UNIVERSIDADE FEDERAL DE MINAS GERAIS - UFMG
SISTEMAS DE INFORMAÇÃO - DCC

João Pedro Braga Ennes

Mineração e Análise de Dados de Bioinformática Estrutural: Construção de um Banco de Dados de Interações Proteína-RNA

Belo Horizonte - MG
2023

JOÃO PEDRO BRAGA ENNES

Mineração e Análise de Dados de Bioinformática Estrutural: Construção de um Banco de Dados de Interações Proteína-RNA

Projeto de Monografia em Sistemas de Informação I apresentado ao Departamento de Ciência da Computação da Universidade Federal de Minas Gerais como requisito para obtenção do título de Bacharel em Sistemas de Informação.

Prof. Orientadora: Raquel Cardoso de Melo Minardi.

Belo Horizonte - MG

2023

RESUMO

A compreensão da relação entre proteínas e RNA é crucial para entender os processos biológicos que regem a vida. Com o avanço da tecnologia, os estudos sobre essas interações vêm sendo guiados pelo uso intensivo de métodos computacionais que possam organizar e analisar as informações. Apesar disso, o acesso a informações estruturais que auxiliem nessas pesquisas, ainda é bem escasso e de difícil acesso. Isso se deve ao fato de não haver um banco de dados referencial, que concentre informações referente às interações proteína-RNA. Tendo em vista o problema, o objetivo dessa pesquisa foi construir um banco de dados que reunisse informações estruturais de interações proteína-RNA, visando auxiliar pesquisas futuras sobre o assunto. A metodologia envolveu a seleção criteriosa de dados do PDB, a análise no ambiente Google Colab com Python e Pandas, seguida pela validação e controle de qualidade. Foram coletadas 309 estruturas, refinando para 198 após validação. A análise exploratória destaca a diversidade nas sequências de proteínas e RNAs. A análise espacial revela padrões de proximidade entre aminoácidos e nucleotídeos. A glicina destaca-se na proximidade de diversas bases, enquanto a adenina demonstra uma presença notável em 11 dos 20 aminoácidos. O estudo conclui que o banco de dados construído e as análises são fundamentais para avançar na compreensão das interações proteína-RNA, destacando possíveis direcionamentos futuros, como a avaliação de modelos preditivos e a busca por padrões específicos de sítios de ligação.

Palavras-chave: Bioinformática, Proteína, RNA, Banco de Dados.

ABSTRACT

A comprehension of the relationship between proteins and RNA is crucial for understanding the biological processes that govern life. With technological advancements, studies on these interactions have been guided by the intensive use of computational methods to organize and analyze information. Despite this, access to structural information that aids in these studies is still scarce and challenging. This is due to the absence of a reference database that consolidates information regarding protein-RNA interactions. Given this issue, the aim of this research was to construct a database that gathers structural information on protein-RNA interactions, with the goal of assisting future research on the subject. The methodology involved the careful selection of data from the PDB, analysis in the Google Colab environment using Python and Pandas, followed by validation and quality control. Initially, 309 structures were collected, refined to 198 after validation. Exploratory analysis highlights the diversity in protein and RNA sequences. Spatial analysis reveals proximity patterns between amino acids and nucleotides. Glycine stands out in the proximity of various bases, while adenine demonstrates a notable presence in 11 out of 20 amino acids. The study concludes that the constructed database and analyses are fundamental for advancing the understanding of protein-RNA interactions, emphasizing potential future directions such as evaluating predictive models and searching for specific binding site patterns.

Keywords: Bioinformatics, Protein, RNA, Database.

SUMÁRIO

1. INTRODUÇÃO

A relação entre proteínas e RNA é fundamental para a compreensão dos processos biológicos que regem a vida. Essa interação desempenha um papel crucial em funções celulares, desde a regulação da expressão gênica até a manutenção da integridade do genoma (PIRES *et al*., 2013). A compreensão dessas interações tem implicações profundas para a pesquisa biomédica e o avanço do conhecimento biológico. No entanto, apesar de sua importância, a disponibilidade de um banco de dados públicos (BDBs) em bioinformática abrangente e acessível que contenha informações detalhadas sobre interações proteína-RNA ainda é limitada.

Juntamente a isso, as pesquisas na área de biomoléculas têm sido amplamente guiadas pelo uso intensivo de métodos computacionais para a organização e análise das informações (DE ARAÚJO *et al*., 2008). Isso faz com que um banco de dados dedicado a interações proteína-RNA seja essencial para avançar nossa compreensão das complexas redes regulatórias que governam a vida celular, por proporcionar um conjunto de dados sólido, para utilização em estudos aplicados com utilização de modelos matemáticos.

Este projeto busca, portanto, construir um banco de dados que se concentre especificamente em armazenar informações sobre sequências e estruturas envolvidas em interações proteína-RNA, onde será possível reunir e disponibilizar estes dados. Neste contexto, serão realizadas coletas de informações a partir do Protein Data Bank (PDB) (BERMAN *et al*., 2002), a fim de reunir dados de estruturas moleculares já sequenciadas e construir um recurso valioso para a pesquisa na área de bioinformática. Este BDBs será um recurso de relevância para a comunidade de pesquisa, permitindo o acesso a informações estruturais que são fundamentais para a compreensão das interações proteína-RNA. Em seguida, com os dados devidamente tratados e armazenados, pretende-se realizar uma análise, visando buscar afinidades entre as interações das diferentes proteínas com o RNA.

Ao longo deste trabalho, será demonstrado que a criação de um banco de dados de interações proteína-RNA não é apenas um passo importante na compreensão da bioinformática estrutural, mas, também, uma ferramenta poderosa para impulsionar descobertas científicas e aplicações clínicas inovadoras. O aprofundamento do conhecimento nesse nicho tem o potencial de revolucionar a pesquisa na área das biomoléculas e oferecer

avanços para pesquisas acadêmicas, além de ser referência para a prática médica e biomédica ao direcionar novas abordagens para diagnósticos, tratamentos e prevenção de doenças.

2. REFERENCIAL TEÓRICO

As interações proteína-RNA desempenham um papel vital nos processos biológicos, influenciando a expressão gênica, a estabilidade do RNA, a tradução e a regulação de uma variedade de funções celulares (PIRES *et al*., 2013). As proteínas interagem com moléculas de RNA para desempenhar funções específicas, muitas vezes reconhecendo sequências estruturais ou sequências específicas de RNA. O estudo dos dados moleculares gerados a partir de análises de sequenciamento das moléculas de DNA e RNA, são essenciais para decifrar características estruturais de segmentos gênicos e de seus produtos protéicos, bem como estabelecer suas interações e compreender processos biológicos, como tradução, processamento de RNA e regulação pós-transcricional (JÚNIOR *et al.*, 2022).

Apesar de possuir tal importância, o sequenciamento estrutural de interações proteína-RNA ainda é uma área pouco explorada. No PDB, banco de dados fundamental nas áreas de biologia estrutural, que armazena estruturas tridimensionais de proteínas, podemos encontrar mais de duzentas mil estruturas experimentais registradas, mas, ao fazer uma filtragem rápida, pode ser notado que apenas sete mil têm alguma relação com uma estrutura da RNA.

Bancos de dados biológicos, como o PDB, são repositórios de informações biológicas que incluem sequências de proteínas, RNA, estruturas tridimensionais, informações de expressão gênica e interações biomoleculares. Esses BDBs são essenciais para a pesquisa de biomoléculas, de forma que permitem o acesso a informações valiosas para análises e descobertas (MARIANO *et al*., 2015). No contexto deste projeto, o foco será a construção de um banco de dados específico para interações proteína-RNA e utilização para análise e busca de similaridades entre tais interações.

Na área de bioinformática, é possível encontrar três tipos de BDBs, que variam de acordo com o conteúdo neles armazenado. O PDB, citado anteriormente, é um exemplo de banco primário que propõe armazenar os dados originais de sequências de nucleotídeos e proteínas utilizando arquivos de texto simples para armazenar as informações. Além deste, temos também bancos secundários que, por sua vez, são utilizados para armazenar resultados de análises feitas a partir de dados primários, esses já utilizam de Sistemas de Gerenciamento de Banco de Dados, como MySQL, PostgreSQL, ORACLE, etc. Um exemplo de banco

secundário também especializado em informações referentes às proteínas é o Protein Information Resources (PIR) (WU *et al*., 2003). Por fim, temos os bancos de dados especializados, que se diferem dos outros por atenderem a um interesse particular de pesquisa, ou seja, são especializados para um particular organismo ou tipo de dado (HERBERT *et al*., 2007).

3. METODOLOGIA

Neste capítulo, serão descritas as etapas metodológicas realizadas no desenvolvimento do projeto de Mineração e Análise de Dados de Bioinformática Estrutural, com foco na construção de um Banco de Dados de Interações Proteína-RNA. As principais fases incluíram a seleção de dados do Protein Data Bank (PDB) e a análise subsequente desses dados usando Google Colab com Python e a biblioteca Pandas.

3.1 Seleção de dados no Protein Data Bank (PDB)

3.1.1 Critérios de seleção

A escolha adequada dos dados é crucial para assegurar a qualidade e relevância das informações no banco de dados. Os critérios de seleção foram definidos com base na natureza da interação proteína-RNA, buscando garantir que os dados refletissem fielmente essas interações.

Os filtros aplicados para a obtenção dos dados incluíram a seguinte condição lógica: "Polymer Entity is Protein AND Polymer Entity is RNA AND Number of Distinct Molecular Entities = 2". Isso foi implementado para garantir que os dados obtidos representassem especificamente interações entre uma entidade polimérica de proteína e uma entidade polimérica de RNA, com exatamente duas entidades moleculares distintas.

3.1.2 Ferramentas utilizadas

Para a extração dos dados do PDB, foi utilizado um script em Python, empregando bibliotecas como Biopython para manipulação de estruturas biológicas e Pandas para a organização e estruturação dos dados. Esse script foi desenvolvido para acessar a API do PDB e fazer o download dos arquivos estruturais que atendiam aos critérios estabelecidos.

3.2 Análise dos dados com Google Colab e Pandas

3.2.1 Ambiente de desenvolvimento

A análise dos dados foi conduzida no ambiente Google Colab, uma plataforma baseada em nuvem que oferece ambientes de execução Jupyter Notebook. A escolha do Google Colab foi motivada pela facilidade de colaboração, integração com o Google Drive para armazenamento de dados e recursos computacionais escaláveis.

O ambiente Python no Google Colab foi configurado com bibliotecas essenciais, como Pandas, NumPy, e Matplotlib, proporcionando um ambiente computacional rico para análises detalhadas.

3.2.2 Análise exploratória dos dados

A análise exploratória dos dados foi realizada utilizando a biblioteca Pandas, permitindo a manipulação eficiente de conjuntos de dados. Essa etapa incluiu a identificação de características relevantes, como características estruturais das interações proteína-RNA, distribuição de comprimentos de sequências e análise de resolução espacial.

3.2.3 Construção do Banco de Dados

Com base nas informações obtidas durante a análise exploratória, procedeu-se à construção do Banco de Dados de Interações Proteína-RNA. Utilizando Pandas, os dados foram organizados em um formato estruturado e eficiente, com ênfase na indexação adequada e integridade dos dados.

3.3 Validação e Controle de Qualidade

A validação do banco de dados foi realizada através da verificação da consistência dos dados obtidos com a literatura existente sobre interações proteína-RNA. Além disso, foram implementados procedimentos de controle de qualidade para identificar possíveis anomalias e removê-las de forma a garantir a integridade dos dados e evitar que as análises fossem enviesadas.

3.4 Limitações do Estudo

É importante ressaltar que este estudo tem algumas limitações, incluindo a dependência da qualidade e representatividade dos dados disponíveis no PDB. Além disso, as análises realizadas estão intrinsecamente ligadas à qualidade das estruturas biológicas depositadas e disponíveis para pesquisa.

4. ANÁLISE E DISCUSSÃO DOS RESULTADOS

Inicialmente, foram coletadas 309 estruturas segundo os padrões citados na etapa de metodologia. Após o processo de validação e controle de qualidade, o banco foi refinado para 198 estruturas que atendiam aos critérios estabelecidos, ou seja, continham uma molécula de proteína e uma molécula de RNA.

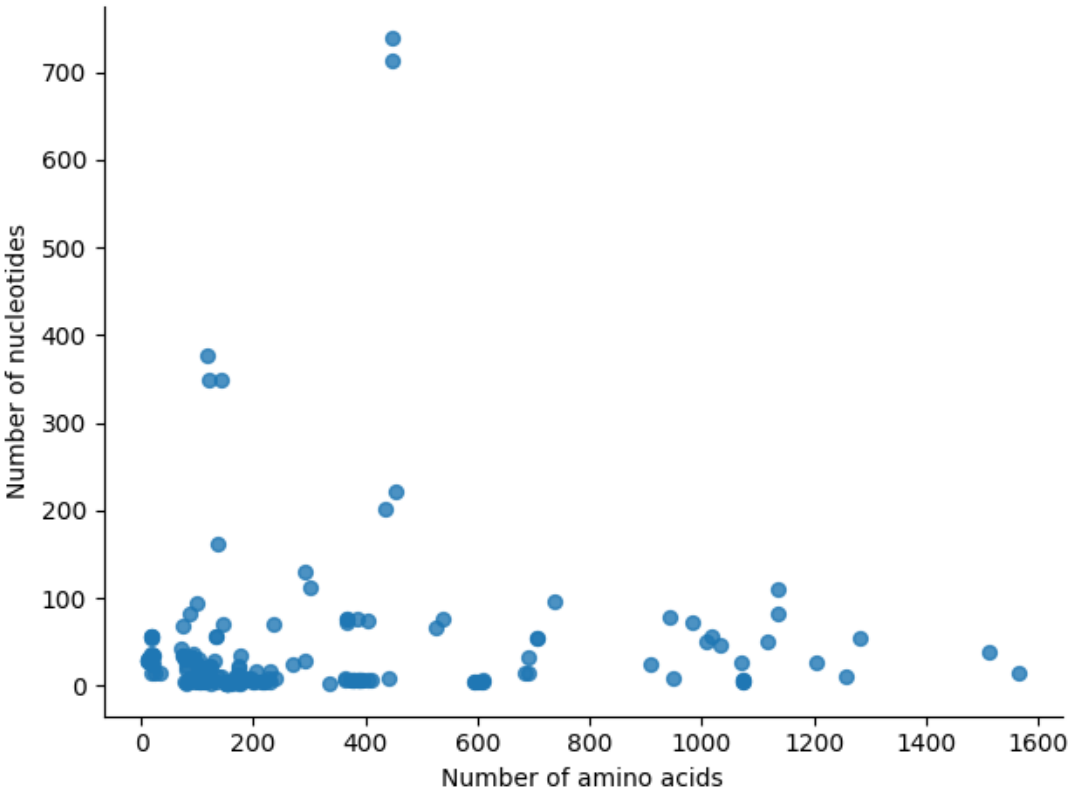
4.1 Características gerais do banco de dados

4.1.1 Composição estrutural

A composição estrutural do banco de dados após a validação revelou uma seleção representativa de 198 estruturas de interações proteína-RNA. A consistência dessas estruturas foi fundamental para garantir a qualidade das análises subsequentes.

4.1.2 Distribuição de tamanho de proteínas e RNAs

Gráfico 1 — Gráfico de dispersão de número de nucleotídeos pelo número de aminoácidos da estrutura.



Fonte: Dados da pesquisa (2023)

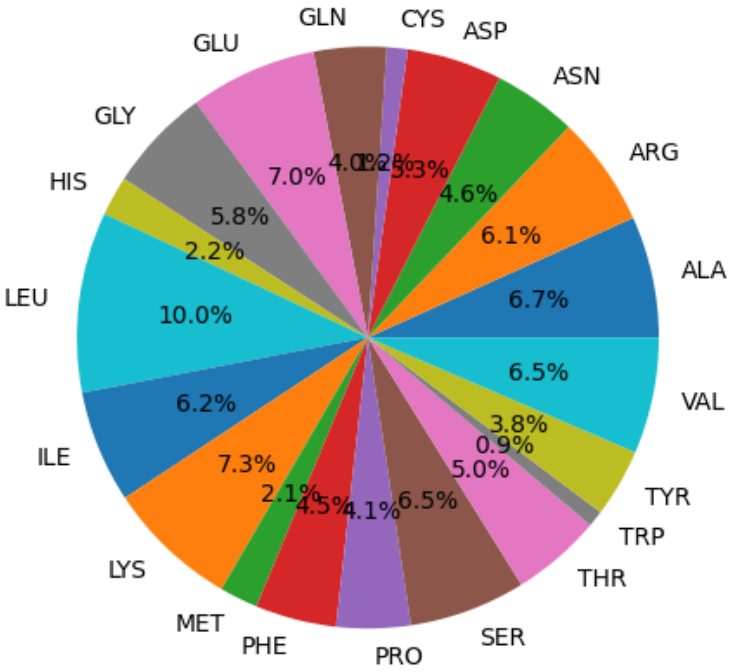
A análise da distribuição de tamanho de proteínas mostrou que a maioria das proteínas no banco possui entre 1 e 400 aminoácidos, sugerindo uma diversidade considerável de tamanhos. Esse resultado é relevante, pois evidencia a representatividade de diferentes proteínas envolvidas em interações com RNAs.

Quanto aos RNAs, a maioria apresentou um tamanho de 1 a 120 nucleotídeos. Essa distribuição também reflete a diversidade de tamanhos de RNAs presentes nas interações estudadas, indicando uma variedade de complexidades nas estruturas formadas.

4.2 Composição de aminoácidos nas proteínas

4.2.1 Perfil geral de aminoácidos

Gráfico 2 — Percentual da frequência de cada aminoácido no banco.



Fonte: Dados da pesquisa (2023)

A análise do perfil de aminoácidos revelou uma distribuição equilibrada no geral. No entanto, alguns aminoácidos se destacaram em termos de frequência no banco de dados. A leucina, por exemplo, apresentou a maior porcentagem, representando 10% do total de aminoácidos nas proteínas estudadas. Esse achado pode sugerir uma preferência ou recorrência na presença de leucina nas interações proteína-RNA analisadas. Por outro lado, o triptofano se destacou como o aminoácido menos frequente, com apenas 0.9% de presença. Essa observação pode indicar uma menor prevalência do triptofano nessas interações específicas.

4.2.2 Implicações biológicas

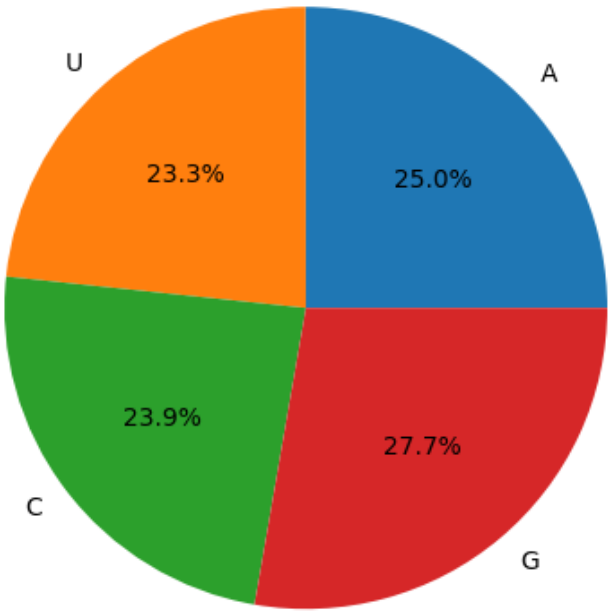
A presença diferencial de aminoácidos pode ter implicações biológicas significativas. A leucina, por exemplo, é conhecida por desempenhar papéis cruciais em interações proteína-RNA, agindo estimulando a fase de iniciação da tradução do RNA mensageiro em proteína (GONÇALVES et al., 2013). A compreensão desses perfis de aminoácidos pode fornecer informações valiosas sobre os padrões de reconhecimento molecular em interações proteína-RNA.

4.4 Composição de Nucleotídeos nos RNAs

4.4.1 Proporção Equilibrada de Nucleotídeos

A análise da composição de nucleotídeos nos RNAs das interações proteína-RNA revelou uma distribuição equilibrada, indicando uma diversidade significativa nas sequências nucleotídicas presentes no banco de dados.

Gráfico 3 — Percentual da frequência de cada nucleotídeo no banco.



Fonte: Dados da pesquisa (2023)

Entre os nucleotídeos, a guanina se destacou com a maior presença, representando 27.7% do total. Esse achado sugere uma prevalência dessa base nitrogenada nas interações estudadas, podendo indicar regiões específicas de ligação ou estabilidade nas estruturas formadas. Contrastando com a guanina, a uracila apresentou a menor presença, totalizando 23.3%. Essa observação aponta para uma distribuição diferencial de nucleotídeos, destacando a importância de considerar não apenas a presença de bases específicas, mas também suas proporções relativas nas interações proteína-RNA.

4.4.2 Implicações Biológicas

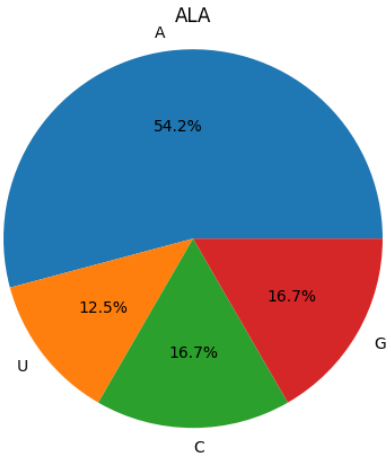
A análise abrangente da composição de nucleotídeos nos RNAs adiciona camadas significativas ao entendimento das interações proteína-RNA. A diversidade equilibrada de nucleotídeos reflete a complexidade das sequências envolvidas, indicando potenciais padrões de reconhecimento molecular.

O destaque para a guanina pode sugerir funções específicas dessa base nitrogenada nas interações estudadas. A compreensão desses padrões é crucial para avançar no conhecimento da regulação genética, processos de tradução e outras vias biológicas relacionadas.

Em conjunto, as análises estruturais e de composição química apresentadas neste capítulo proporcionam uma visão abrangente das interações proteína-RNA no contexto bioinformático estrutural. Esses resultados não apenas validam a qualidade do banco de dados construído, mas também abrem portas para investigações mais aprofundadas sobre as implicações funcionais e biológicas dessas interações.

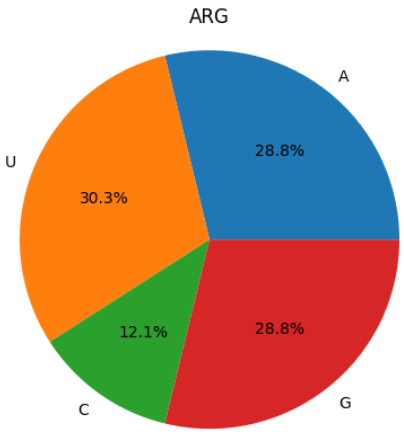
4.6 Análise de Relações entre Aminoácidos e Nucleotídeos

Gráfico 4 — Percentual da frequência de nucleotídeos próximos a Alanina.



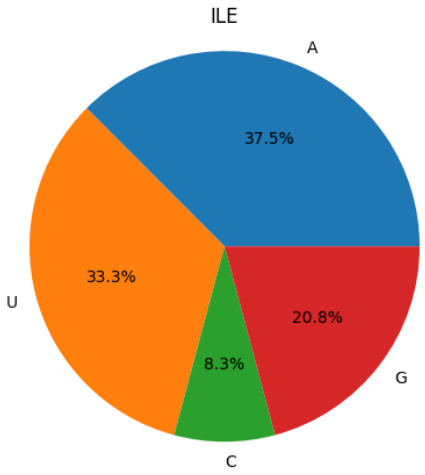
Fonte: Dados da pesquisa (2023)

Gráfico 5 — Percentual da frequência de nucleotídeos próximos a Arginina.



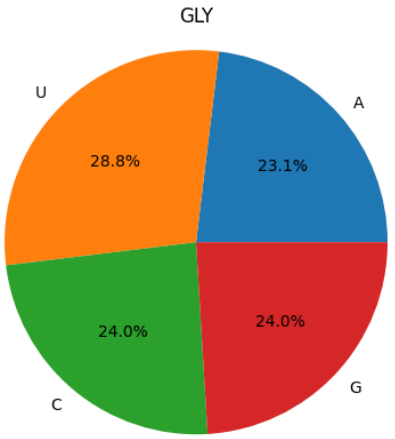
Fonte: Dados da pesquisa (2023)

Gráfico 6 — Percentual da frequência de nucleotídeos próximos a Isoleucina.



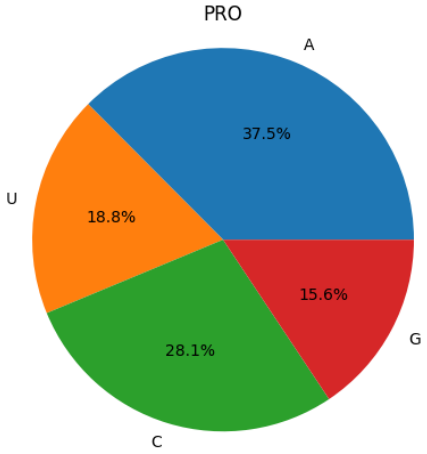
Fonte: Dados da pesquisa (2023)

Gráfico 7 — Percentual da frequência de nucleotídeos próximos a Glicina.



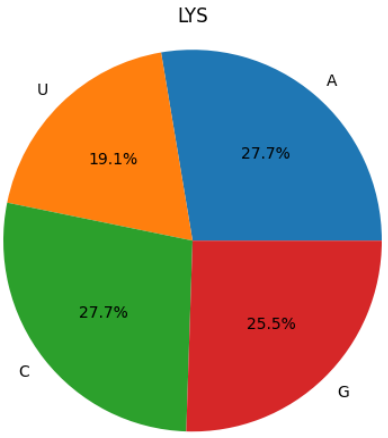
Fonte: Dados da pesquisa (2023)

Gráfico 8 — Percentual da frequência de nucleotídeos próximos a Prolina.



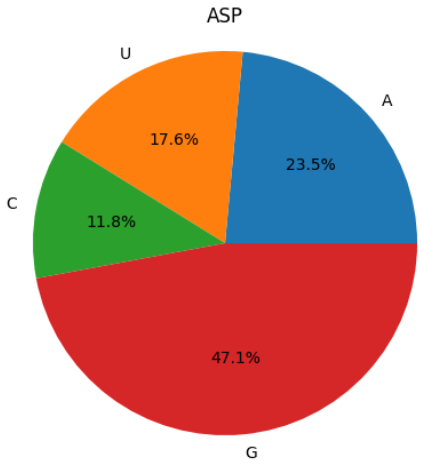
Fonte: Dados da pesquisa (2023)

Gráfico 9 — Percentual da frequência de nucleotídeos próximos a Lisina.



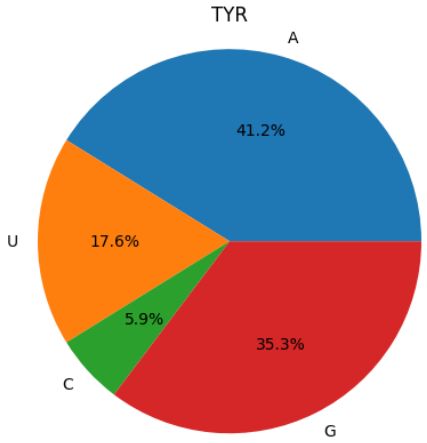
Fonte: Dados da pesquisa (2023)

Gráfico 10 — Percentual da frequência de nucleotídeos próximos ao Ácido aspártico.



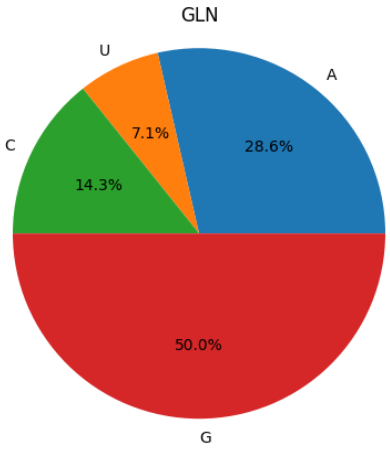
Fonte: Dados da pesquisa (2023)

Gráfico 11 — Percentual da frequência de nucleotídeos próximos a Tirosina.



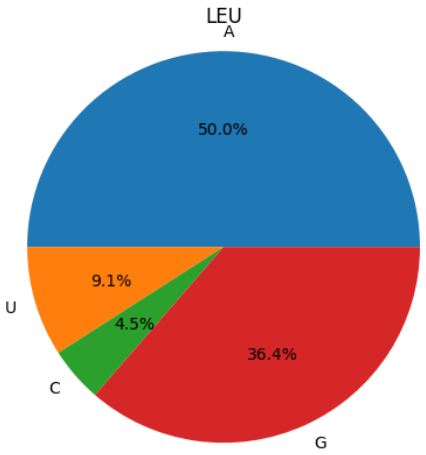
Fonte: Dados da pesquisa (2023)

Gráfico 12 — Percentual da frequência de nucleotídeos próximos a Glicina.



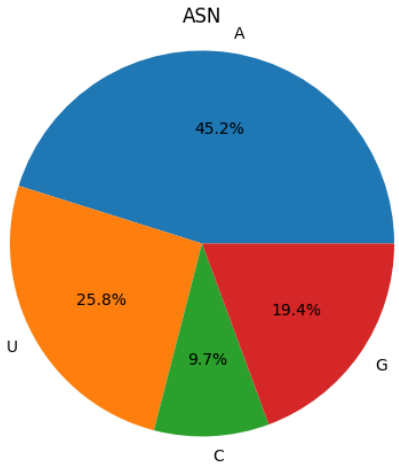
Fonte: Dados da pesquisa (2023)

Gráfico 13 — Percentual da frequência de nucleotídeos próximos a Leucina.



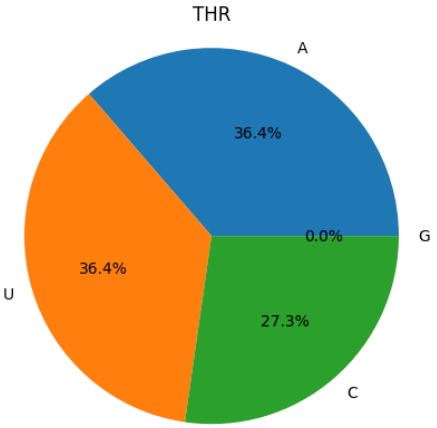
Fonte: Dados da pesquisa (2023)

Gráfico 14 — Percentual da frequência de nucleotídeos próximos a Asparagina.



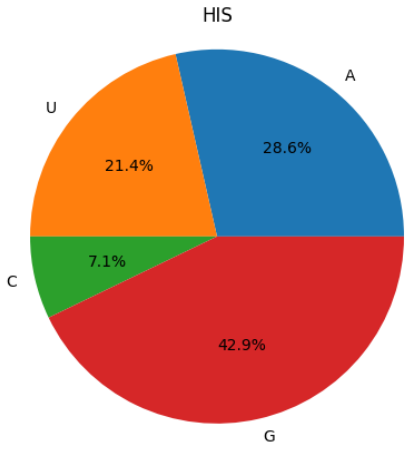
Fonte: Dados da pesquisa (2023)

Gráfico 15 — Percentual da frequência de nucleotídeos próximos a Treonina.



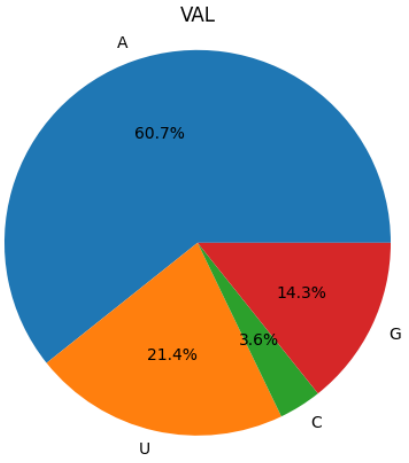
Fonte: Dados da pesquisa (2023)

Gráfico 16 — Percentual da frequência de nucleotídeos próximos a Histidina.



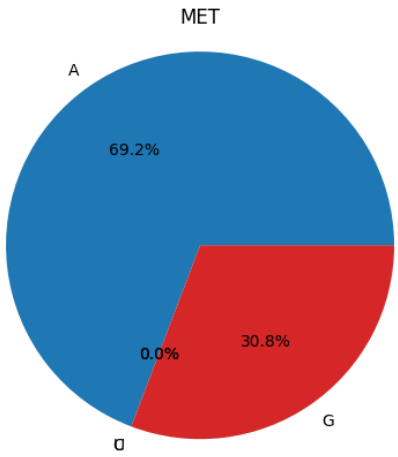
Fonte: Dados da pesquisa (2023)

Gráfico 17 — Percentual da frequência de nucleotídeos próximos a Valina.



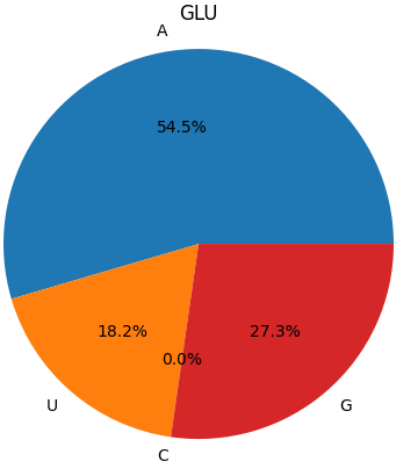
Fonte: Dados da pesquisa (2023)

Gráfico 18 — Percentual da frequência de nucleotídeos próximos a Metionina.



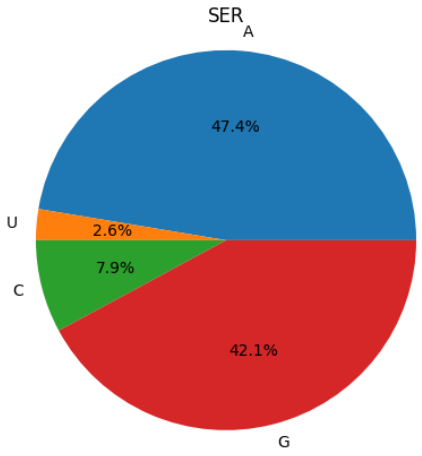
Fonte: Dados da pesquisa (2023)

Gráfico 19 — Percentual da frequência de nucleotídeos próximos ao Ácido glutâmico.



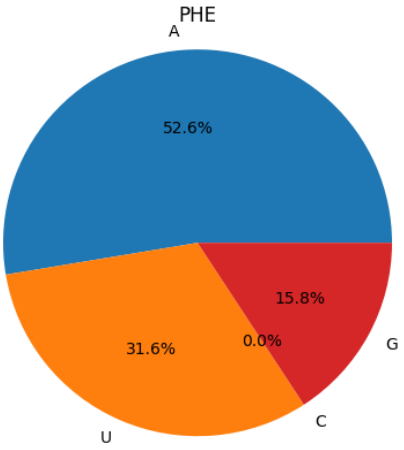
Fonte: Dados da pesquisa (2023)

Gráfico 20 — Percentual da frequência de nucleotídeos próximos a Serina.



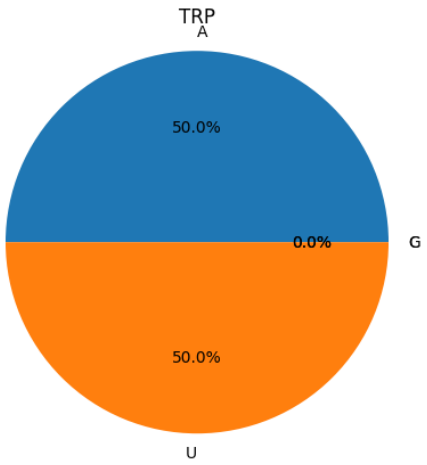
Fonte: Dados da pesquisa (2023)

Gráfico 21 — Percentual da frequência de nucleotídeos próximos a Fenilalanina.



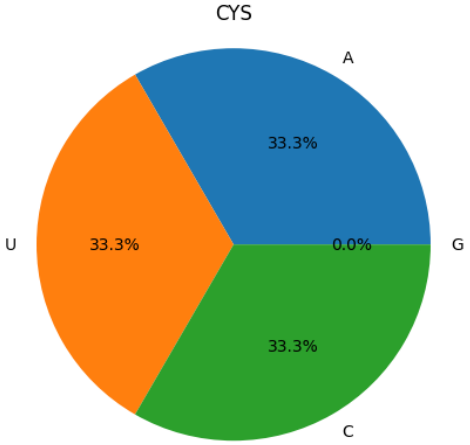
Fonte: Dados da pesquisa (2023)

Gráfico 22 — Percentual da frequência de nucleotídeos próximos ao Triptofano.



Fonte: Dados da pesquisa (2023)

Gráfico 23 — Percentual da frequência de nucleotídeos próximos a Cisteína.



Fonte: Dados da pesquisa (2023)

4.6.1 Metodologia de Análise Espacial

O próximo passo da pesquisa visou explorar possíveis relações espaciais entre aminoácidos e nucleotídeos nas interações proteína-RNA. Foi utilizado um método que envolveu a identificação das posições de cada nucleotídeo e do carbono alfa de cada aminoácido. Em seguida, foi traçado um raio de 5 angstroms ao redor de cada aminoácido para analisar a frequência de cada nucleotídeo nesse perímetro.

4.6.2 Destaques na Frequência de Nucleotídeos

Foram observados resultados notáveis ao analisar a frequência de nucleotídeos nas proximidades dos aminoácidos. Em 11 dos 20 aminoácidos estudados, destacou-se uma maior presença de adenina comparada aos outros nucleotídeos.

- Adenina (A) em Destaque: Alanina, Asparagina, Ácido glutâmico, Isoleucina, Leucina, Lisina, Metionina, Prolina, Serina, Tirosina e Valina.

- Guanina (G) e Uracila (U) Empatadas: Em três casos (Ácido aspártico, Glutamina e Histidina), a guanina e a uracila apresentaram frequências semelhantes (Treonina, Glicina e Arginina).

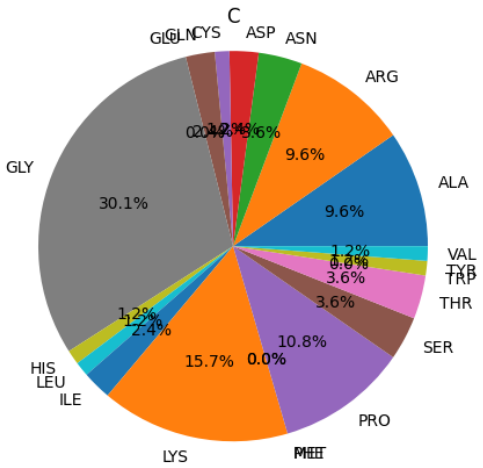
- Ausência de Citosina (C) como Predominante: Em nenhum dos casos, a citosina foi a base nitrogenada mais predominante na vizinhança dos aminoácidos.

4.6.3 Casos Específicos

Em dois casos específicos, Cisteína e Triptofano, não foi possível identificar uma quantidade relevante de nucleotídeos na vizinhança, com apenas 3 e 2 registros, respectivamente. Essa observação pode sugerir peculiaridades na interação desses aminoácidos com os nucleotídeos nas estruturas analisadas.

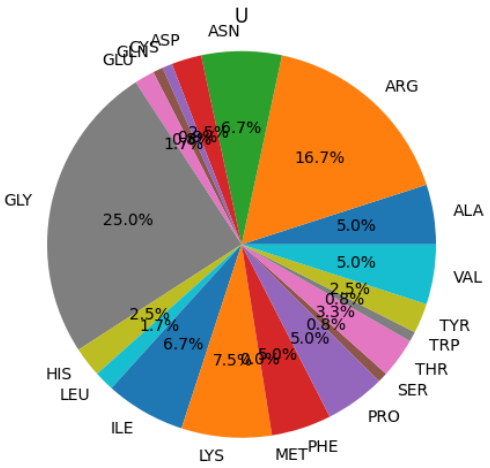
4.8 Análise Recíproca: Frequência de Aminoácidos em Proximidade de Nucleotídeos

Gráfico 24 — Percentual da frequência de aminoácidos próximos a Citosina.



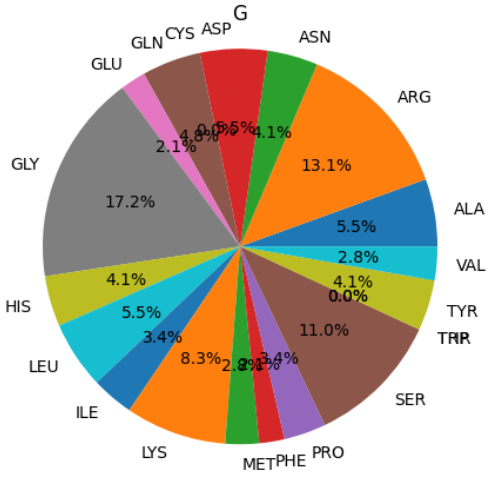
Fonte: Dados da pesquisa (2023)

Gráfico 25 — Percentual da frequência de aminoácidos próximos a Uracila.



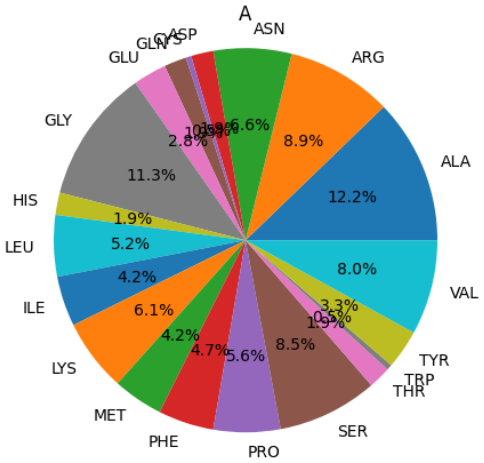
Fonte: Dados da pesquisa (2023)

Gráfico 26 — Percentual da frequência de aminoácidos próximos a Guanina.



Fonte: Dados da pesquisa (2023)

Gráfico 27 — Percentual da frequência de aminoácidos próximos a Adenina.



Fonte: Dados da pesquisa (2023)

4.8.1 Metodologia de Análise Espacial Inversa

Para explorar as interações recíprocas, o enfoque da análise foi invertido, centrando agora nos nucleotídeos como base e investigando a frequência de cada um dos 20 aminoácidos nas suas proximidades. O procedimento consistiu em traçar um raio de 5 angstroms ao redor de cada nucleotídeo e verificar a presença de cada aminoácido nesse perímetro.

4.8.2 Destaques na Frequência de Aminoácidos

Os resultados desta etapa revelaram padrões interessantes na preferência de aminoácidos em proximidade aos nucleotídeos. A glicina (GLY) destacou-se como o aminoácido mais presente em três casos: citosina (30.1%), uracila (25%), e guanina (17.2%). Adicionalmente, na adenina foi notada uma presença significativa de 11.3%, perdendo apenas para a arginina (ARG), que apresentou 12.2%.

4.9 Implicações e Considerações

A elevada presença de glicina sugere sua propensão para interações com diversas bases nitrogenadas, indicando uma possível flexibilidade conformacional desse aminoácido em interações proteína-RNA. A adenina, embora não seja o nucleotídeo mais frequente,

demonstra uma presença notável, destacando-se como um potencial participante nessas interações.

5. CONSIDERAÇÕES FINAIS

O presente trabalho teve como objetivo central a construção de um Banco de Dados de Interações Proteína-RNA, utilizando dados provenientes do PDB e aplicando análises detalhadas por meio do ambiente Google Colab com Python e Pandas. Ao longo do desenvolvimento deste projeto, diversas etapas metodológicas foram realizadas para garantir a qualidade e relevância das informações obtidas.

A criação desse banco de dados e as análises realizadas não representam apenas um avanço na compreensão da bioinformática estrutural, mas também abrem caminho para aplicações inovadoras na pesquisa acadêmica e biomédica. O conhecimento aprofundado nesse nicho tem o potencial de influenciar significativamente diagnósticos, tratamentos e prevenção de doenças. Dessa forma, incentiva-se pesquisadores a explorar esse banco de dados e aprofundar as análises aqui iniciadas para desvendar novas nuances nas interações proteína-RNA.

Em suma, os direcionamentos futuros devem se concentrar não apenas na expansão do banco de dados, mas também na sofisticação das análises realizadas. A integração de abordagens preditivas, a consideração de padrões específicos e a exploração de técnicas avançadas podem levar a descobertas mais profundas e aplicações mais precisas no campo da bioinformática estrutural. Essas iniciativas visam continuar a contribuir significativamente para a compreensão das complexas interações entre proteínas e RNA.

Além dos resultados apresentados, este trabalho abre portas para diversas oportunidades de pesquisa e aprimoramento. Considerando o potencial do Banco de Dados de Interações Proteína-RNA construído, destacam-se algumas áreas específicas para direcionamentos futuros: Avaliação de modelos preditivos, como o Alphafold (JUMPER *et al*., 2021), para geração de interfaces de interações proteína-RNA. Investigar se essas interfaces são consistentes com estruturas nativas ou reais; Incorporação de métodos de previsão de ligação específicos para RNA, a fim de tentar identificar possíveis sítios para interação com proteínas; Explorar a existência de padrões específicos de sítios de ligação para diferentes tipos de RNA, como transportador ou mensageiro; Integração contínua de dados estruturais, visando manter o banco de dados atualizado com novas estruturas obtidas por técnicas avançadas de resolução.

REFERÊNCIAS BIBLIOGRÁFICAS

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/Estudosdebiologia/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.

JÚNIOR, Adenivaldo Lima Filgueira *et al*. OS BANCOS DE DADOS DE INFORMAÇÃO BIOLÓGICA E SUA POTENCIAL APLICABILIDADE ÀS CIÊNCIAS MÉDICAS: UMA REVISÃO. **Visão Acadêmica**, v. 23, n. 1, 2022.

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.
GONÇALVES, Luciana Aparecida. A suplementação de leucina com relação à massa muscular em humanos. **Revista brasileira de nutrição esportiva**, v. 7, n. 40, p. 3, 2013.

JUMPER, John; EVANS, Rhett; PRITZEL, Alexander; *et al*. Highly accurate protein structure prediction with AlphaFold. **Nature**, v. 596, n. 7873, p. 583–589, 2021. Disponível em: <https://www.nature.com/articles/s41586-021-03819-2>. Acesso em: 4 dez. 2023.

ANEXOS E APÊNDICES

APÊNDICE A - Drive com estruturas coletadas para o banco de dados.

https://drive.google.com/drive/folders/1uYS2UIxQaLGzkbtwWFD4OmUqR6c4I8Is?usp=sharing

APÊNDICE B - Repositório com código das análises.

https://github.com/EnnesJp/Monografia-Sistemas-de-Informacao