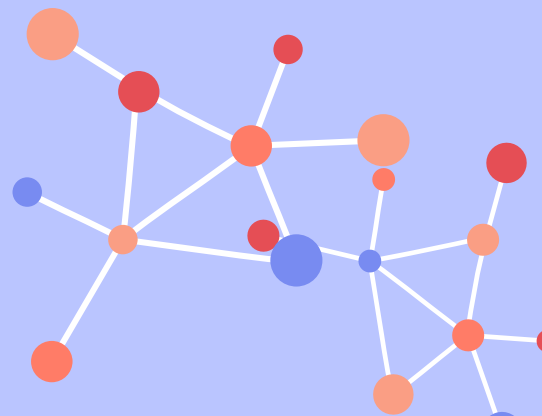




Mineração e Análise de Dados de Bioinformática Estrutural: Análise de Estruturas de Interações Proteína-RNA Modeladas por Métodos Computacionais

João Pedro Braga Ennes

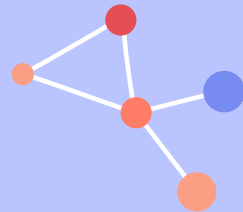




01

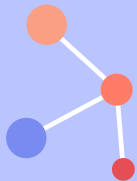
INTRODUÇÃO

BANCO DE DADOS **PROTÍNA-RNA**



No último semestre construímos um banco de dados de estruturas de interações proteína-rna, com informações do PDB, obtidas de forma experimental.

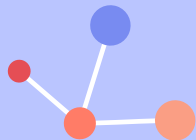
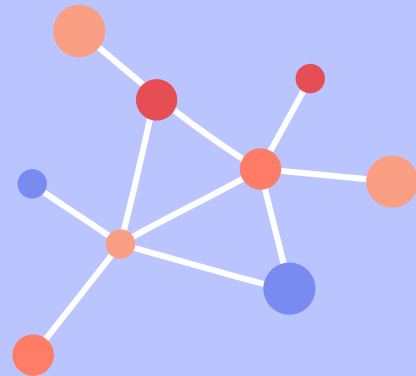
Inicialmente coletamos informações de 309 estruturas e após realizar algumas análises, removemos algumas que não se encaixavam no nosso grupo de interesse e obtivemos ficando no final com 170 estruturas.





02

ATUALIZANDO DADOS



MÉTODO DE COLETA

Advanced Search Query Builder

[Help](#)

Full Text ?

Structure Attributes ?

[Help](#)

AND	Polymer Entity Type	x	▼	is	▼	Protein	▼	+ NOT	Count	x
	Polymer Entity Type	x	▼	is	▼	RNA	▼	+ NOT	Count	x
	Number of Distinct Molecular Entities	x	▼	=	▼	2		+ NOT	Count	x
AND / OR		Add Attribute		Add Subquery		Remove Subquery				
Add Subquery										

Chemical Attributes ?

Sequence Similarity ?

Sequence Motif ?

Structure Similarity ?

Structure Motif ?

Chemical Similarity ?

Return Structures ▼ ? grouped by No Grouping ▼ ?

Include Computed Structure Models (CSM) ? ☐

325 Clear

 Search

Chemical Similarity ?

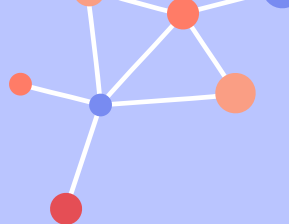
Return Structures ▼ ? grouped by No Grouping ▼ ?

Include Computed Structure Models (CSM) ? ☐

309 Clear

 Search

BANCO DE DADOS ESTRUTURAL

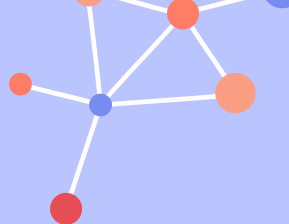


Após filtrar nossas estruturas de interesse, com apenas uma molécula de RNA e uma de Proteína, ficamos com um total de 204 estruturas



	Structure	Number of amino acids
0	1a4t	19
1	1aud	101
2	1biv	17
3	1d6k	94
4	1ekz	76
...
199	8pdl	351
200	8pdm	351
201	8pzp	470
202	8qgt	1289
203	8sxu	726
204 rows × 2 columns		

BANCO DE DADOS ESTRUTURAL



Após o processo de validação e controle de qualidade, ficamos com um total de **175** estruturas no nosso banco atualizado, cinco a mais do que na etapa anterior.

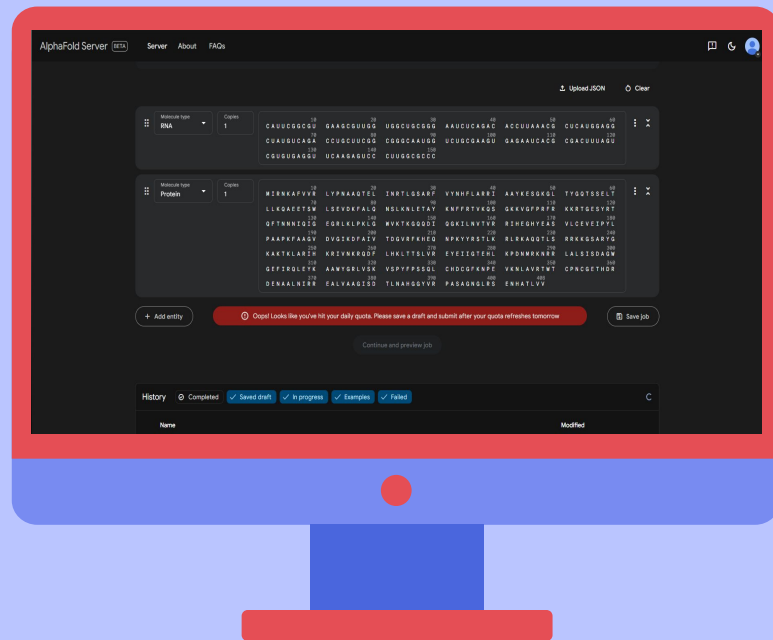


	Structure	Number of amino acids	Number of nucleotides
0	1a4t	19	15
1	1aud	101	30
2	1biv	17	28
3	1d6k	94	37
4	1ekz	76	30
...
170	8fti	737	97
171	8pdl	351	7
172	8pdm	351	7
173	8pzp	470	12
174	8sxu	726	10

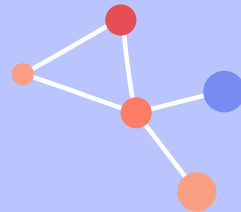
175 rows × 3 columns

03

MODELANDO ESTRUTURAS



ALPHAFOLD 3



Inicialmente utilizaríamos o RosettaFoldNA para realização deste estudo, porém no último dia 8 de Maio, foi lançado o **AlphaFold 3** com a possibilidade de modelar estruturas de DNA e RNA.

AI

AlphaFold 3 predicts the structure and interactions of all of life's molecules

Introducing AlphaFold 3, a new AI model developed by Google DeepMind and Isomorphic Labs. By accurately predicting the structure of proteins, DNA, RNA, ligands and more, and how they interact, we hope it will transform our understanding of the biological world and drug discovery.

May 08, 2024 · 6 min read

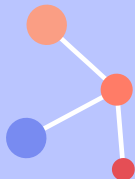


Google DeepMind
AlphaFold team



Isomorphic Labs

Share



APHAFOLD SERVER

AlphaFold Server BETA [Server](#) [About](#) [FAQs](#) 🗨️ 🌙 👤

Remaining jobs: 0

AlphaFold Server allows you to model a structure consisting of many biological molecules [Learn more](#) ▼

[📁 Upload JSON](#) [🔄 Clear](#)

⋮

Molecule type
RNA

▼

Copies
1

10 20 30 40 50 60

CAUUCGGCGU GAAGCGUUG UGGCUGCGGG AAUCUCAGAC ACCUUAACG CUCAUGGAGG

70 80 90 100 110 120

CUAUGUCAGA CCUGCUUCGG CGGGCAAUGG UCUGCGAAGU GAGAAUCACG CGACUUUAGU

130 140 150

CGUGUGAGGU UCAAGAGUCC CUUGGCGCCC

⋮ ⌵

⋮

Molecule type
Protein

▼

Copies
1

10 20 30 40 50 60

MIRNKAFFVR LYPNAAQTEL INRTLGSARF VYNHFLARRI AAYKESGKGL TYGQTSSSELT

70 80 90 100 110 120

LLKQAEETSW LSEVDKFALQ NSLKNLETAY KNFFRTVKQS GKKVGFPRFR KKRTGESYRT

130 140 150 160 170 180

QFTNNNIQIG EGRLKLPKLG WVKTKGQDDI QGKILNVTVR RIHEGHYEAS VLCEVEIPYL

190 200 210 220 230 240

PAAPKFAAGV DVGIKDFAIV TDGVRFKHEQ NPKYYRSTLK RLRKAQQTLS RRRKKGSARYG

250 260 270 280 290 300

KAKTKLARII KRIVNKRQDF LHKLTTSLVR EYEIIGTEHL KPDNMRKNRR LALSISDAGW

310 320 330 340 350 360

GEFIRQLEYK AAWYGRLVSK VSPYFPSSQL CHDCGFKNPE VKNLAVRTWT CPNCGETHDR

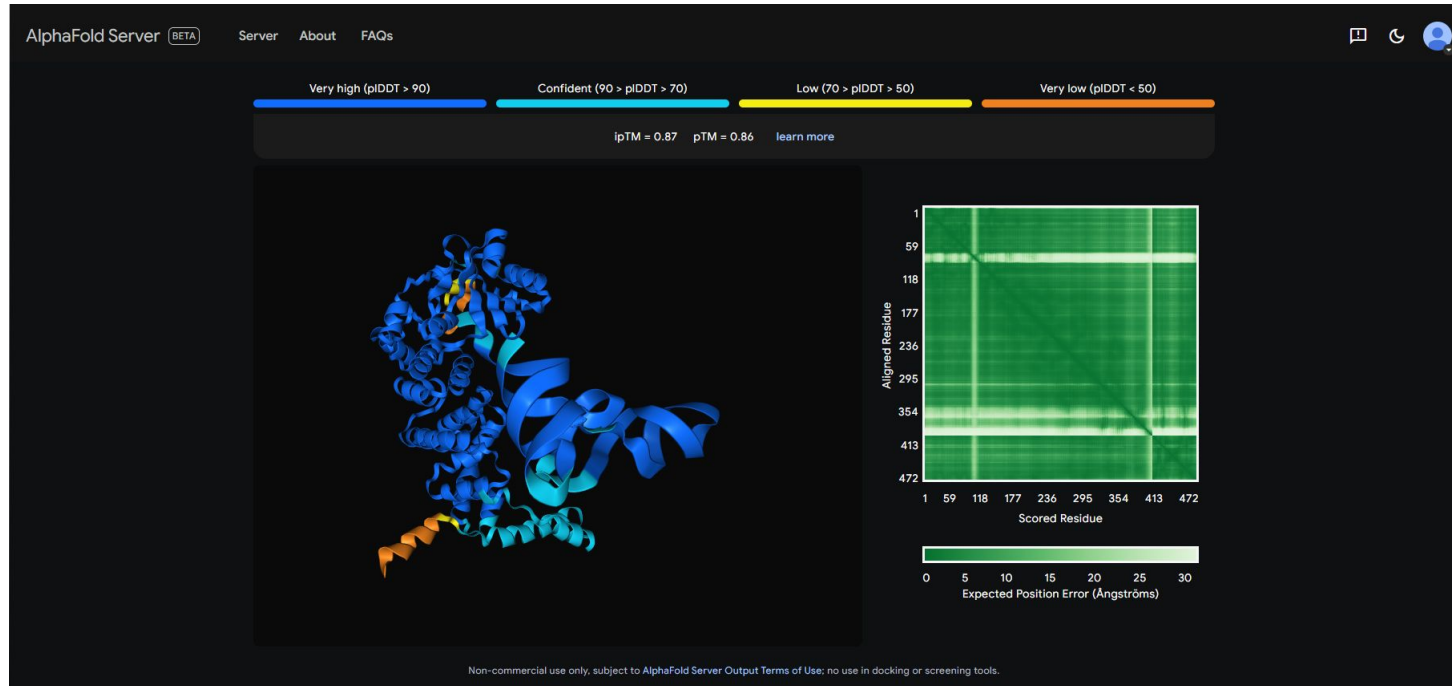
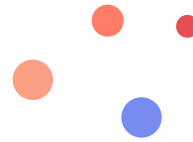
370 380 390 400

DENAALNIRR EALVAAGISD TLNAHGGYVR PASAGNGLRS ENHATLVV

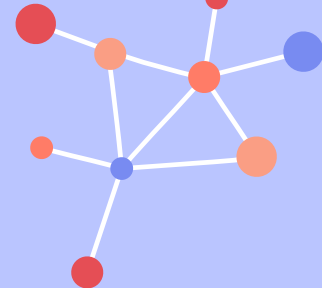
⋮ ⌵

[+ Add entity](#) ⓘ Oops! Looks like you've hit your daily quota. Please save a draft and submit after your quota refreshes tomorrow [📁 Save job](#)[Continue and preview job](#)

MODELAGEM DE ESTRUTURAS



TOTAL DE ESTRUTURAS MODELADAS



Com o AlphaFold Server, foi possível modelar **155** estruturas.

A queda de 20 estruturas se deu por 3 motivos principais:

- Máximo de tokens por job (1)
- Mínimo de tokens por RNA (15)
- Falha durante o processo de modelagem (4)



History











Completed

Saved draft

In progress

Examples

Failed

Name	Modified
 5MPL	2024-05-19 16:11
 5MPG	2024-05-19 16:11
 5FN1	2024-05-19 16:10
 5M8I	2024-05-19 16:10
 5J2W	2024-05-19 16:10
 5J1O	2024-05-19 16:09
 5BYM	2024-05-19 16:09
 5JOM	2024-05-19 16:08
 4UFT	2024-05-19 16:07
 4XOB	2024-05-19 16:07

Items per page: 10

1 – 10 of 159

<

>

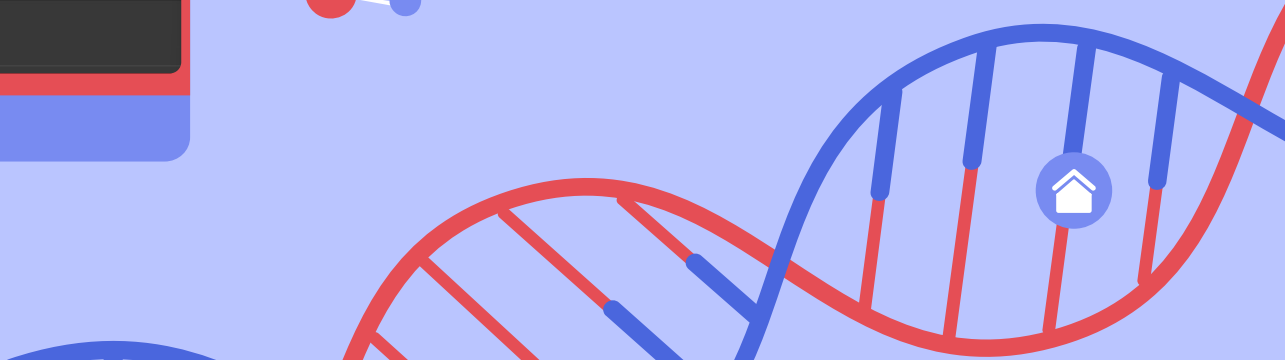
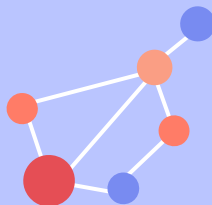
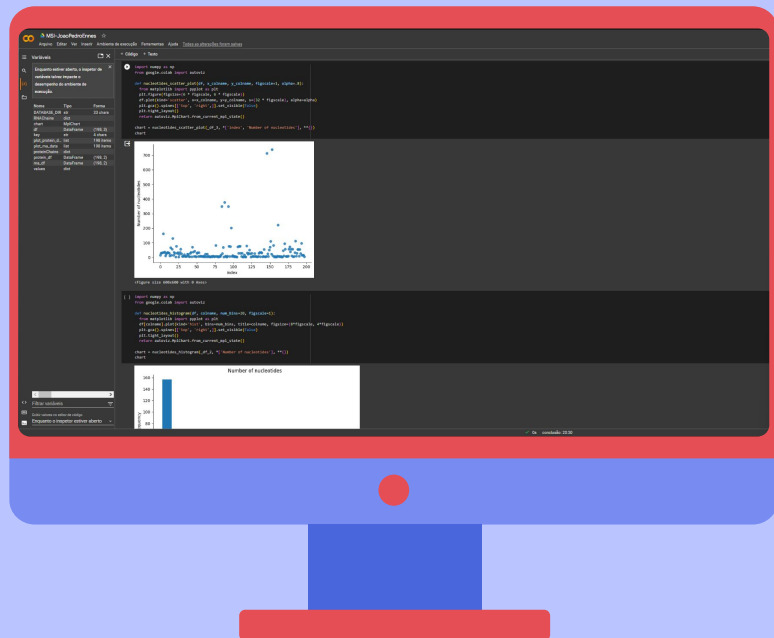
TRATAMENTO DOS DADOS

- Remoção de cabeçalho
- Remoção de resíduos gerados durante o processo de sequenciamento

3	ATOM	3	C4'	G	A	4	11.190	6.700	5.875	1.00	0.00	C
4	ATOM	4	O4'	G	A	4	12.365	5.959	5.435	1.00	0.00	O
5	ATOM	5	C3'	G	A	4	10.264	6.644	4.657	1.00	0.00	C
6	ATOM	6	O3'	G	A	4	9.300	7.705	4.620	1.00	0.00	O
7	ATOM	7	C2'	G	A	4	11.235	6.678	3.482	1.00	0.00	C
8	ATOM	8	O2'	G	A	4	11.451	7.949	2.868	1.00	0.00	O
9	ATOM	9	C1'	G	A	4	12.534	6.100	4.016	1.00	0.00	C
10	ATOM	10	N9	G	A	4	12.901	4.816	3.367	1.00	0.00	N
11	ATOM	11	C8	G	A	4	12.912	3.591	3.921	1.00	0.00	C
12	ATOM	12	N7	G	A	4	13.279	2.621	3.121	1.00	0.00	N
13	ATOM	13	C5	G	A	4	13.534	3.248	1.930	1.00	0.00	C
14	ATOM	14	C6	G	A	4	13.958	2.683	0.720	1.00	0.00	C
15	ATOM	15	O6	G	A	4	14.185	1.483	0.537	1.00	0.00	O
16	ATOM	16	N1	G	A	4	14.106	3.675	-0.294	1.00	0.00	N
17	ATOM	17	C2	G	A	4	13.870	5.055	-0.145	1.00	0.00	C
18	ATOM	18	N2	G	A	4	14.093	5.754	-1.293	1.00	0.00	N
19	ATOM	19	N3	G	A	4	13.460	5.557	1.052	1.00	0.00	N
20	ATOM	20	C4	G	A	4	13.311	4.622	2.034	1.00	0.00	C
21	ATOM	21	H5'	G	A	4	10.263	5.022	6.937	1.00	0.00	H
22	ATOM	22	H5''	G	A	4	11.408	6.001	7.924	1.00	0.00	H
23	ATOM	23	H4'	G	A	4	11.546	7.732	6.079	1.00	0.00	H
24	ATOM	24	H3'	G	A	4	9.785	5.642	4.630	1.00	0.00	H
25	ATOM	25	H2'	G	A	4	10.853	5.939	2.744	1.00	0.00	H
26	ATOM	26	HO2''	G	A	4	11.048	7.896	1.998	1.00	0.00	H

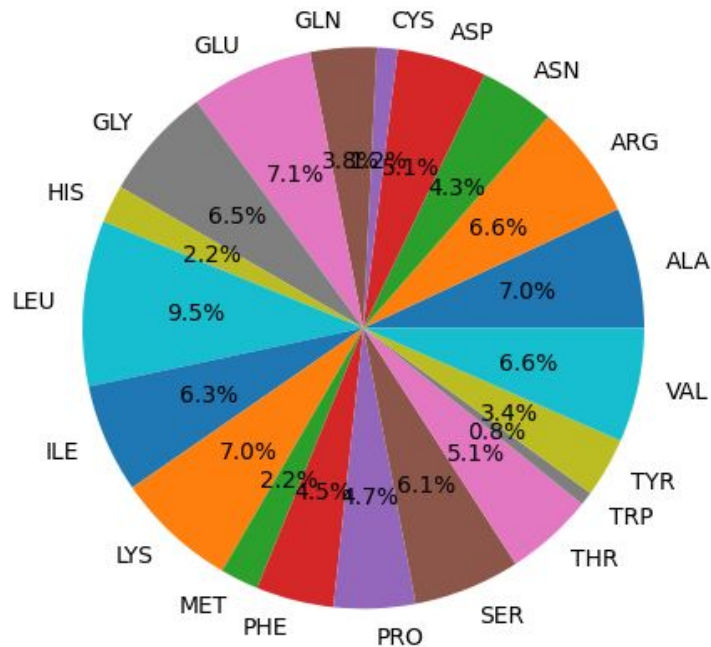
ANALISES

04

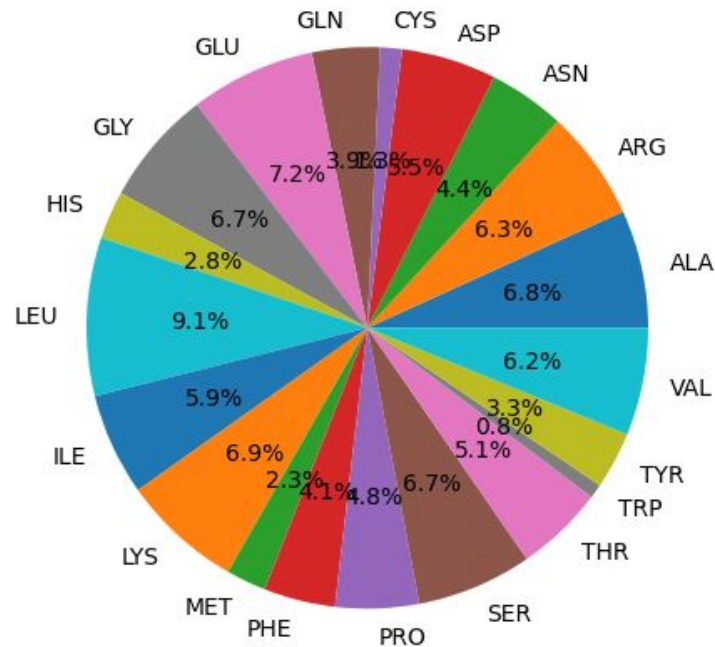


Frequência por Aminoácido

PDB

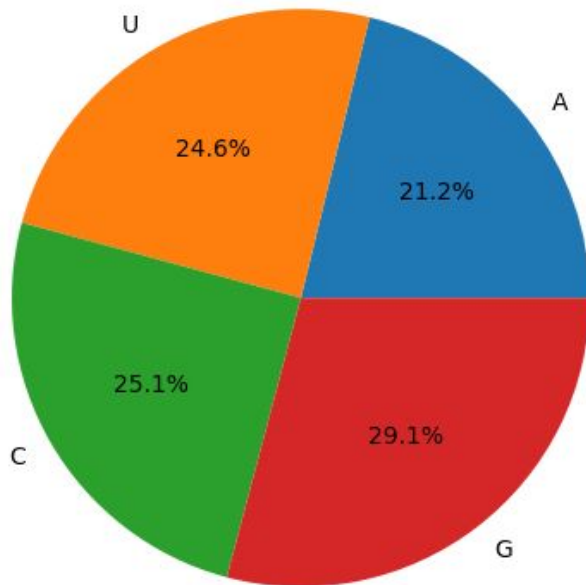


AlphaFold

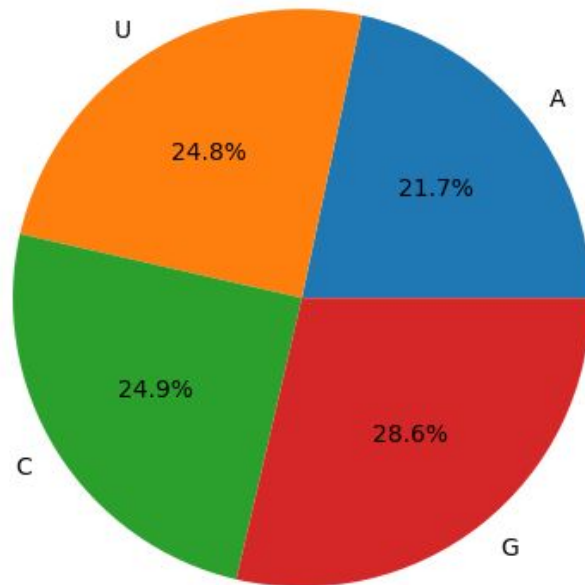


Frequência por Nucleotídeo

PDB

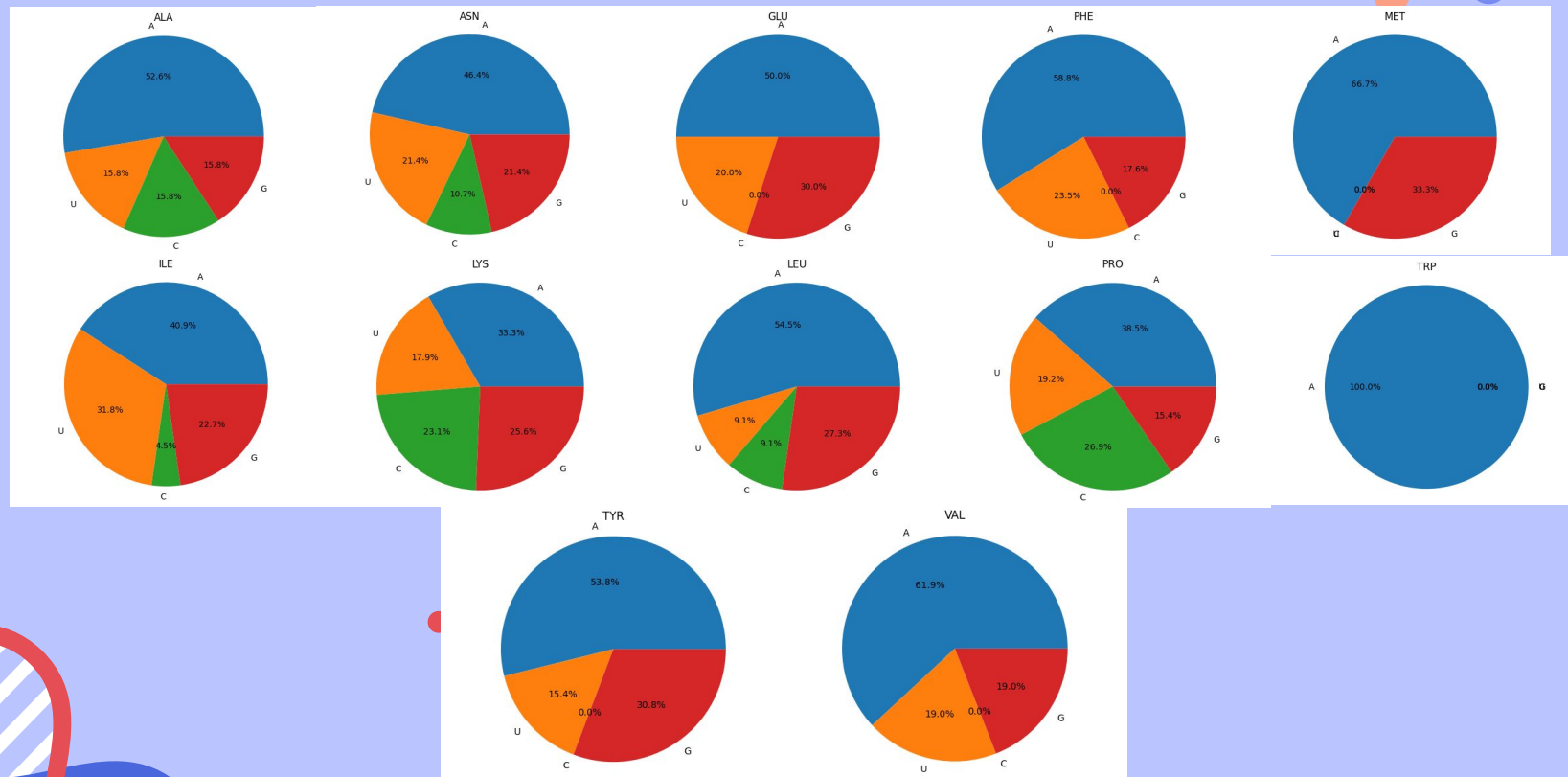
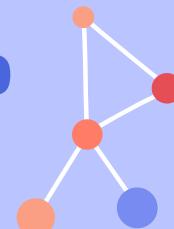


AlphaFold



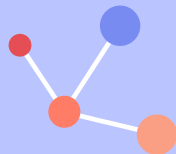
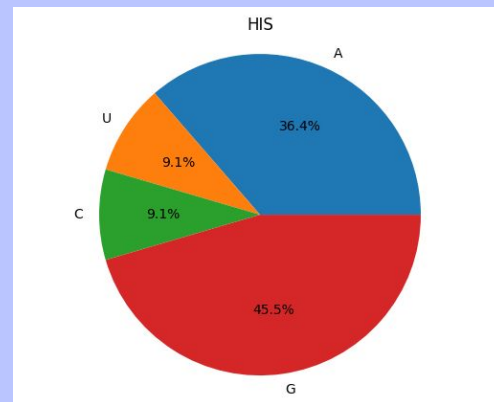
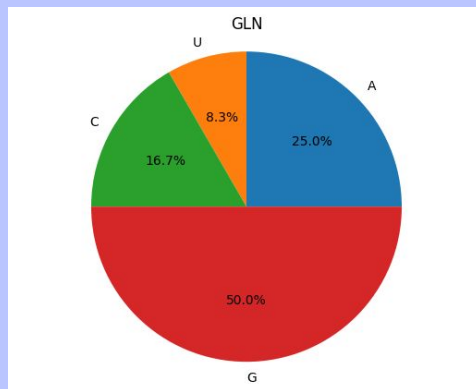
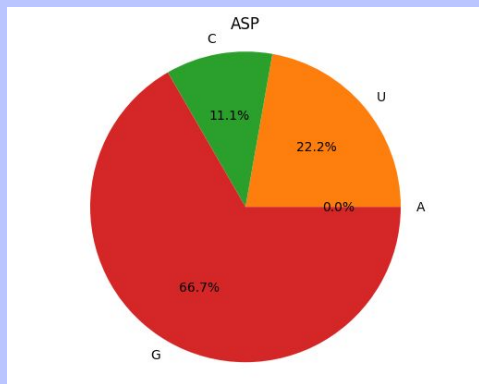
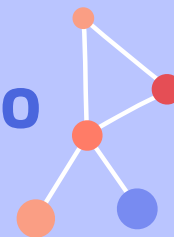
Frequência de Nucleotídeo por Aminoácido

Adenina - PDB



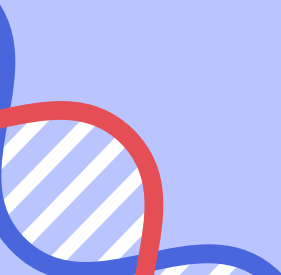
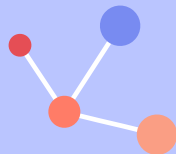
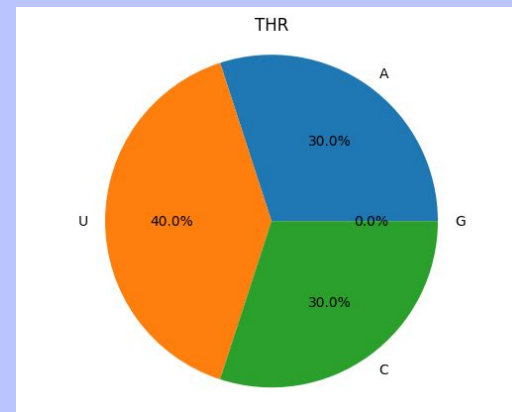
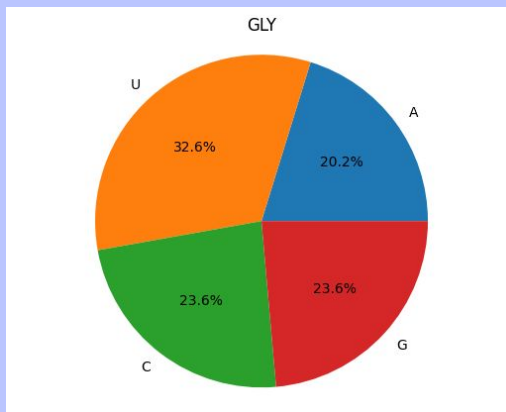
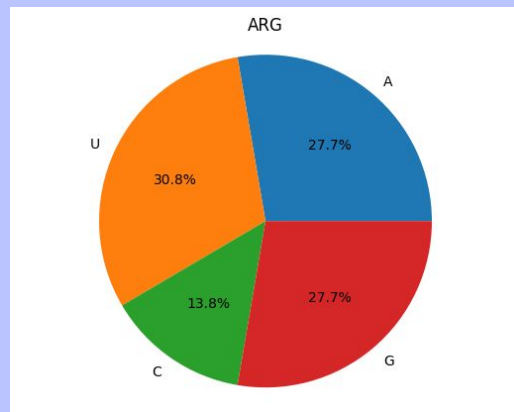
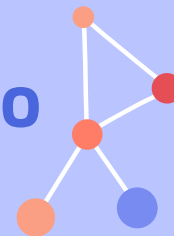
Frequência de Nucleotídeo por Aminoácido

Guanina - PDB



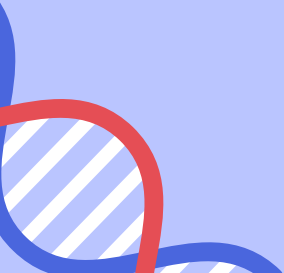
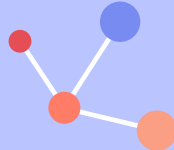
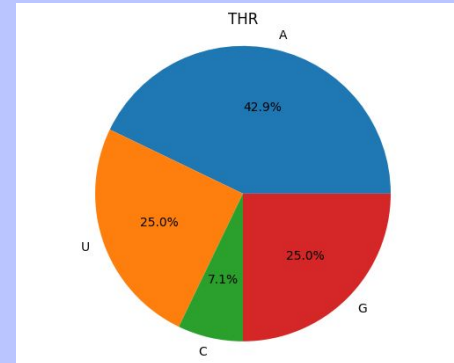
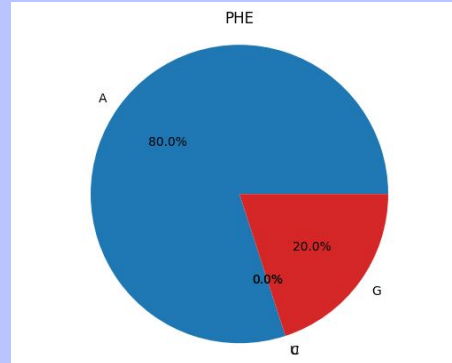
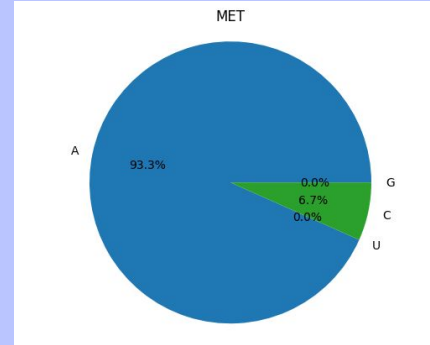
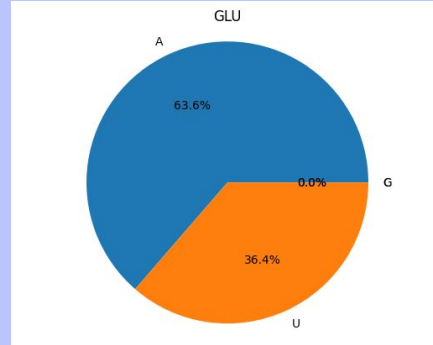
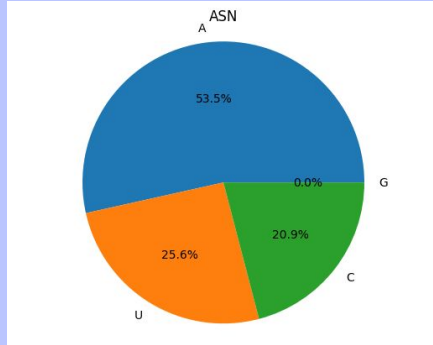
Frequência de Nucleotídeo por Aminoácido

Uracila - PDB



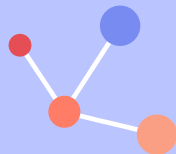
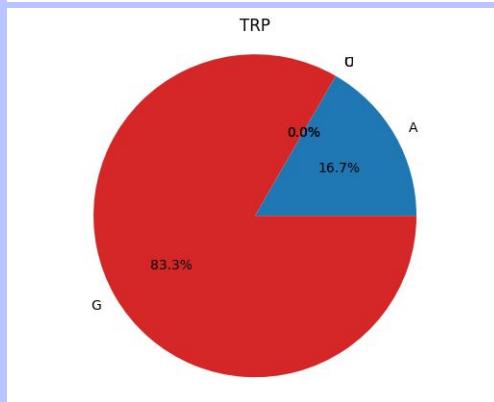
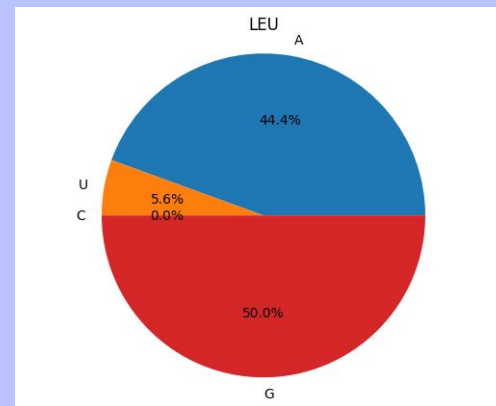
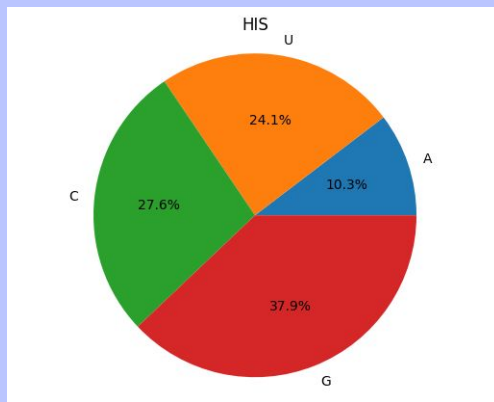
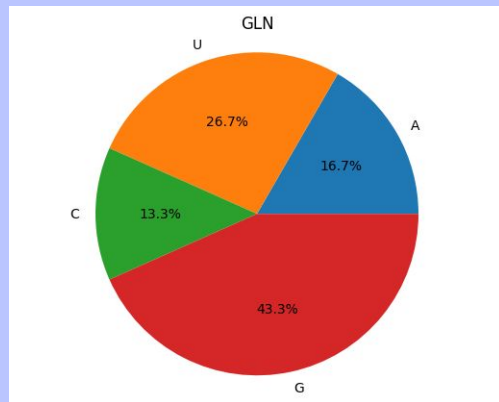
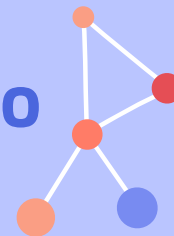
Frequência de Nucleotídeo por Aminoácido

Adenina - AlphaFold



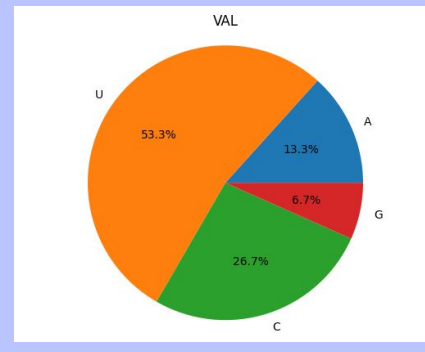
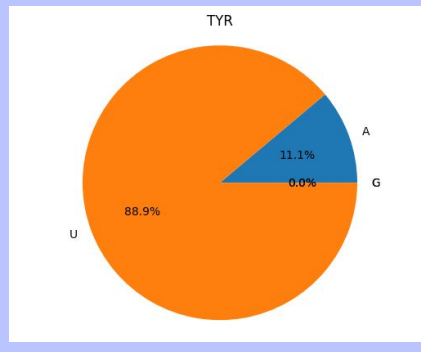
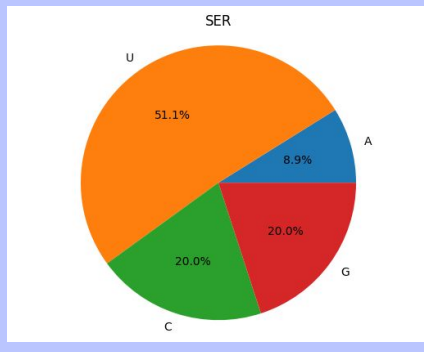
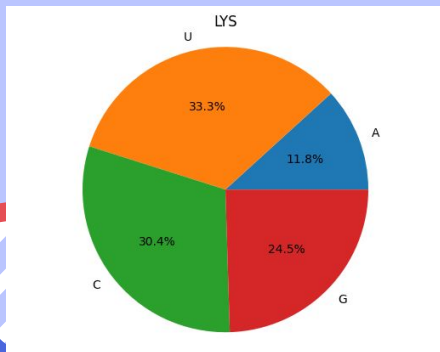
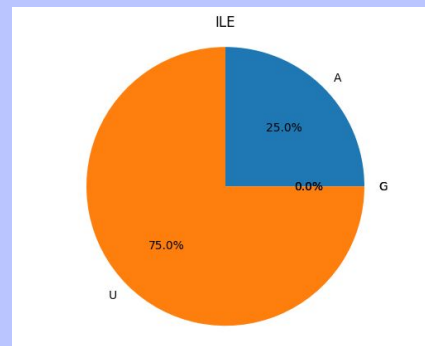
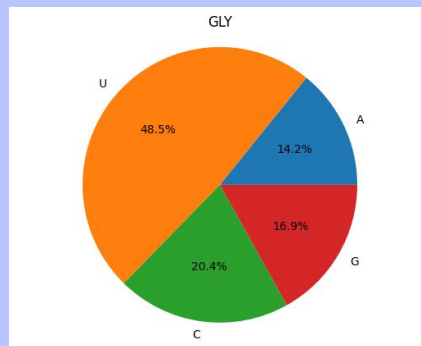
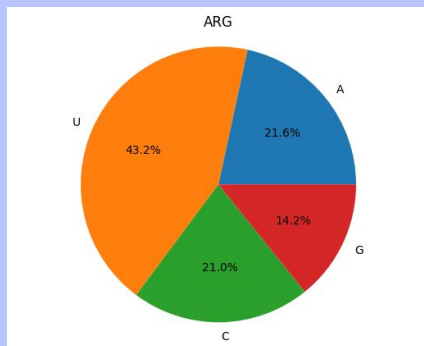
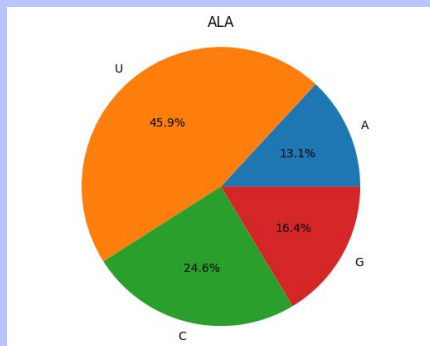
Frequência de Nucleotídeo por Aminoácido

Guanina - AlphaFold



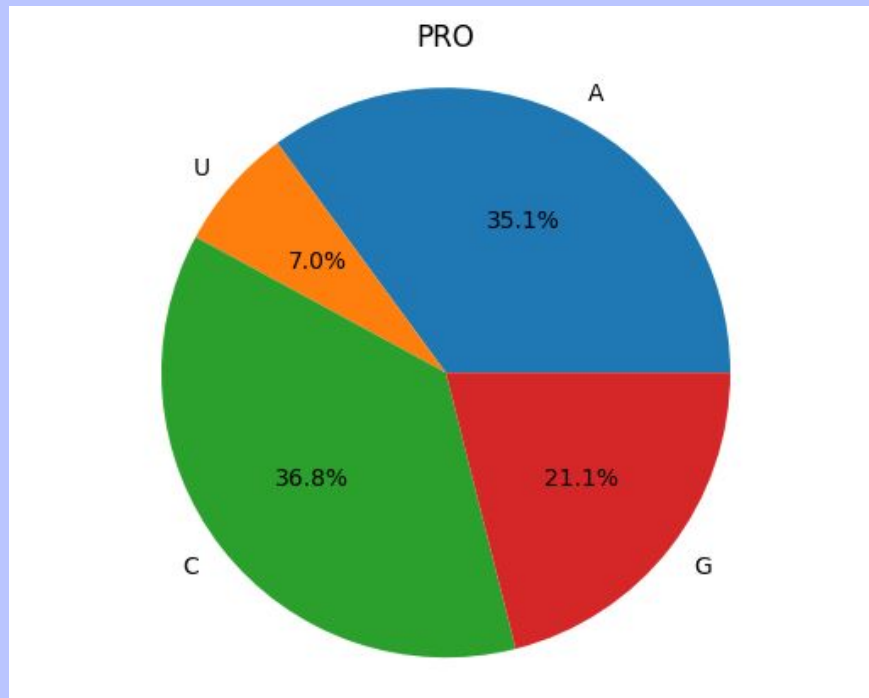
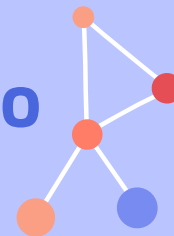
Frequência de Nucleotídeo por Aminoácido

Uracila - AlphaFold



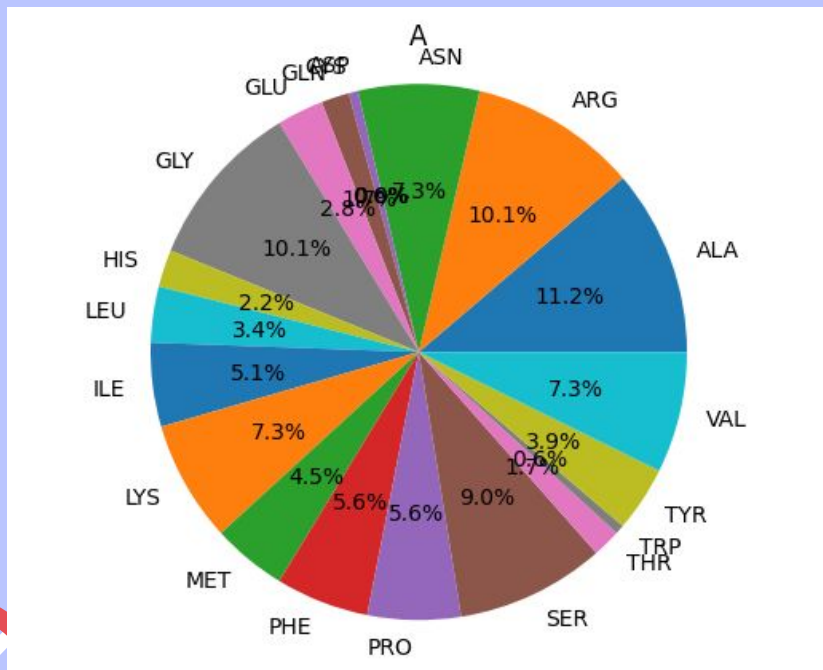
Frequência de Nucleotídeo por Aminoácido

Citosina - AlphaFold

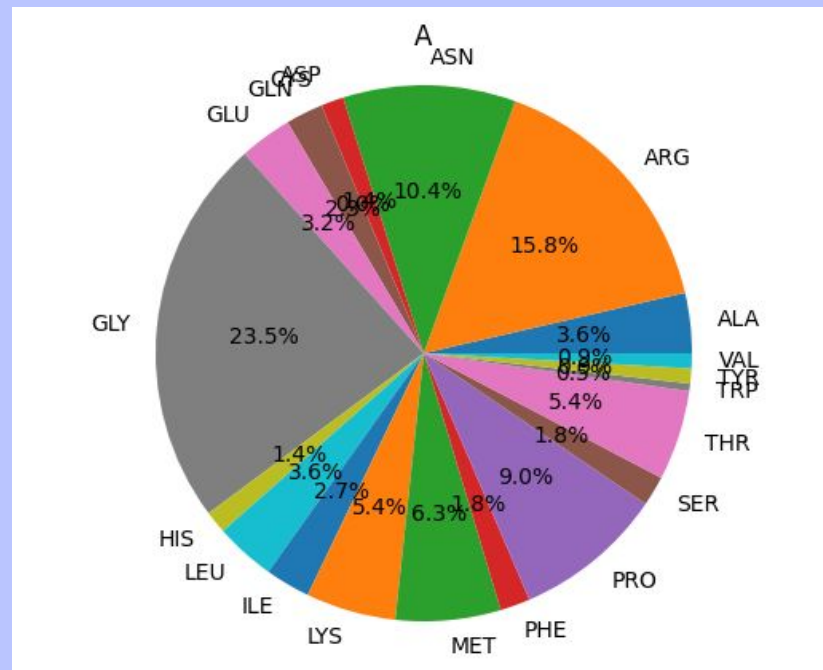


Frequência de Aminoácido por Nucleotídeo

PDB

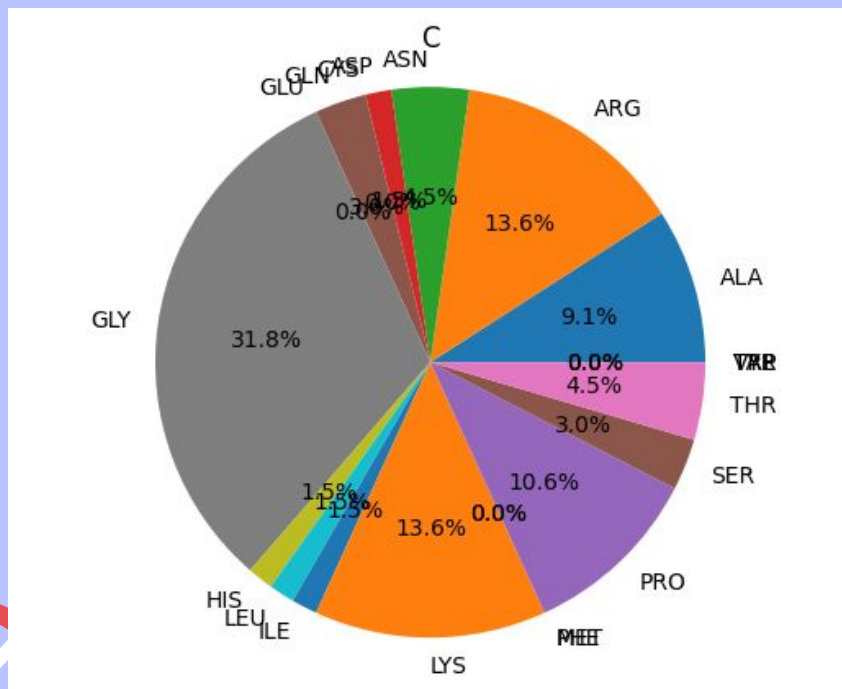


AlphaFold

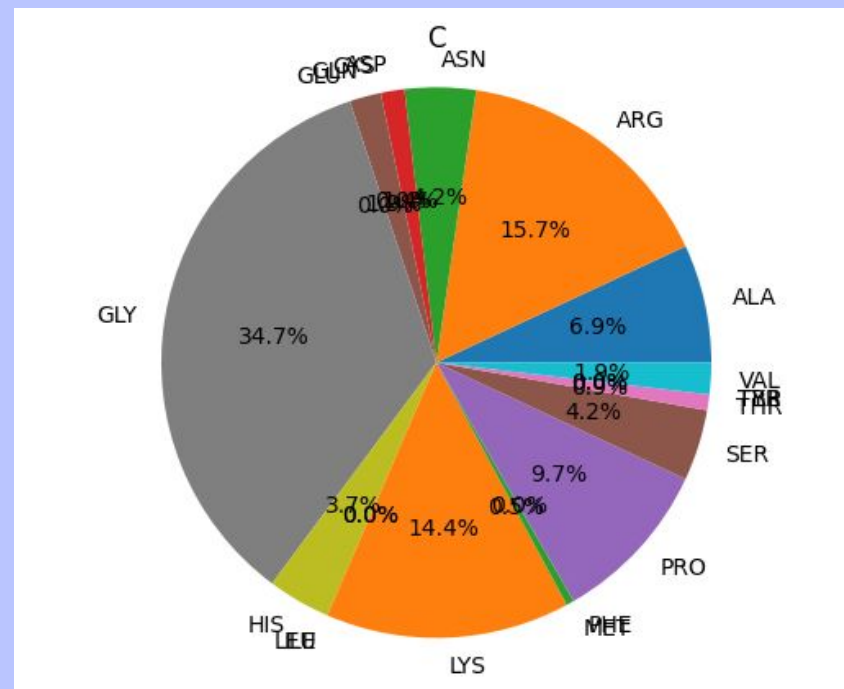


Frequência de Aminoácido por Nucleotídeo

PDB

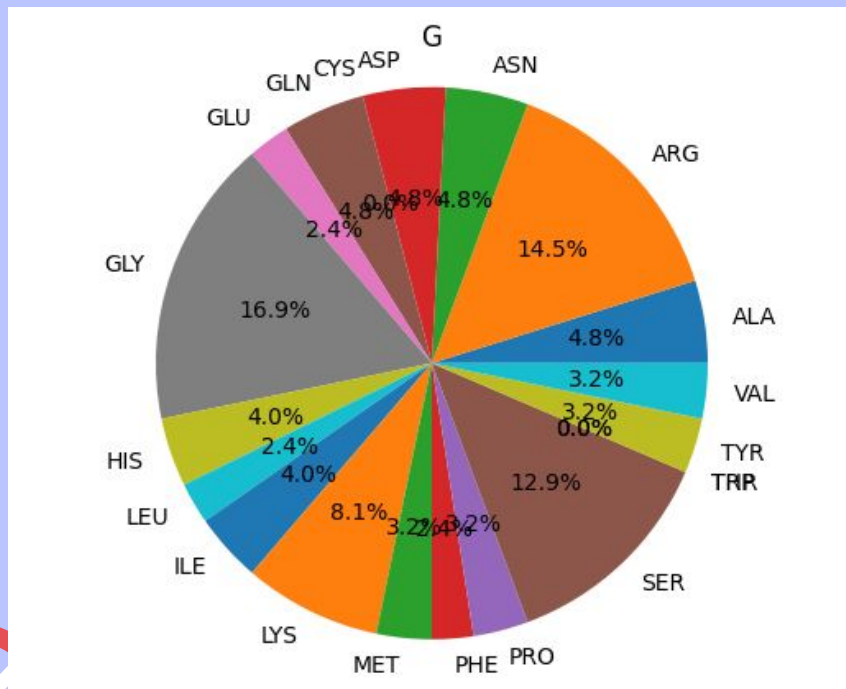


AlphaFold

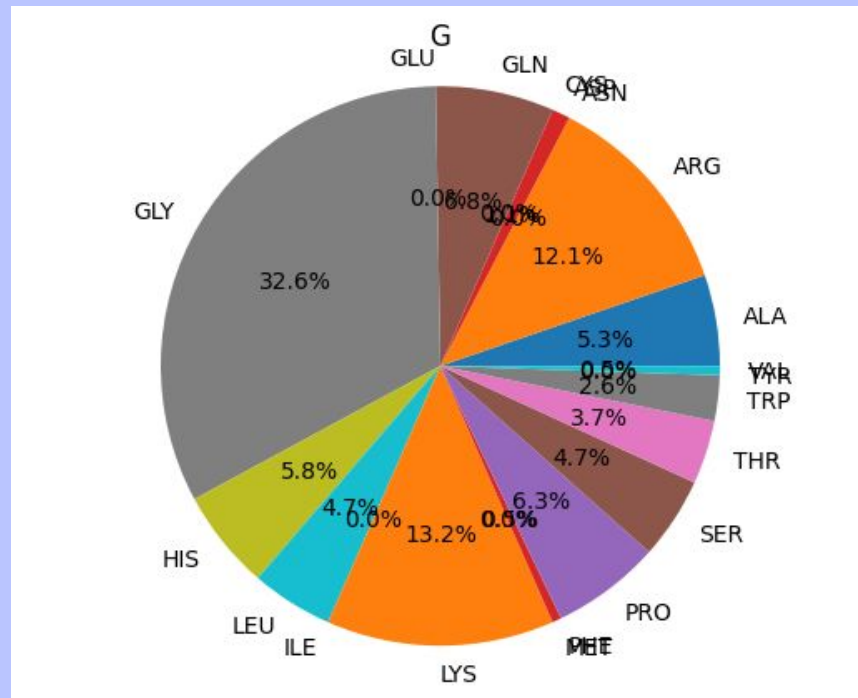


Frequência de Aminoácido por Nucleotídeo

PDB

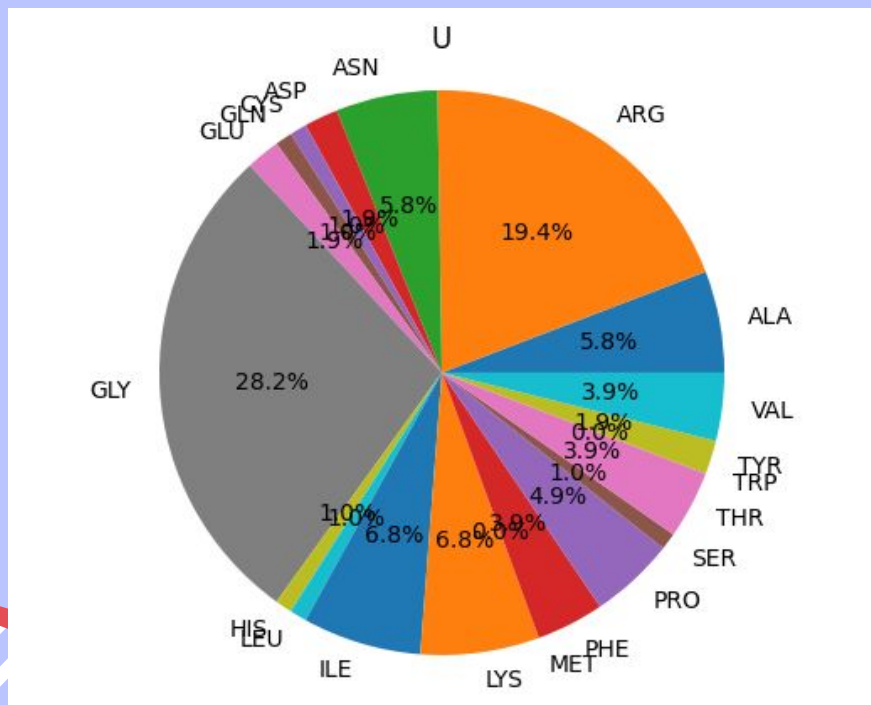


AlphaFold

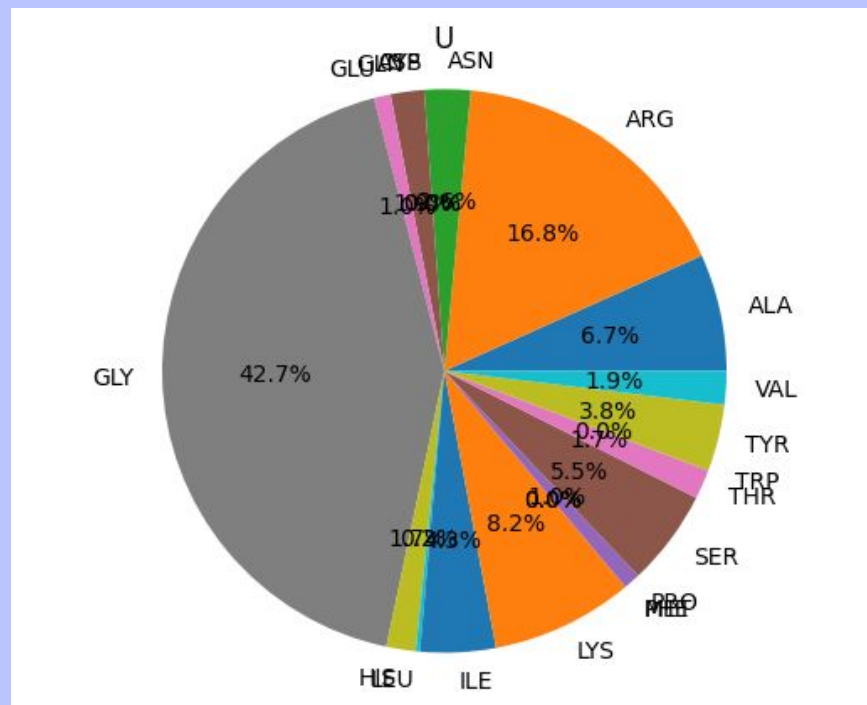


Frequência de Aminoácido por Nucleotídeo

PDB



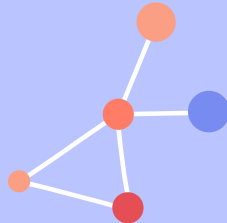
AlphaFold





Trabalhos Futuros

- Prever sítio de ligação com RNA
- Tipo de rna (padrão de sítio para transportador/mensageiro)





REFERÊNCIAS BIBLIOGRÁFICAS

Abramson, J., Adler, J., Dunger, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature (2024). <https://doi.org/10.1038/s41586-024-07487-w>

BERMAN, Helen M; BATTISTUZ, Tammy; BHAT, Talapady N; et al. The Protein Data Bank. Acta Crystallographica Section D-biological Crystallography, v. 58, n. 6, p. 899–907, 2002. Disponível em: <<https://scripts.iucr.org/cgi-bin/paper?an0594>>. Acesso em: 11 set. 2023.

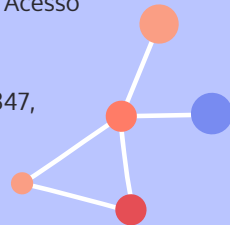
DE ARAÚJO, Nilberto Dias et al. A era da bioinformática: seu potencial e suas implicações para as ciências da saúde. Estudos de biologia, v. 30, n. 70/72, 2008. Disponível em: <<https://biblat.unam.mx/hevila/Estudiosdebiologia/2008/vol30/no70-72/16.pdf>>. Acesso em: 12 set. 2023.

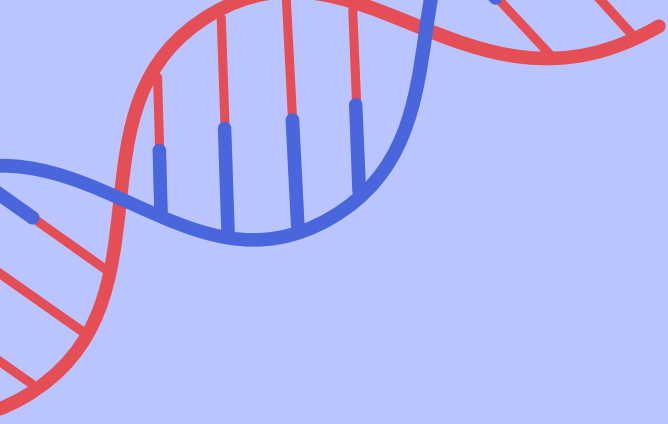
HERBERT, Katherine G; JUNILDA SPIROLLARI; WANG, Jianli; et al. Bioinformatic Databases. Wiley Encyclopedia of Computer Science and Engineering, 2007. Disponível em: <<https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470050118.ecse561>>. Acesso em: 13 set. 2023.

MARIANO, D. C. B.; BARROSO, J. R. P. M. ; CORREIA, T. S. ; de MELO-MINARDI, R. C. . Introdução à Programação para Bioinformática com Biopython. 3. ed. North Charleston, SC (EUA): CreateSpace Independent Publishing Platform, 2015. v. 1. 230p .

PIRES, Douglas E V; RAQUEL; CARLOS; et al. aCSM: noise-free graph-based signatures to large-scale receptor-based ligand prediction. Bioinformatics, v. 29, n. 7, p. 855–861, 2013. Disponível em: <<https://academic.oup.com/bioinformatics/article/29/7/855/253252>>. Acesso em: 11 set. 2023.

WU, Cathy H; YEH, Lai-Su L; HUANG, Hongzhan; et al. The Protein Information Resource. Nucleic Acids Research, v. 31, n. 1, p. 345–347, 2003. Disponível em: <<https://academic.oup.com/nar/article/31/1/345/2401247>>. Acesso em: 13 set. 2023.





Obrigado!

