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 e Análise de Estruturas de Interações Proteína-RNA Modeladas por Métodos Computacionais

Belo Horizonte - MG
2023

JOÃO PEDRO BRAGA ENNES

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

Projeto de Monografia em Sistemas de Informação II 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 interação entre proteínas e RNA é fundamental para a compreensão dos processos biológicos essenciais, incluindo a regulação da expressão gênica e a manutenção da integridade do genoma. No entanto, a disponibilidade de um banco de dados público contendo informações sobre essas interações ainda é limitada. Este estudo visa preencher essa lacuna, construindo um banco de dados dedicado às interações proteína-RNA, utilizando métodos computacionais avançados para modelar e analisar as estruturas dessas interações. Na primeira fase do projeto, foi criado um banco de dados contendo informações detalhadas sobre sequências e estruturas de interações proteína-RNA, coletadas do Protein Data Bank (PDB). Este banco de dados oferece um recurso valioso para a pesquisa em bioinformática, fornecendo dados estruturais essenciais para a compreensão da redes regulatórias que governam a vida celular. Na segunda fase, foram realizadas comparações entre as estruturas coletadas do PDB e um novo conjunto de estruturas geradas por métodos computacionais, especificamente o AlphaFold 3. Este modelo computacional permitiu a geração de novas estruturas com base em sequências de nucleotídeos e aminoácidos previamente coletadas. A análise comparativa dessas estruturas visou identificar padrões e implicações biológicas importantes. Os resultados destacam as diferenças na composição estrutural de proteínas e RNAs, e na frequência de aminoácidos próximos a nucleotídeos. As implicações biológicas desses achados foram discutidas, enfatizando a importância de tais interações na funcionalidade celular. Em conclusão, este estudo contribui significativamente para a bioinformática estrutural, fornecendo um banco de dados robusto e análises detalhadas que facilitam a compreensão das interações proteína-RNA.

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

ABSTRACT

The interaction between proteins and RNA is fundamental for understanding essential biological processes, including the regulation of gene expression and the maintenance of genome integrity. However, the availability of a public database containing information about these interactions is still limited. This study aims to fill this gap by constructing a dedicated database for protein-RNA interactions, using advanced computational methods to model and analyze the structures of these interactions. In the first phase of the project, a database was created containing detailed information on sequences and structures of protein-RNA interactions collected from the Protein Data Bank (PDB). This database provides a valuable resource for bioinformatics research, offering essential structural data for understanding the regulatory networks that govern cellular life. In the second phase, comparisons were made between structures collected from the PDB and a new set of structures generated by computational methods, specifically AlphaFold 3. This computational model allowed the generation of new structures based on previously collected nucleotide and amino acid sequences. The comparative analysis of these structures aimed to identify important biological patterns and implications. The results highlight differences in the structural composition of proteins and RNAs and the frequency of amino acids near nucleotides. The biological implications of these findings were discussed, emphasizing the importance of such interactions in cellular functionality. In conclusion, this study significantly contributes to structural bioinformatics by providing a robust database and detailed analyses that facilitate the understanding of protein-RNA interactions.

Keywords: Bioinformatics, Interaction, Protein, RNA, Database, AlphaFold, pDB.

SUMÁRIO

1. INTRODUÇÃO....3
2. REFERENCIAL TEÓRICO....5
3. METODOLOGIA....6
3.1 Atualização do banco de dados....6
3.2 Modelagem de estruturas....6
3.3 Armazenamento dos modelos....7
3.4 Análise das estruturas geradas....7
3.4.1 Ambiente de desenvolvimento....7
3.4.2 Pré-processamento de dados....8
3.4.3 Análise exploratória dos dados....8
3.4.4 Análise espacial....8
3.4.5 Visualização de dados....8
3.5 Visualização e Interpretação dos Resultados....9
3.6 Limitações do Estudo....9
4. ANÁLISE E DISCUSSÃO DOS RESULTADOS....9
4.1 Características gerais dos bancos de dados....10
4.2 Composição de aminoácidos nas proteínas....10
4.4 Composição de Nucleotídeos nos RNAs....11
4.6 Análise de Relações entre Aminoácidos e Nucleotídeos....12
4.6.1 Destaques na Frequência de Nucleotídeos....16
4.6.2 Casos Específicos....18
4.8 Análise Recíproca: Frequência de Aminoácidos em Proximidade de Nucleotídeos....18
5. CONSIDERAÇÕES FINAIS....20
ANEXOS E APÊNDICES....23

1. INTRODUÇÃO

A interação entre proteínas e RNA é essencial para a compreensão dos processos biológicos que sustentam a vida. Essas interações desempenham papéis cruciais em várias funções celulares, como a regulação da expressão gênica e a manutenção da integridade do genoma (PIRES *et al*., 2013). A compreensão detalhada dessas interações têm implicações significativas para a pesquisa biomédica e o avanço do conhecimento na área biológica. No entanto, a disponibilidade de um banco de dados público abrangente em bioinformática, contendo 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.

Tendo isso em vista, na primeira etapa desse projeto, foi construído um banco de dados voltado especificamente para o armazenamento de informações relacionadas às sequências e estruturas envolvidas nas interações proteína-RNA, com o objetivo de reunir e disponibilizar esses dados. Nesse contexto, foram realizadas coletas de informações a partir do *Protein Data Bank* (PDB) (BERMAN *et al*., 2002), a fim de agregar dados de estruturas moleculares previamente sequenciadas e construir um recurso valioso para a pesquisa em bioinformática. Este banco de dados se apresenta como um recurso de relevância para a comunidade científica, permitindo o acesso a informações estruturais cruciais para a compreensão das interações proteína-RNA.

Esta segunda etapa da monografia busca, portanto, realizar comparações entre as estruturas previamente coletadas do PDB e o novo banco de estruturas geradas com auxílio da utilização de métodos computacionais. Nesse contexto, para construção do novo banco, foram utilizadas as sequências de nucleotídeos e aminoácidos referentes às estruturas reunidas na primeira metade do projeto, para modelar novas estruturas com o *AlphaFold* 3 (ABRAMSON *et al*., 2024), a fim de reunir dados gerados computacionalmente, de estruturas moleculares já sequenciadas de maneira experimental. Tal combinação possibilita a realização de análises valiosas para o avanço da pesquisa na área de bioinformática e das técnicas de modelagem estrutural.

Este projeto demonstra que a criação de um banco de dados de interações proteína-RNA é um passo crucial para a compreensão da bioinformática estrutural e uma ferramenta poderosa para impulsionar descobertas científicas e aplicações clínicas inovadoras. Além disso, a união destes grandes repositórios de informação com a utilização de novas técnicas e métodos computacionais de predição de estruturas, que permitem uma análise muito mais rápida desses dados, podem potencializar e acelerar exponencialmente os ganhos proporcionados por essa área de estudo. Tudo isso faz com que o aprofundamento do conhecimento nesse campo tenha o potencial de revolucionar a pesquisa em biomoléculas e oferecer avanços significativos para a prática médica e biomédica, direcionando novas abordagens para diagnósticos, tratamentos e prevenção de doenças .

2. REFERENCIAL TEÓRICO

A interação entre proteínas e RNA desempenha um papel fundamental nos processos biológicos, afetando a expressão gênica, a estabilidade do RNA e a regulação de diversas funções celulares (PIRES *et al*., 2013). As proteínas se ligam às moléculas de RNA para executar funções específicas, muitas vezes reconhecendo sequências estruturais ou sequências particulares de RNA. A análise dos dados moleculares provenientes do sequenciamento de DNA e RNA é crucial para entender as características estruturais dos segmentos gênicos e de seus produtos protéicos, além de elucidar suas interações e compreender processos biológicos como tradução, processamento de RNA e regulação pós-transcricional (JÚNIOR *et al*., 2022).

Apesar de sua importância, a investigação do sequenciamento estrutural das interações proteína-RNA ainda é pouco explorada. No PDB, um banco de dados essencial para a biologia estrutural, que contém estruturas tridimensionais de proteínas, existem mais de duzentas mil estruturas experimentais registradas. No entanto, com uma rápida filtragem, é possível notar que apenas cerca de oito mil estão relacionadas a uma estrutura de RNA.

Os bancos de dados biológicos, como o PDB, são depósitos de informações biológicas que incluem sequências de proteínas, RNA, estruturas tridimensionais, dados de expressão gênica e interações biomoleculares. Esses bancos são fundamentais para a pesquisa biomolecular, fornecendo acesso a informações valiosas para análises e descobertas (MARIANO *et al*., 2015). Na primeira fase deste projeto, foi construído um banco de dados específico para interações proteína-RNA, que permitiu a realização de análises de interações moleculares e comparação de estruturas nativas, obtidas de forma experimental, e estruturas

geradas com auxílio de ferramentas de modelagem, que utilizam métodos computacionais de predição.

Na área de bioinformática, os avanços da predição computacional vêm permitindo a obtenção de informações estruturais com base na sequência de aminoácidos de uma proteína cuja estrutura ainda não foi determinada experimentalmente (SILVA *et al*., 2021). Anteriormente, essa forma de predição era considerada um desafio, contudo, devido ao progresso dos algoritmos computacionais ao longo dos anos e à disponibilidade crescente de estruturas protéicas conhecidas, tornou-se viável, resultando em previsões plausíveis e com alto nível de precisão em muitos casos. Entretanto, é importante reconhecer que os métodos de predição estrutural computacional apresentam limitações que requerem uma avaliação cuidadosa para determinar a confiabilidade dos modelos gerados. Portanto, é de suma importância realizar uma avaliação desses modelos preditivos, a fim de estimar o nível de confiança a ser atribuído às estruturas preditas.

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 análise de estruturas modeladas por métodos computacionais. As principais fases incluíram a atualização do banco de dados construído na primeira etapa do projeto com estruturas do PDB (Banco A), a modelagem das estruturas com auxílio do *AlphaFold* 3, hospedado no *AlphaFold Server*, a criação de um novo banco para armazenar os modelos gerados (Banco B) e a análise subsequente desses dados usando *Google Colab* com Python e as bibliotecas Pandas, MatplotLib e Numpy.

3.1 Atualização do banco de dados

A partir do *script* em Python construído na primeira etapa deste estudo, foi feita uma atualização do Banco A, visando incluir novas estruturas cadastradas no PDB no intervalo entre as etapas do projeto. A escolha dos dados seguiu os mesmos critérios de seleção definidos na primeira parte da monografia, escolhidos 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.

3.2 Modelagem de estruturas

Para a modelagem das estruturas de interações proteína-RNA, inicialmente havia sido escolhido o *Rosetta3* (LEAVER-FAY *et al*., 2011), uma ferramenta de previsão e design de estrutura de proteínas. Porém com o decorrer do projeto e após o lançamento, no dia oito de maio, da nova versão do *Alphafold*, optamos por alterar a ferramenta. Essa decisão foi tomada devido, principalmente, à disponibilização do *Alphafold Server* que permitiu que o processo de modelagem fosse realizado em nuvem, o que reduziu o custo computacional da máquina utilizada para gerar as novas estruturas.

Portanto, para a etapa de modelagem, foi utilizado o *AlphaFold* 3, um algoritmo de inteligência artificial desenvolvido pela DeepMind, para prever a estrutura tridimensional de proteínas com base em suas sequências de aminoácidos, e que agora em sua terceira versão, permite também a predição de estruturas tridimensionais de DNA e RNA, com base em sua sequência de nucleotídeos. A modelagem foi realizada no *AlphaFold Server*, que consiste em uma interface *web* que permite a geração de complexos moleculares em nuvem, sem que sejam necessários altos recursos computacionais por parte do usuário ou qualquer experiência em aprendizado de máquina.

As sequências de proteínas e RNAs foram submetidas para prever as estruturas de suas interações respeitando o limite diário de 20 *jobs* por usuário, estipulado pela plataforma, e também o limite máximo de 5000 *tokens* por *job*.

3.3 Armazenamento dos modelos

As estruturas tridimensionais geradas pelo *AlphaFold* 3 foram armazenadas em um novo banco de dados, organizado de forma a facilitar a recuperação e análise posterior. Este banco de dados inclui metadados relevantes, como identificadores das sequências, localização espacial dos átomos, condições experimentais e qualidade da previsão estrutural.

3.4 Análise das estruturas geradas

3.4.1 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, Seaborn e Matplotlib, proporcionando um ambiente computacional rico para análises detalhadas.

3.4.2 Pré-processamento de dados

Os dados armazenados no Banco B foram carregados no ambiente do *Google Colab* utilizando a biblioteca Pandas. Assim como havia sido feito com o Banco A, foi realizado um pré-processamento para limpeza dos dados, para que fossem removidas todas as informações que não fossem relevantes para as análises das estruturas, presentes nos arquivos.

3.4.3 Análise exploratória dos dados

A análise exploratória dos dados foi realizada utilizando as bibliotecas NumPy e Pandas, que permitiram 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, para posterior comparação com as estruturas coletadas no PDB.

3.4.4 Análise espacial

Para realizar a análise espacial das proteínas, 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.

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.

3.4.5 Visualização de dados

Utilizando a biblioteca Matplotlib e Seaborn, foram gerados gráficos para visualizar as distribuições e relações entre diferentes variáveis. Gráficos de dispersão, histogramas e heatmaps foram utilizados para identificar tendências e outliers.

3.5 Visualização e Interpretação dos Resultados

Os resultados das análises feitas nos bancos de dados obtidos através do PDB e dos modelos do *AlphaFold*, foram visualizados utilizando gráficos para melhor interpretação e comparação. Comparação de distâncias entre átomos e número de interações por aminoácidos e por nucleotídeos, foram apresentadas para destacar as diferenças e similaridades entre as duas fontes de dados.

3.6 Limitações do Estudo

É importante ressaltar que este estudo tem algumas limitações, incluindo a dependência da qualidade e representatividade dos dados disponíveis no PDB, assim como a qualidade dos modelos gerados pelo *AlphaFold Server* que ainda está em estágio Beta. Além disso, as análises realizadas estão intrinsecamente ligadas à qualidade das estruturas biológicas depositadas, modeladas e disponíveis para pesquisa.

4. ANÁLISE E DISCUSSÃO DOS RESULTADOS

A partir da atualização dos dados, o Banco A, que inicialmente era composto por 309 estruturas de interações proteína-RNA geradas de forma experimental, teve um aumento de aproximadamente 5% e conta agora com 325 estruturas. Após o processo de validação e

controle de qualidade, o Banco A foi refinado para 175 estruturas que atendiam aos critérios estabelecidos, ou seja, um aumento de cinco estruturas em relação à primeira etapa da pesquisa. Apesar desse aumento, as novas estruturas não geraram impacto relevante nas análises feitas previamente.

Devido a algumas limitações do *AlphaFold Server*, 20 sequências do nosso Banco A de 175 não foram modeladas. Dentre elas, 15 possuíam apenas três nucleotídeos, o que impossibilitava o processo por não atender o número mínimo de quatro nucleotídeos, exigido pela plataforma. Além disso, uma das sequências possuía mais de 5000 *tokens*, o que também impossibilitou a modelagem, por ultrapassar o limite máximo permitido pela ferramenta. Por fim, quatro outras sequências, apesar de estarem dentro dos limites da plataforma, falharam durante a modelagem e por isso também não puderam ser incluídas no estudo.

Portanto, ao final do processo de modelagem com *AlphaFold* 3, cada sequência de entrada gerou cinco novos modelos de estruturas, fazendo com que o Banco B contasse com 755 estruturas, representando 155 sequências distintas.

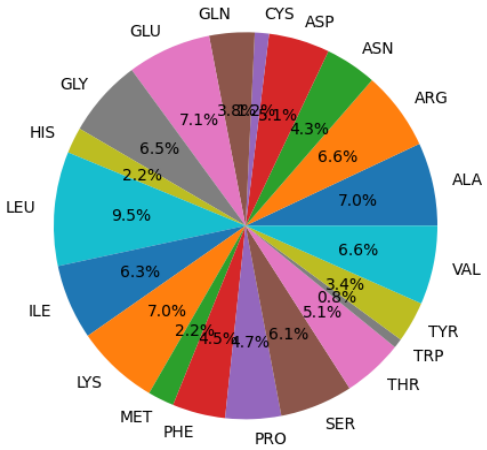
4.1 Características gerais dos bancos de dados

Para que as comparações fossem feitas de maneira a evitar ao máximo qualquer tipo de enviesamento dos dados, o Banco A foi reduzido às mesmas 155 sequências modeladas pelo *AlpahFold*. A consistência dessas estruturas foi fundamental para garantir a qualidade das análises subsequentes.

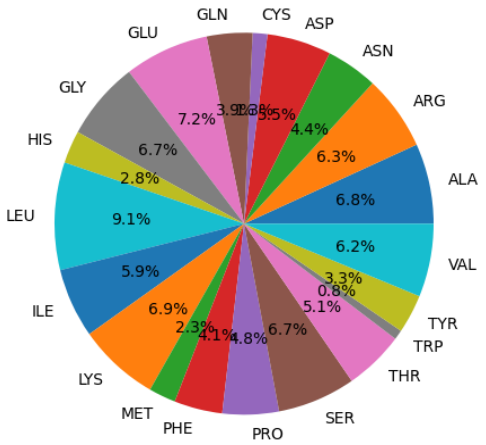
4.2 Composição de aminoácidos nas proteínas

Gráfico 1 — Percentual da frequência de cada aminoácido no Banco A.

Gráfico 2 — Percentual da frequência de cada aminoácido no Banco B.



Fonte: Dados da pesquisa (2024)



Fonte: Dados da pesquisa (2024)

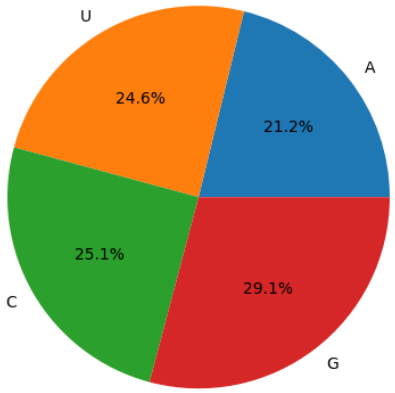
Em uma primeira análise rasa, é possível notar que houve uma pequena diferença na frequência de aminoácidos entre os bancos de estruturas nativas e modeladas computacionalmente. Contudo, com o aprofundamento do estudo, foi identificado que isso se deve ao fato de que, cerca de 67 arquivos recuperados do PDB, não possuem a sequência da proteína totalmente resolvida.

Durante o processo de determinação experimental da estrutura de uma proteína, pode ocorrer que apenas uma parte da sequência seja resolvida com clareza, enquanto outras regiões permaneçam indeterminadas. Isso pode ser devido a várias razões, como a flexibilidade intrínseca de certas regiões da proteína, dificuldades em cristalização, ou limitações técnicas nas técnicas de imagem, como a cristalografia de raios-X ou a ressonância magnética nuclear (RMN). Essas limitações podem impedir a obtenção de uma estrutura completa e precisa de toda a sequência da proteína, restringindo a análise e a compreensão total de sua função e interações moleculares.

Entretanto, a análise do perfil de aminoácidos revelou que a distribuição se manteve equilibrada no geral. Além disso, outros padrões encontrados na primeira etapa da pesquisa, ainda se mantiveram, como a leucina ainda se destacando com a maior porcentagem de aparição, tendo no Banco A 9,5% e no Banco B e 9,1%. Da mesma forma, o triptofano continuou sendo o aminoácido com menor presença, com um percentual de 0.8% em ambos os bancos.

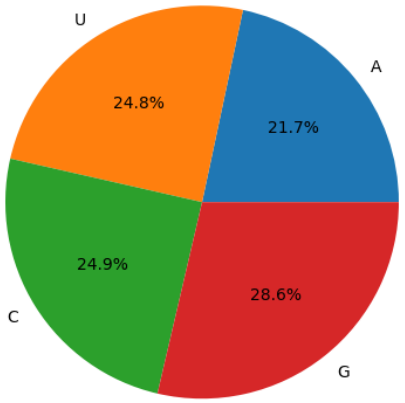
4.4 Composição de Nucleotídeos nos RNAs

Gráfico 3 — Percentual da frequência de cada nucleotídeo no Banco A.



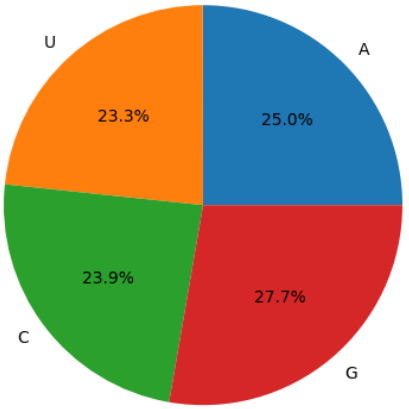
Fonte: Dados da pesquisa (2024)

Gráfico 4 — Percentual da frequência de cada nucleotídeo no Banco B.



Fonte: Dados da pesquisa (2024)

Gráfico 5 — Percentual da frequência de cada nucleotídeo no Banco A.



Fonte: Dados da pesquisa (2023)

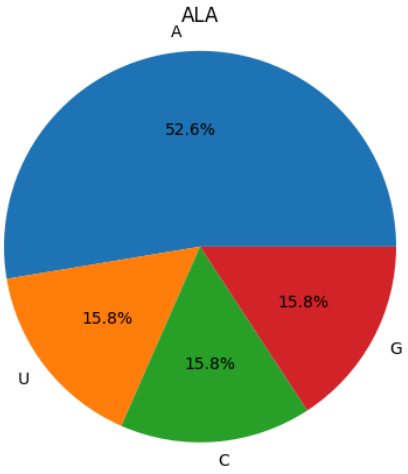
Assim como aconteceu com a frequência de aminoácidos, nos nucleotídeos também foi possível notar uma pequena divergência entre os dados dos bancos A e B. Essa diferença também se deve ao fato de que, devido a algumas limitações encontradas nos métodos experimentais de resolução de estruturas, determinadas sequências não puderam ser totalmente resolvidas com clareza, o que não acontece nos métodos computacionais.

Apesar das limitações supracitadas, entre os nucleotídeos, a guanina manteve o destaque com a maior presença, representando 29.1% do total no Banco A e 28,6% no Banco

B. Esse achado sugere uma prevalência dessa base nitrogenada nas interações estudadas, podendo indicar regiões específicas de ligação ou estabilidade nas estruturas formadas. Diferente do que apontava o primeiro estudo realizado, onde a uracila apresentou a menor presença, totalizando 23.3%, agora a adenina se encontra com o menor percentual, apresentando 21,2% no Banco A e 21,7% no Banco B.

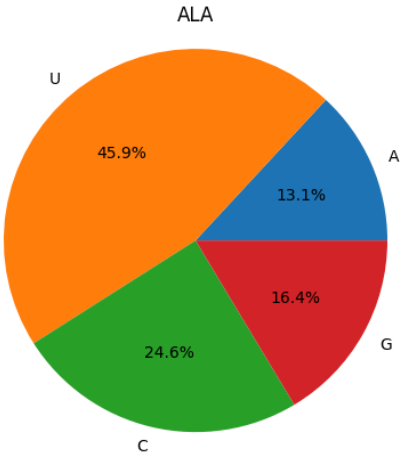
4.6 Análise de Relações entre Aminoácidos e Nucleotídeos

Gráfico 6 — Percentual da frequência de nucleotídeos próximos a Alanina Banco A.



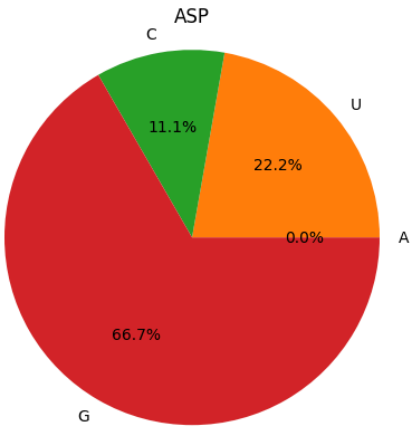
Fonte: Dados da pesquisa (2024)

Gráfico 7 — Percentual da frequência de nucleotídeos próximos a Alanina Banco B.



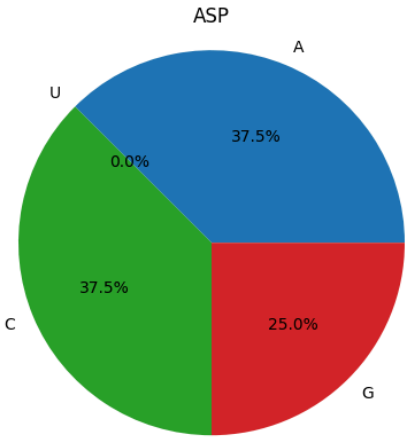
Fonte: Dados da pesquisa (2024)

Gráfico 8 — Percentual da frequência de nucleotídeos próximos ao Ácido aspártico Banco A.



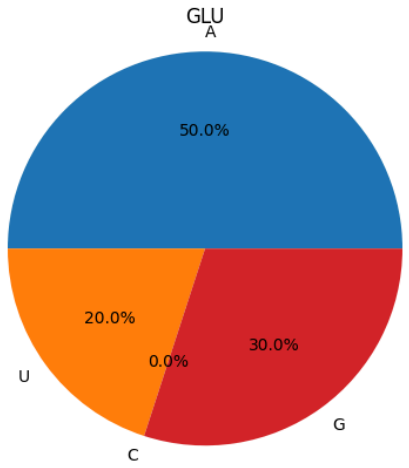
Fonte: Dados da pesquisa (2024)

Gráfico 9 — Percentual da frequência de nucleotídeos próximos ao Ácido aspártico Banco B.



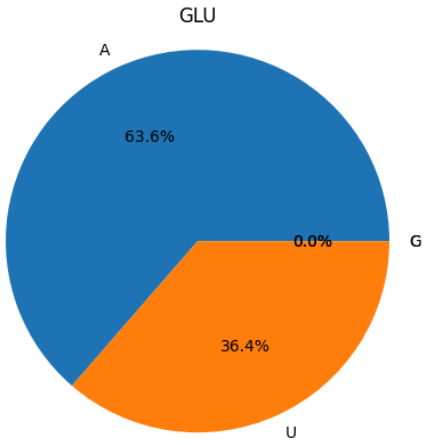
Fonte: Dados da pesquisa (2024)

Gráfico 10 — Percentual da frequência de nucleotídeos próximos ao Ácido glutâmico Banco A.



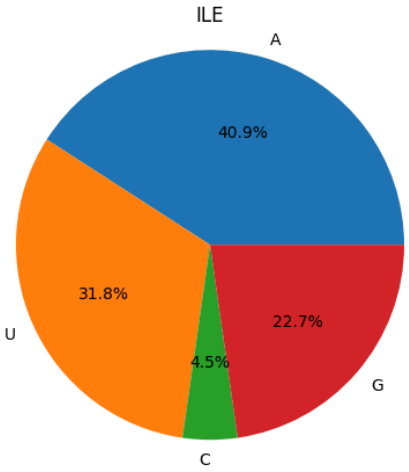
Fonte: Dados da pesquisa (2024)

Gráfico 11 — Percentual da frequência de nucleotídeos próximos ao Ácido glutâmico Banco B.



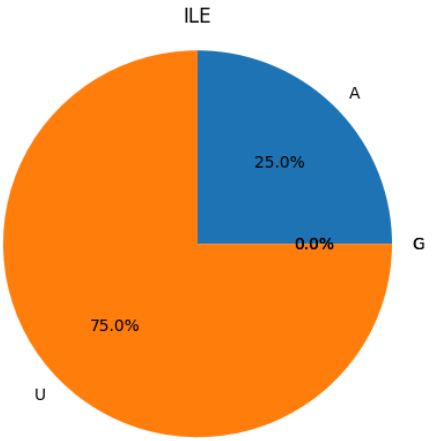
Fonte: Dados da pesquisa (2024)

Gráfico 12 — Percentual da frequência de nucleotídeos próximos ao Isoleucina Banco A.



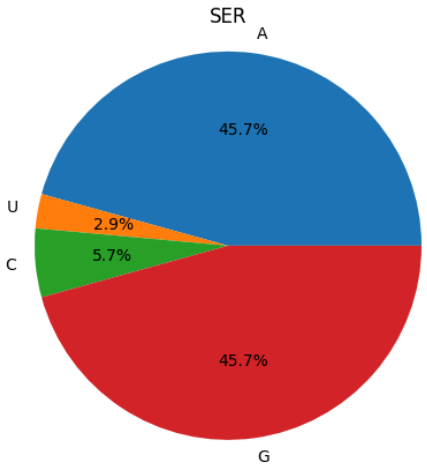
Fonte: Dados da pesquisa (2024)

Gráfico 13 — Percentual da frequência de nucleotídeos próximos ao Isoleucina Banco B.



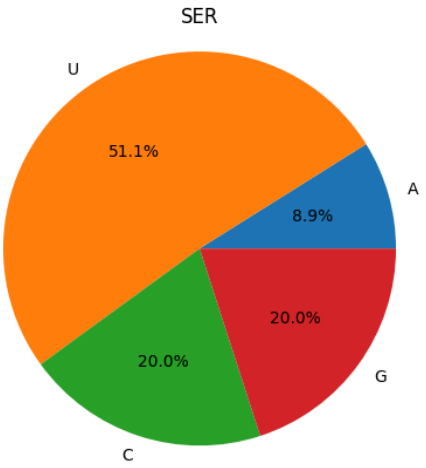
Fonte: Dados da pesquisa (2024)

Gráfico 14 — Percentual da frequência de nucleotídeos próximos ao Serina Banco A.



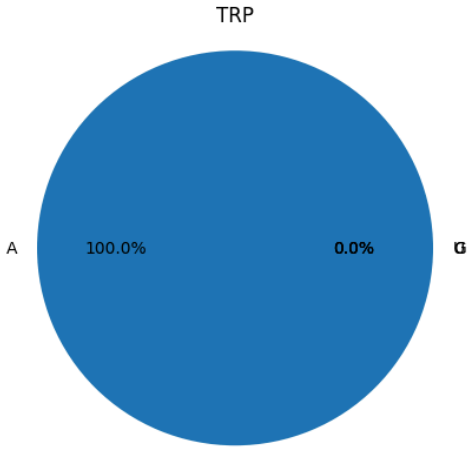
Fonte: Dados da pesquisa (2024)

Gráfico 15 — Percentual da frequência de nucleotídeos próximos ao Serina Banco B.



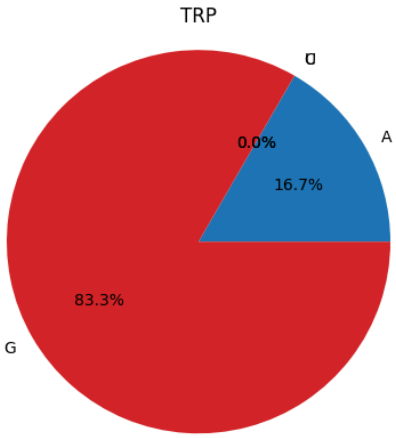
Fonte: Dados da pesquisa (2024)

Gráfico 16 — Percentual da frequência de nucleotídeos próximos ao Triptofano Banco A.



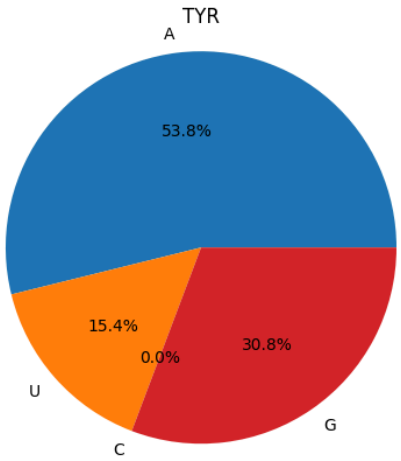
Fonte: Dados da pesquisa (2024)

Gráfico 17 — Percentual da frequência de nucleotídeos próximos ao Triptofano Banco B.



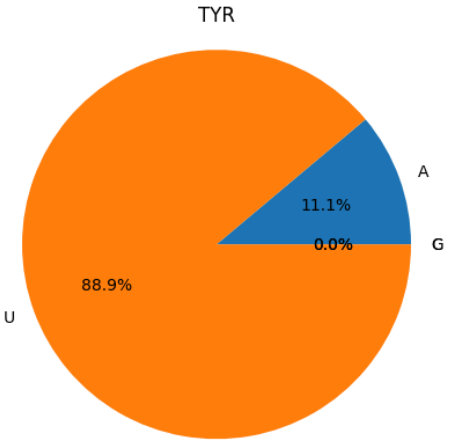
Fonte: Dados da pesquisa (2024)

Gráfico 18 — Percentual da frequência de nucleotídeos próximos ao Tirosina Banco A.



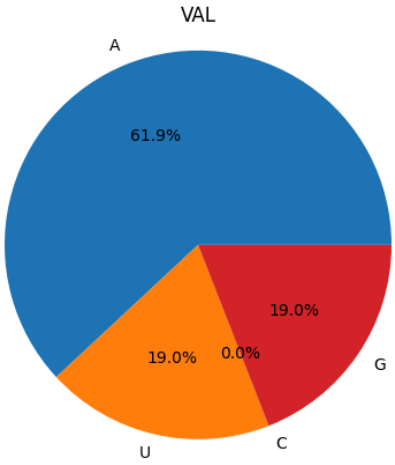
Fonte: Dados da pesquisa (2024)

Gráfico 19 — Percentual da frequência de nucleotídeos próximos ao Tirosina Banco B.



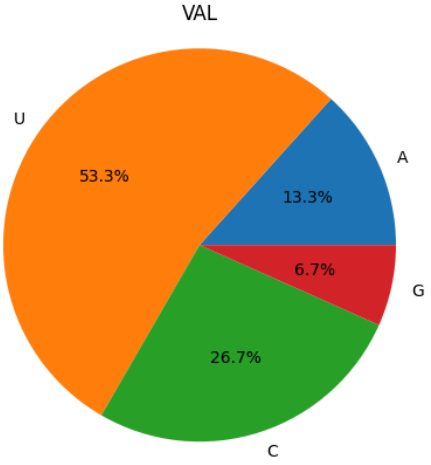
Fonte: Dados da pesquisa (2024)

Gráfico 20 — Percentual da frequência de nucleotídeos próximos ao Valina Banco A.



Fonte: Dados da pesquisa (2024)

Gráfico 21 — Percentual da frequência de nucleotídeos próximos ao Valina Banco B.



Fonte: Dados da pesquisa (2024)

4.6.1 Destaques na Frequência de Nucleotídeos

Foram observadas diferenças notáveis ao analisar a frequência de nucleotídeos nas proximidades dos aminoácidos e compará-las entre os dois bancos. No Banco A, em 12 dos 20 aminoácidos estudados, destacou-se uma maior presença de adenina comparada aos outros nucleotídeos. De forma oposta, no Banco B, a adenina teve predominância em apenas cinco dos aminoácidos, perdendo o seu posto para a uracila, que apareceu em 8 dos 20 casos, e foi o nucleotídeo mais frequente nas proximidades dos aminoácidos.

No Banco A, a adenina (A) se destaca como a base nitrogenada mais prevalente em várias ocasiões. Especificamente, ela é predominante nas proximidades dos aminoácidos alanina, asparagina, ácido glutâmico, isoleucina, leucina, lisina, metionina, fenilalanina, prolina, triptofano, tirosina e valina. Por outro lado, a guanina (G) e a uracila (U) apresentaram frequências semelhantes em três aminoácidos distintos: ácido aspártico, glutamina e histidina. Além disso, observou-se que em nenhum caso a citosina (C) foi a base nitrogenada predominante na vizinhança dos aminoácidos estudados. Esse perfil sugere uma clara predominância da adenina e uma ausência notável da citosina como base majoritária em torno dos aminoácidos no Banco A.

No Banco B, a uracila (U) emergiu como a base nitrogenada mais proeminente, sendo predominante em oito aminoácidos: alanina, arginina, glicina, isoleucina, lisina, serina, triptofano e valina. Em contraste, a adenina (A) teve uma queda significativa em sua predominância, aparecendo em apenas seis casos: asparagina, ácido glutâmico, metionina, fenilalanina e treonina. A guanina (G) manteve um padrão semelhante ao observado no Banco A, porém com uma leve elevação no número de predominâncias, aparecendo em diferentes aminoácidos como leucina, glicina, histidina e triptofano. Notavelmente, a citosina (C) foi a menos frequente entre todos os nucleotídeos, com a prolina sendo o único aminoácido em que a citosina teve essa predominância, de acordo com os dados gerados por modelos computacionais. Esses achados indicam um papel mais diversificado das bases nitrogenadas no Banco B em comparação com o Banco A, com uma supressão na predominância notável da adenina e se mantendo um raro destaque da citosina.

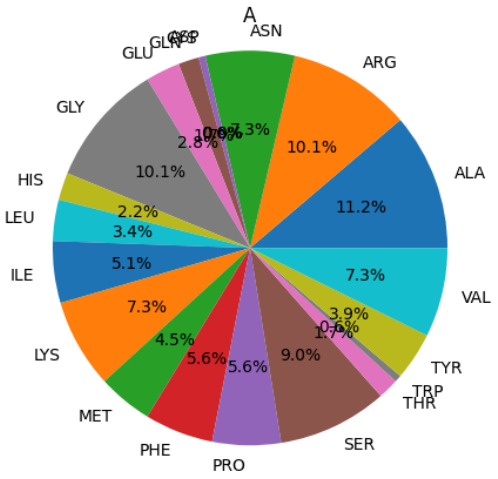
A observação de diferenças notáveis na frequência de nucleotídeos nas proximidades dos aminoácidos entre os dois bancos de dados pode ser atribuída a diversas hipóteses. No Banco A, a predominância de certos nucleotídeos nas proximidades dos aminoácidos estudados pode refletir as condições biológicas e experimentais específicas sob as quais essas estruturas foram resolvidas. Esses contextos podem incluir fatores como a conformação natural das biomoléculas, interações estabilizadas por condições fisiológicas ou preferências específicas de ligação em ambientes celulares reais. Em contraste, as diferentes frequências apresentadas no Banco B, se devem possivelmente às diferenças nos algoritmos preditivos e nas bases de dados de treinamento utilizados pelo *AlphaFold*. Esses modelos computacionais podem enfatizar diferentes aspectos das interações proteína-RNA, levando a variações na previsão das interações nucleotídicas em relação aos dados experimentais. Além disso, as simplificações e suposições inerentes aos métodos computacionais podem introduzir vieses que se manifestam nas frequências de nucleotídeos observadas.

4.6.2 Casos Específicos

No caso da Cisteína, apesar de serem identificadas algumas poucas ocorrências de proximidades a alguns nucleotídeos no Banco A, não foi possível identificar nenhuma aparição no Banco B e, por conta disso, não foi atribuída a nenhum dos casos citados. Essa observação pode sugerir peculiaridades na interação desses aminoácidos com os nucleotídeos nas estruturas analisadas. Além disso, outros aminoácidos que tiveram empates entre as suas maiores presenças, também não foram atribuídos a nenhum dos grupos.

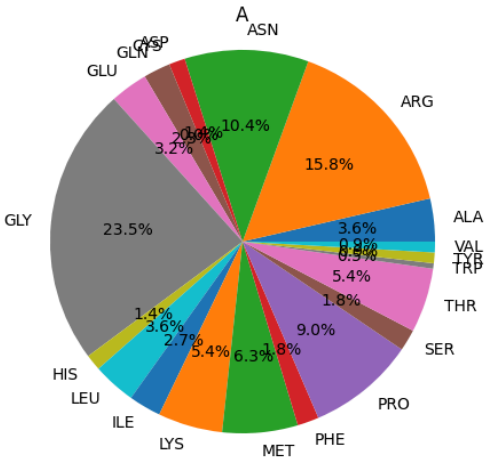
4.8 Análise Recíproca: Frequência de Aminoácidos em Proximidade de Nucleotídeos

Gráfico 22 — Percentual da frequência de aminoácidos próximos a Adenina no Banco A.



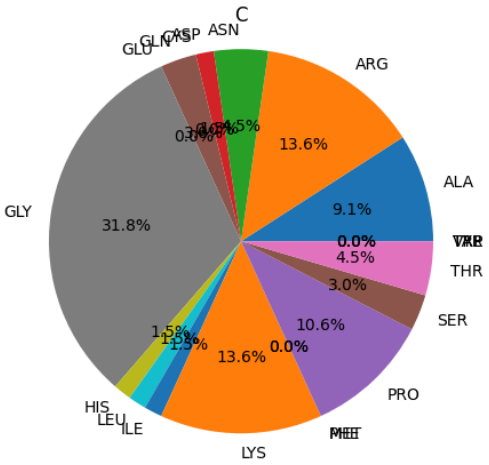
Fonte: Dados da pesquisa (2024)

Gráfico 23 — Percentual da frequência de aminoácidos próximos a Adenina no Banco B.



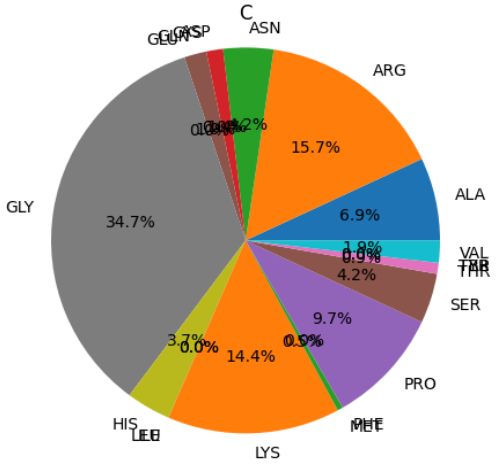
Fonte: Dados da pesquisa (2024)

Gráfico 24 — Percentual da frequência de aminoácidos próximos a Citosina no Banco A.



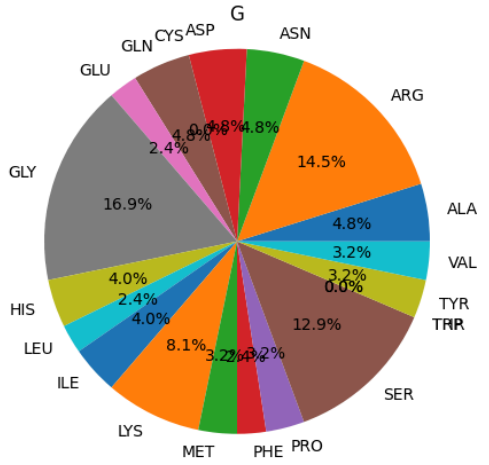
Fonte: Dados da pesquisa (2024)

Gráfico 25 — Percentual da frequência de aminoácidos próximos a Citosina no Banco B.



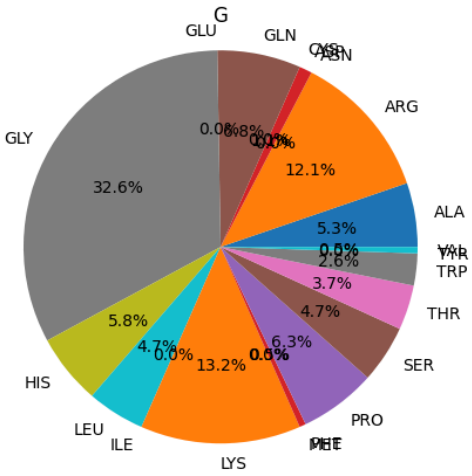
Fonte: Dados da pesquisa (2024)

Gráfico 26 — Percentual da frequência de aminoácidos próximos a Guanina no Banco A.



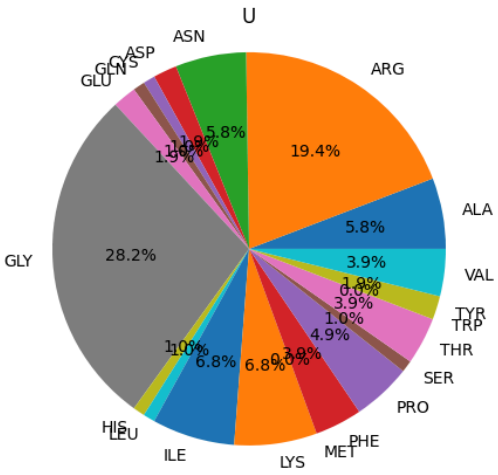
Fonte: Dados da pesquisa (2024)

Gráfico 27 — Percentual da frequência de aminoácidos próximos a Guanina no Banco B.



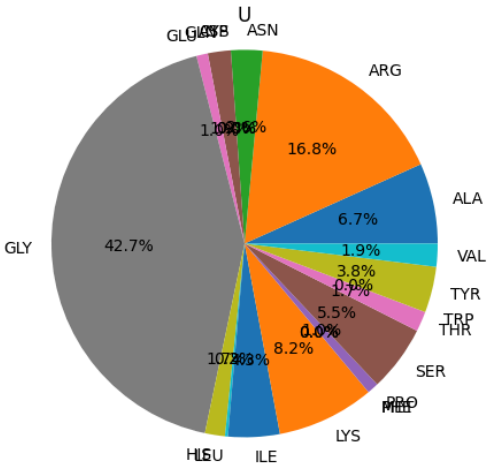
Fonte: Dados da pesquisa (2024)

Gráfico 28 — Percentual da frequência de aminoácidos próximos a Uracila no Banco A.



Fonte: Dados da pesquisa (2024)

Gráfico 29 — Percentual da frequência de aminoácidos próximos a Uracila no Banco B.



Fonte: Dados da pesquisa (2024)

Apesar das diferenças apresentadas entre os bancos, não houve uma discrepância tão grande na análise recíproca das frequências, provavelmente devido ao alto número de aminoácidos e baixo número de nucleotídeos, o que faz com que uma mudança pequena não gere tanto impacto na porcentagem final. A glicina (GLY), que no Banco A se destacou como o aminoácido mais presente na citosina, uracila e guanina, com 31,8%, 28,2% e 16.9%, respectivamente, no Banco B, se mostra ainda mais predominante e inclui a adenina também a lista com 23,5% de presença.

No Banco A a frequência elevada de glicina pode refletir preferências naturais de interação determinadas por contextos celulares específicos e condições experimentais. A glicina, sendo um aminoácido pequeno e flexível, pode se acomodar facilmente em diferentes contextos estruturais, favorecendo a formação de interações estáveis com esses nucleotídeos. No Banco B a maior predominância de glicina, incluindo a adenina, pode ser resultado das características dos algoritmos preditivos do *AlphaFold* que podem amplificar tendências observadas nos dados de treinamento. Além disso, as simplificações e generalizações inerentes aos modelos computacionais podem levar a uma superestimação da presença de glicina em contextos onde sua flexibilidade é favorecida, incluindo novas interações que não são tão frequentemente observadas em condições experimentais reais.

5. CONSIDERAÇÕES FINAIS

A presente monografia destacou a importância das interações proteína-RNA na biologia celular e a necessidade de recursos bioinformáticos robustos para aprofundar a compreensão dessas interações. A criação de um banco de dados específico para interações proteína-RNA, aliada ao uso de métodos computacionais avançados como o *AlphaFold* 3, demonstrou ser uma abordagem eficaz para aumentar o volume e a precisão dos dados disponíveis para pesquisa.

O estudo revelou que, apesar das limitações dos métodos experimentais na resolução de estruturas, a modelagem computacional pode oferecer uma alternativa poderosa e complementar. A comparação entre estruturas nativas do PDB e estruturas modeladas pelo *AlphaFold* 3 permitiu identificar padrões e diferenças significativas, enriquecendo o entendimento das características estruturais das interações proteína-RNA.

Os resultados indicaram que, enquanto a frequência de aminoácidos e nucleotídeos variou entre os bancos de dados nativos e modelados, a distribuição geral se manteve equilibrada, com pequenas divergências atribuídas às limitações experimentais e às capacidades dos métodos computacionais. Além disso, a prevalência de certos nucleotídeos em proximidade a aminoácidos específicos aponta para possíveis regiões de ligação ou estabilidade estrutural que merecem investigação adicional.

Este trabalho também ressaltou a importância da avaliação cuidadosa dos modelos preditivos computacionais para garantir a confiabilidade das estruturas geradas. A utilização de ferramentas como o *Google Colab*, com bibliotecas avançadas de análise de dados, mostrou-se eficiente para a análise e visualização dos resultados, facilitando a interpretação e comparação dos dados estruturais.

Em suma, a integração de dados experimentais e computacionais em bioinformática estrutural não apenas amplia o conhecimento sobre interações proteína-RNA, mas também impulsiona a pesquisa biomédica, oferecendo novas perspectivas para o diagnóstico, tratamento e prevenção de doenças. Juntamente a isso, o avanço contínuo em técnicas de modelagem estrutural e a criação de bancos de dados mais abrangentes são essenciais para o progresso científico nesta área. Este estudo contribui para essa evolução, demonstrando o potencial transformador da bioinformática na compreensão e aplicação dos processos biológicos fundamentais.

REFERÊNCIAS BIBLIOGRÁFICAS

ABRAMSON, Josh; ADLER, Jonas; DUNGER, Jack; *et al*. *Accurate structure prediction of biomolecular interactions with AlphaFold* 3. ***Nature*** 630, 493–500 (2024). Disponível em: <https://www.nature.com/articles/s41586-024-07487-w>. Acesso em: 23 jul. 2024.

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.

CHENG, Clarence Yu; CHOU, Fang-Chieh; DAS, Rhiju. *Modeling complex RNA tertiary folds with Rosetta*. In: ***Methods in enzymology***. Academic Press, 2015. p. 35-64.

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.

LEAVER-FAY, Andrew *et al*. *ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules*. In: ***Methods in enzymology***. Academic Press, 2011. p. 545-574.

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.

SILVA, Letícia X; BASTOS, Luana L; SANTOS, Lucianna H. Modelagem computacional de proteínas. **BIOINFO-Revista Brasileira de Bioinformática e Biologia Computacional**, v. 1, n. 8, 2021. Disponível em: <https://bioinfo.com.br/wp-content/uploads/2021/07/08_Modelagem-computacional-de-proteinas-_-BIOINFO_compressed.pdf>. Acesso em: 05 jan. 2024.

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.

ANEXOS E APÊNDICES

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

https://drive.google.com/drive/folders/1shknv6kS9OCCHmLarlCkdBItUkWqVfSV?usp=sharing

APÊNDICE B - Repositório com código das análises.

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