de focarmos apenas na perspectiva da ciência nas últimas duas grandes guerras mundiais, também poderíamos discutir como a ciência precisou se adaptar à pandemia da COVID-19 diante das decisões políticas. Além disso, há ainda a relevância da oceanografia, química e física no contexto da Guerra Fria. Engana-se quem pensa que nossas pesquisas, por serem aparentemente

imparciais, não impactam na sociedade ou não são impactadas pelos eventos sociais da época. Sempre haverá influência da sociedade, seja nas perguntas que fazemos ou nas respostas que buscamos e suas possibilidades de aplicação. A ciência e a sociedade estão intrinsecamente entrelaçadas, nossas pesquisas causam impactos sociais, tanto quanto a sociedade impacta nossas pesquisas e juntas elas seguem moldando o curso da história e delineando nosso futuro.

#### Saiba mais 26.1

Este artigo está disponível em <https://bioinfo.com.br/perspectiva-historica-de-grandess-eventos-da-ciencia-da-computacao-e-na-biologia-molecular-segunda-guerra-e-guerra-fria/>

## 26.5 Referências

- [1] Kondratieff, Nikolai D. "The long waves in economic life." Review (Fernand Braudel Center) (1979): 519-562.
- [2] Coccia, Mario. "A theory of the general causes of long waves: War, general purpose technologies, and economic change." Technological Forecasting and Social Change"128 (2018): 287-295. [3] French, Robert M. "The Turing Test: the first 50 years." Trends in cognitive sciences 4.3 (2000): 115-122.
- [4] Hodges, Andrew. "Alan turing." (2002).
- [5] Mowery, David C., and Nathan Rosenberg. "The US national innovation system." National innovation systems: A comparative analysis (1993): 29-75.
- [6] Wang, Jessica. American science in an age of anxiety: Scientists, anticomunism, and the Cold War. Univ of North Carolina Press, 1999.
- [7] Manchester, Keith L. "Historical Opinion: Erwin Chargaff and his 'rules' for the base composition of DNA: why did he fail to see the possibility of complementarity?." Trends in biochemical sciences 33.2 (2008): 65-70.
- [8] Watson, James D., and Francis Crick. "A structure for deoxyribose nucleic acid." (1953): 737.
- [9] Danylova, T. V., and S. V. Komisarenko. "Standing on the shoulders of giants: James Watson, Francis Crick, Maurice Wilkins, Rosalind Franklin and the birth of molecular biology." Ukr Biochem J 92.4 (2020): 154-165.

- [10] Crick, Francis H. "On protein synthesis." *Symp Soc Exp Biol.* Vol. 12. No. 138-63. 1958.
- [11] Petsko, Gregory A. "War and peace." *Genome biology* 4.5 (2003): 1-2.
- [12] Pannier, Alice. "From one exceptionalism to another: France's strategic relations with the United States and the United Kingdom in the post-Cold War era." *Journal of strategic studies* 40.4 (2017): 475-504.
- [13] Hauben, Michael, and R. Hauben. "Behind the net: the untold history of the ARPANET and computer science." *Netizens: on the history and impact of Usenet and the internet* (2006).
- [14] Lammle, Todd. "Introduction to TCP/IP." (2020): 63-104.
- [15] Townes, Miles. "The spread of TCP/IP: How the Internet became the Internet." *Millennium* 41.1 (2012): 43-64.
- [16] Kulski, Jerzy K. "Next-generation sequencing—an overview of the history, tools, and "Omic" applications." *Next generation sequencing-advances, applications and challenges* 10 (2016): 61964.
- [17] Heather, J. M., Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics*, 107(1), 1-8.
- [18] Pauli, Jonathan N., Shawn A. Steffan, and Seth D. Newsome. "It is time for IsoBank." *BioScience* 65.3 (2015): 229-230.
- [19] Strasser, Bruno J. "GenBank—Natural History in the 21st Century?" *Science* 322.5901 (2008): 537-538.
- [19] V. Chakrabarty, Diamond. Intellectual Property Strategy in Bioinformatics and Biochips 85.
- [20] Strasser, B. J. (2008). GenBank—Natural History in the 21st Century?. *Science*, 322(5901), 537-538.
- [21] Mullis, Kary B. "The unusual origin of the polymerase chain reaction." *Scientific American* 262.4 (1990): 56-65.
- [23] Lindee, Susan. "Scaling up: human genetics as a Cold War network." *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 47 (2014): 185-190.