UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
CENTRO DE DESENVOLVIMENTO SUSTENTÁVEL DO SEMIÁRIDO
UNIDADE ACADÊMICA DE ENGENHARIA DE BIOTECNOLOGIA E BIOPROCESSOS
CURSO DE ENGENHARIA DE BIOTECNOLOGIA E BIOPROCESSOS

ALISSON CLEMENTINO DA SILVA

AVALIAÇÃO DO USO DE INTELIGÊNCIA ARTIFICIAL APLICADA À MINERAÇÃO DE DADOS TERMODINÂMICOS DE PROTEÍNAS

SUMÉ - PB

2023

ALISSON CLEMENTINO DA SILVA

AVALIAÇÃO DO USO DE INTELIGÊNCIA ARTIFICIAL APLICADA À MINERAÇÃO DE DADOS TERMODINÂMICOS DE PROTEÍNAS

Monografia apresentada ao Curso de Engenharia de Biotecnologia e Bioprocessos do Centro de Desenvolvimento Sustentável do Semiárido da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Engenharia de Biotecnologia e Bioprocessos.

Orientador: Professor Dr. Bruno Rafael Pereira Nunes.

Professora Dra. Joycimara Santos Xavier.

SUMÉ - PB

2023

S586a Silva, Alisson Clementino da.
Avaliação do uso de inteligência artificial aplicada à mineração de dados termodinâmicos de proteínas. / Alisson Clementino da Silva. - 2023.

55 f.

Orientador: Prof. Dr. Bruno Rafael Pereira Nunes. Co-orientadora. Profa. Dra. Joycimara Santos Xavier.
Monografia - Universidade Federal de Campina Grande; Centro de Desenvolvimento Sustentável do Semiárido; Curso de Engenharia de Biotecnologia e Bioprocessos.

1. Mineração de dados. 2. Inteligência artificial. 3. CAprendizado de máquina. 4. Termodinâmica de proteínas – dados. 5. Curadoria de dados. 6. ThermoMutDB – base de dados. I. Nunes, Bruno Rafael Pereira. II. Xavier, Joycimara Santos. Título.

CDU: 60(043.1)

Elaboração da Ficha Catalográfica:

Johnny Rodrigues Barbosa
Bibliotecário-Documentalista
CRB-15/626

ALISSON CLEMENTINO DA SILVA

AVALIAÇÃO DO USO DE INTELIGÊNCIA ARTIFICIAL APLICADA À MINERAÇÃO DE DADOS TERMODINÂMICOS DE PROTEÍNAS

Monografia apresentada ao Curso de Engenharia de Biotecnologia e Bioprocessos do Centro de Desenvolvimento Sustentável do Semiárido da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Engenharia de Biotecnologia e Bioprocessos.

BANCA EXAMINADORA:

Professor Dr. Bruno Rafael Pereira Nunes
Orientador – UATEC/CDSA/UFCG

Professora Dra. Joycimara Santos Xavier
Co-orientadora – UFVJM

Professora Dra. Thaís Gaudêncio do Rego
Examinadora Externa I – DINF/CI/UFPB

Mestra Bruna Moreira
Examinadora Externa II – University of Melbourne

Professor Dr. Rafael Trindade Maia
Examinador Interno – UAEDUC/CDSA/UFCG

Trabalho aprovado em: 10 de julho de 2023.

SUMÉ - PB

“Não que eu seja tanta coisa

Sou um grão de areia na imensidão

Mas cabe quase o mundo inteiro no meu peito

Carrego todas as memórias

Todos os sabores que daqui provei

Levo comigo os abraços que ganhei

Mas se tiver que definir em uma só palavra

Resumir a minha história numa só canção

Se dessa vida eu levasse um só nome

Ele é ***Cristo.****”*

Tudo o que eu vivi - Vocal Livre

Grandes são as obras do SENHOR, dignas de estudo para quem as ama. Salmos 111.2

Aos meus pais Osman e Elmaísa, por todo suporte, amor e cuidado em todos os momentos da minha vida até aqui. Por todos os pastéis e sanduíches vendidos debaixo de sol e chuva para me sustentar longe de casa.

Ao meu irmão Alex, a quem desejo servir de exemplo e cuidar até o fim da minha vida. À Atiliane, por todo amor, confiança e suporte durante esse tempo tão importante para mim.

Aos meus avós Luis (*in memoriam)* e Dona Nita*, Orlando (in memoriam)* e Mundinha. Por todos os conselhos, todo amor e apoio incondicional, de duas mulheres que me inspiram. Também as minha tias e tios que diretamente ou indiretamente contribuíram nessa conquista.

Aos meus amigos mais íntimos; Miqueias, Juliana, Lilian, Isabela, jefter e Ana Paula por tudo quanto fizeram e suportaram junto a mim.

A juventude Aceviana de cidade de Itaporanga - PB e Juventudes batista das cidades Sumé, Monteiro e Caraúbas - Pb, por todo amor fraternal e cumplicidade.

Ao Prof. Dr. Bruno Rafael Pereira Nunes, por toda disponibilidade, companheirismo e dedicação.

A Profª. Drª. Joicymara Xavier, por proporcionar toda a experiência vivenciada e aos colegas do Laboratório de Bioinformática e Inteligência Artificial.

A Profª. Drª. Glauciane Danusa Coelho, por todo ensino, amor, cuidado, café e acolhimento.

Aos colegas do Laboratório de Microbiologia do CDSA, em especial, ao Mestre Cícero Anthonyelson.

Agradeço também aos demais servidores do CDSA e toda a equipe da UAEB pelas contribuições dadas e pelo tempo cedido.

SOU GRATO.

RESUMO

O alinhamento da biotecnologia moderna com a bioinformática tem fornecido importantes informações para a descoberta e o desenvolvimento de novos fármacos. A realização de estudos de mutagênese, a partir de abordagens computacionais, tem tentado prever os efeitos de mutações *missense* em proteínas, que estão relacionadas a doenças graves, por meio de suas estruturas tridimensionais. Para tal feito, preditores computacionais de estabilidade, que avaliam os efeitos da mutação, precisam de um grande volume de dados termodinâmicos para serem capazes de predizer os efeitos estruturais causados à proteína. Um dos problemas recorrentes é a falta de estruturação e padronização dos dados utilizados, que demanda muito tempo de trabalho humano para solucionar. Sendo assim, a utilização de inteligência artificial torna possível a mineração e gerenciamento de dados em menor tempo, auxiliando o processo de design de novos fármacos. Esta pesquisa apresenta o treinamento de um modelo de machine learning, na plataforma LitSuggest, para recuperação de referências que contenham dados termodinâmicos de proteínas, depositadas no repositório PubMed. Um total de 14 referências foram classificadas pelo modelo e selecionadas em curadoria manual, totalizando 283 novas mutações e 2.901 novos dados adicionados no ThermoMutDB.

Palavras-chave: Mutações *missense*; Aprendizado de máquina; Curadoria de dados; ThermoMutDB

SILVA, Alisson Clementino da. **Evaluation of the use of artificial intelligence applied to protein thermodynamic data mining.** 2023. 55f. Trabalho de Conclusão de Curso (Monografia), Curso de Engenharia de Biotecnologia e Bioprocessos, Centro de Desenvolvimento Sustentável do Semiárido, Universidade Federal de Campina Grande - Sumé - Paraíba - Brasil, 2023.

ABSTRACT

The alignment of modern biotechnology with bioinformatics has provided important information for the discovery and development of new drugs. Mutagenesis studies from computational approaches have attempted to predict the effects of missense mutations on proteins that are related to serious diseases through their three-dimensional structures. For this purpose, computational stability predictors which evaluate the effects of mutation need a large volume of thermodynamic data to be able to predict the structural effects caused to the protein. One of the recurring problems is the lack of structuring and standardization of the data used, which takes a lot of human time to solve. Thus, the use of artificial intelligence makes data mining and management possible in less time, assisting in the design process of new drugs. This research presents the training of a machine learning model, on the LitSuggest platform, to retrieve references containing thermodynamic data of proteins, deposited in the PubMed repository. A total of 14 references were classified by the model and selected in manual curation, totaling 283 new mutations and 2,901 new data added to ThermoMutDB.

Keywords: Missense mutations; Machine learning; Data curation; ThermoMutDB

LISTA DE FIGURAS

LISTA DE ABREVIATURAS E SIGLAS

mRNA - fita de ácido ribonucleico mensageiro

pré-mRNA - fita de ácido ribonucleico mensageiro com regiões não codificantes

RNA - ácido ribonucleico

tRNA - ácido ribonucléico transportador

DNA - ácido desoxirribonucleico

DSC - calorimetria de varredura diferencial

RMN - ressonância magnética nuclear

CD - dicroísmo circular

ML - machine learning

SUMÁRIO

1 INTRODUÇÃO

A biotecnologia tem sido crucial nos últimos anos em diversas áreas, seu caráter multidisciplinar que envolve a aplicação de variadas formas de vida (microrganismos, plantas e animais), para obtenção de processos e produtos de interesse para a sociedade, tem permitido inúmeras conexões e impulsionado avanços na medicina, na agricultura, na indústria e no cuidado com o meio ambiente. Ao longo do tempo, novas ferramentas e aplicações foram incorporadas ao seu escopo, a partir do avanços concomitantes em genética, microbiologia, química, fisiologia, biologia molecular, bioquímica e entre outras, sendo a biotecnologia moderna o resultado de suas convergências (GUIDO; ANDRICOPULO; OLIVA, 2010)

A bioinformática é uma área também multidisciplinar, que envolve computação, biologia, física, química e outras. O seu alinhamento com a biotecnologia tem promovido consideráveis avanços técnico-científicos e novos desafios ainda mais complexos. Um exemplo das aplicações biotecnológicas da bioinformática é a descoberta e desenvolvimento de novos fármacos produzidos a partir de microorganismos, que se destacam em uma subárea denominada de biologia sintética, que permite a montagem de genomas inteiros, com uma infraestrutura mínima (VERLI, 2014)

Nos últimos anos, a biotecnologia tem aumentado a sua participação no desenvolvimento e produção de fármacos, já representando cerca de 10 a 15% do mercado farmacêutico, revolucionando a investigação e o desenvolvimento de novos medicamentos, bem como a fabricação destes em escala industrial. O processo que começa com a triagem virtual e identificação da proteína que deve ser regulada, passa por *docking* molecular e tem seus parâmetros de interação, energia e posicionamento atômico otimizados em softwares específicos (GUIDO; ANDRICOPULO; OLIVA, 2010; VERLI, 2014).

Os estudos de mutagênese em proteínas fornecem muitas informações sobre os efeitos de mutações pontuais que podem alterar a estabilidade de uma proteína e orientar as iniciativas de design de medicamentos farmacêuticos que visam combater os efeitos das doenças. Geralmente, as informações são descritas em protocolos e artigos científicos depositados em repositórios, apresentados em tabelas e gráficos plotados a partir de ensaios experimentais. Dessa forma, quando surge a necessidade de acessar esses dados, como suporte para desenvolvimento de novas drogas, ou para que os preditores computacionais de estabilidade, que avaliam se uma mutação é susceptível de estabilizar ou desestabilizar uma

proteína a partir do consumo de dados como estes, um problema é evidenciado (FREITAS, 2020; TORRES-FREIRE; GOLGHER; CALLIL, 2014).

Apesar da maioria das referências que descrevem mutações e dados termodinâmicos estarem depositadas em um mesmo repositório, o grande volume de referências similares dificulta o trabalho dos pesquisadores que precisam somente dos dados padronizados e estruturados. Sendo assim, o desenvolvimento de espaços como a base de dados ThermoMutDB, que é dedicada a curadoria e gerenciamento de dados termodinâmicos de proteínas, se faz necessário. Porém, um outro problema encontrado é o tempo necessário para busca e mineração dos textos que contém os dados de interesse, que normalmente custam meses com base no trabalho humano (ALLOT et al., 2021; DOMINGUES, 2003; XAVIER et al., 2021).

Diante do exposto, a mineração de textos e dados, surge como uma alternativa que torna possível a automatização da recuperação de literatura que tenha os dados de interesse, a partir da criação de algoritmos e técnicas computacionais, que sejam capazes de conectar um conjunto de entradas a um resultado específico (DOMINGUES, 2003; FREITAS, 2020). Nessa pesquisa, a ferramenta LitSuggest, que é uma plataforma que permite o treinamento de modelos de aprendizado de máquina para recuperação de referências do repositório PubMed, será testada como via de recuperação de referências que contenham dados termodinâmicos de proteínas, para serem submetidos a curadoria manual e disponibilizado no ThermoMutDB.

2 OBJETIVOS

2.1 OBJETIVO GERAL

Avaliar a utilização da ferramenta LitSuggest como potencial via de recuperação de literatura contendo dados termodinâmicos de proteínas.

2.2 OBJETIVOS ESPECÍFICOS

- Validação de um modelo para recuperação de dados termodinâmicos;
- Análise da viabilidade de automatização total da recuperação de literatura para curadoria manual;
- Submeter novas referências ao ThermoMutDB.

3 REFERENCIAL TEÓRICO

3.1 PROTEÍNAS E PROTEOMA

A bioquímica moderna é o resultado mais expressivo do alinhamento entre a genética, a química e a física. Os conhecimentos gerados a partir do estudo da natureza molecular do material genético e da caracterização do processo de transmissão da informação genética tornaram-se base para o que é denominado "Dogma Central da Biologia Molecular". Foi a partir dos anos 50 que a colaboração entre a biologia molecular e bioquímica se intensificou, culminando em importantes descobertas que foram fundamentais para criar hipóteses sobre os meios de expressão gênica e elucidar a síntese de proteínas que ocorre nas células, processo este que se configurou como um dos grandes desafios da biologia molecular (RIBEIRO, 2014; NELSON, COX, 2014).

O dogma, termo este que é mantido por razões históricas, se refere a uma descrição concatenada do fluxo bidirecional de informações genéticas, no qual sua efetivação ocorre sempre de ácidos nucléicos para proteínas. O DNA codifica para a produção de ácido ribonucleico no processo de transcrição e o RNA codifica para a produção de proteína no processo de tradução, sendo estas o resultado expresso das informações traduzidas. Dessa forma, a especificidade das funções biológicas, por muitos anos, passaram a ser atribuídas às proteínas por serem abundantes e mediarem todo tipo de reação bioquímica. Contudo, a desconstrução desse entendimento moldou o fim do século XIX e começo do século XX com profundas mudanças nas bases da genética (SOLHA, 2005; NELSON; COX, 2014)

3.2 DE MENDEL À WATSON E CRICK

As ideias fundamentais da genética começaram a ser compiladas teoricamente por volta de 1865, a partir dos estudos de Johann Gregor Mendel (1822-1884), que analisou experimentalmente o desenvolvimento de híbridos a partir do cruzamento de ervilhas e objetivou verificar a forma que as características visíveis eram transmitidas e diferenciadas dos genitores após algumas gerações de descendentes. Para isso, escolheu ervilhas do gênero *Pisum* que são facilmente cultivadas, possuem um curto ciclo reprodutivo e apresentam muita produtividade. Assim, Mendel postulou a existência de fatores herdados, caracteres dominantes ou recessivos, que são determinantes para o aparecimento de alguns traços fenotípicos observados nas plantas. Hoje sabemos que tais fatores eram os genes (MARTINS, 2002; OLIVEIRA; SANTOS; BELTRAMINI, 2004)

Entrementes, estudos citológicos publicados por Friedrich Miescher (1844-1895) em 1869, caracterizaram a composição química de uma substância isolada do núcleo de glóbulos brancos que foi denominada de nucleína, sendo esta, determinante para o entendimento do

núcleo celular como componente essencial das células, para caracterização dos processos de mitose e meiose e reconhecimento das estruturas cromossômicas (OLIVEIRA; SANTOS; BELTRAMINI, 2004). Assim, evidências de que um gene daria origem a uma proteína começaram surgir, baseados nessa pesquisa e associados aos escritos de Mendel, que pouco mais de três décadas após sua publicação foram "redescobertos" e reconhecidos experimentalmente em vários organismos (RIBEIRO, 2014). Desta forma, no fim do século XIX, surge a teoria cromossômica da herança, fruto da união entre citologia e genética, que disserta sobre o comportamento dos cromossomos durante os processos de divisão celular com base em previsões mendelianas, uma vez que essas estruturas começaram a ser apontadas como detentoras dos caracteres herdados (SOLHA, 2005).

Por conseguinte, pesquisas de diferentes campos da biologia foram destaques na busca pelo entendimento da natureza dos fatores mendelianos, que acabaram sendo fortemente propagados em eventos científicos e introduzidos na Inglaterra pelo biólogo William Bateson (1861-1925) que se interessava pelo esclarecimento das análises de Mendel quanto aos fenômenos observáveis, mas também buscava entendimento sobre as causalidades da hereditariedade (MARTINS, 2002). Doravante, as hipóteses sobre a natureza física dos genes tornaram-se recorrentes, pois até então, não se sabia qual a composição de tal estrutura e não havia entendimento sobre seus efeitos, até que os bioquímicos George Beadle e Edward Tatum, em 1941, apresentaram uma conexão que relacionava a presença de enzimas com o genes, sendo eles seus "reguladores". Essa característica foi identificada por eles a partir de experimentos realizados com a levedura *Neurospora*, evidenciando que havia relação entre deficiências metabólicas e características genéticas, que também foi identificada em estudos anteriores, de Archibald Garrod, em 1902. A relação estabelecida entre as enzimas e genes ficou conhecida como a hipótese "um gene - uma enzima" (CECCATTO, 2010; SOLHA, 2005).

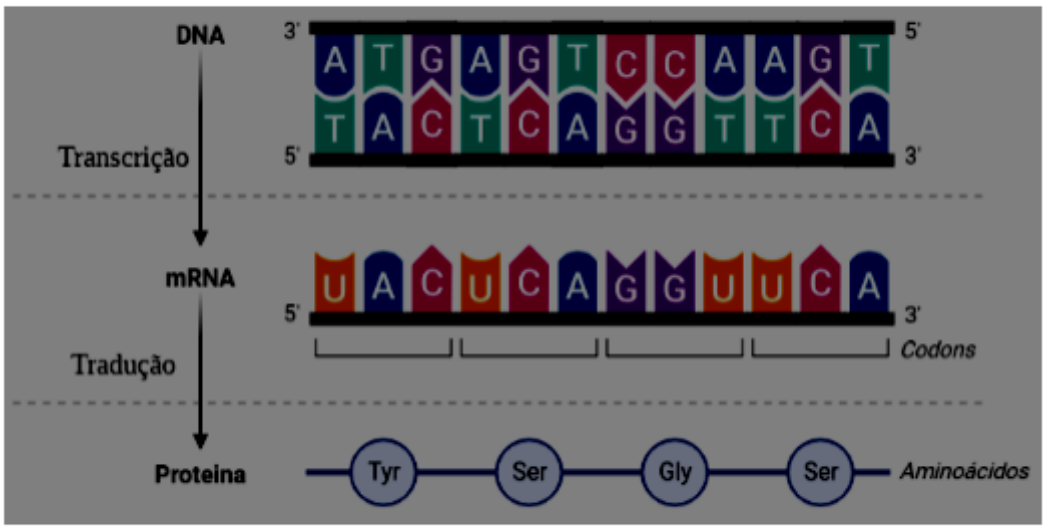
Após a repercussão dos estudos de Garrod e as contribuições de Beadle e Tatum, a ideia de que um gene é responsável pela produção de uma proteína, foi disseminada e explorada por outros pesquisadores, que tendo poucos esclarecimentos sobre a estrutura físico-química do material genético, fizeram delas moléculas intensamente estudadas e cogitadas como o possível suporte químico da hereditariedade, já que desempenham funções regulatórias e estruturais nos seres vivos (SOLHA, 2005; NELSON; COX, 2014). Contudo, a partir de 1944, importantes estudos que caracterizaram o princípio transformante nas células eucarióticas, começaram a teorizar que o DNA era responsável pelo armazenamento das informações herdadas e repasse das mesmas, não atribuindo tal ciência às proteínas, que eram apontadas como feitorias de tal ciência. Com isso, o esclarecimento da estrutura do ácido desoxirribonucleico tomou muitíssima importância e instigou uma rede colaborativa de grandes proporções que avançou em diversas frentes científicas ao longo dos anos para desvendar a natureza da transmissão das informações biológicas (THIEMAN, 2003;

RIBEIRO, 2014).

Foi então que em 1952, Alfred Hershey e Martha Chase trabalhando com isótopos radioativos, demonstraram que o componente viral utilizado como mecanismo para infectar as células humanas é o ácido nucleico (ALVES; SOUZA, 2013). Nesse período, as evidências reunidas já apontavam inegavelmente para o DNA como sendo a estrutura central da hereditariedade, mas a ideia contrariava a perspectiva dominante de que as proteínas desempenhavam o papel de armazenamento e transferência de informações, de modo que houve resistência dentro da comunidade acadêmica (ARAÚJO e MARTIN, 2008; THIEMAN, 2003). Com o alvorecer dessas concepções, admitiu-se que elucidar a estrutura molecular do ácido desoxirribonucleico traria o entendimento de sua função biológica e de como o fluxo de informações acontece no interior da célula.

Então, em 1953, por influência das metodologias de Linus Pauling, análise das medidas cristalográficas vindas principalmente de William Astbury e Rosalind Elsie Franklin (1920-1958) e pela concatenação de todo o conhecimento gerado até aquele momento por pesquisas tangenciais, os cientistas Francis Crick e James Watson em suas publicações, tornam-se reconhecidos pela elucidação da molécula de DNA e ainda propuseram um modelo para a estrutura físico-química da molécula, onde monômeros formariam cadeias longas unidas por uma pentose e um grupo fosfato com a presença de bases nitrogenadas que se ligam ao carbono 1’ da pentose e ainda contaria com ligações de hidrogênio, que daria sustento a forma helicoidal da fita dupla antiparalela, havendo ainda um pareamento restrito entre os nucleotídeos (A - T e C - G), anteriormente evidenciado por Erwin Chargaff (1905-2002) em seus estudos da composição do DNA que ficaram conhecidos como a regra de Chargaff (RIBEIRO, 2014; SOLHA, 2005; OLIVEIRA; SANTOS; BELTRAMINI, 2004; THIEMAN, 2003).

Figura 1 - Esquematização do Dogma Central da Biologia Molecular



Fonte: Autoria própria (Figura criada com o BioRender.com)

O modelo foi idealizado a partir da análise das imagens cristalográficas, dos demais dados obtidos das amostras de DNA e das metodologias utilizadas por Pauling. Ao analisar as variáveis desse processo, Watson e Crick chegaram a um arranjo molecular teórico que foi capaz de satisfazer o pareamento proposto e que correspondesse diretamente aos dados descritos por Wilkins e Franklin. A alta densidade molecular e a proporcionalidade das bases nitrogenadas, sugeriu que o ácido nucleico deveria ser composto de uma dupla fita anti paralela, que se mantém unida por ligações fosfodiéster entre nucleotídeos adjacentes e pela sua complementaridade que permite a formação de ligações de hidrogênio com o lado oposto, estabelecendo uma distância de 20 Angstrons entre fitas, conferindo à molécula resistência química e possibilitando a sua continuidade (RIBEIRO, 2014; CECCATTO, 2010; NELSON; COX, 2014). Dessa forma, a base físico-química da hereditariedade foi encontrada nos ácidos nucleicos e na sua particular conformação sequencial, o que justifica a diversidade de organismos existentes, uma vez que a reprodução e os processos evolutivos atrelados a ela, em termos gerais, depende do material genético que é o responsável por armazenar e transmitir as informações das gerações passadas (SOLHA, 2005; THIEMAN, 2003).

3.3 DO DNA À PROTEÍNAS

O dogma então se tornou importante para o entendimento das funções biológicas por apontar para os ácidos nucleicos, que são macromoléculas na base das informações genéticas. O DNA é formado por nucleotídeos, composto por um grupo fosfato ligado através de uma ligação fosfodiéster a uma pentose (desoxirribose no DNA e ribose no RNA) que, por sua vez, está ligada às bases nitrogenadas (purinas: adenina (A) e guanina (G) e pirimidinas: citosina (C) e timina (T)). A partir do DNA, temos a formação do RNA que é composto por uma fita simples de nucleotídeos que ao chegar no citosol participa de processos tradicionais codificando proteínas, que são resultantes da efetivação da informação biológica (NELSON; COX, 2014).

As proteínas são o produto da cadeia de reações que traduzem as informações biológicas. A palavra grega *protos*, que pode significar “a mais importante”, é a origem etimológica da palavra proteína e remete ao princípio de que essas grandes moléculas biológicas se tornam as mais presentes e atuantes em todos os seres vivos desempenhando diversas funções. São longos polímeros compostos por um esqueleto covalente, sustentado por centenas de ligações que permitem múltiplas angulações e conformações quase incontáveis. São essenciais à vida, constituem mais da metade da massa seca total de uma célula e sua síntese tem uma importância fundamental para a manutenção e crescimento celular, norteando praticamente todas as reações fisiológicas dos organismos vivos. (FOOD INGREDIENTS BRASIL Nº 28 - 2014; FROTA, et al, 2021; NELSON; COX, 2014).

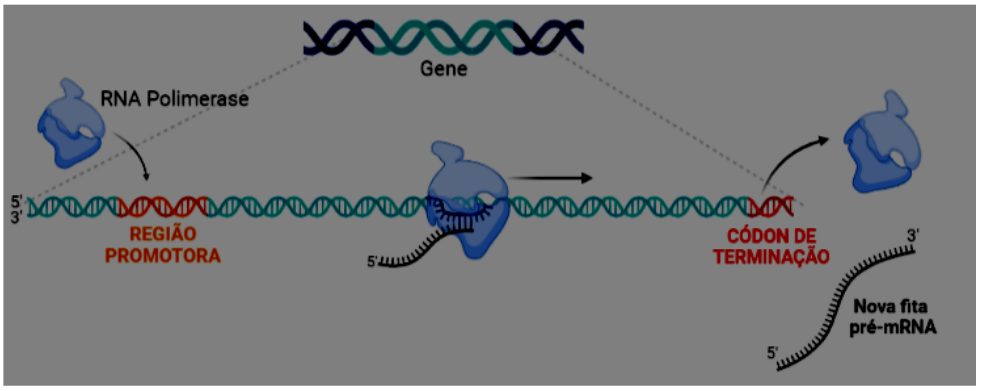
Os aminoácidos são os monômeros que formam uma proteína. Sendo seus menores componentes, eles apresentam uma região comum denominada de esqueleto peptídico, formado por radicais orgânicos que se ligam ao mesmo átomo de carbono (um átomo de hidrogênio, um grupo amina, um grupo carboxílico e a cadeia lateral). O genoma humano codifica um total de 22 aminoácidos, sendo 20 para proteínas, os quais se diferem uns dos outros a partir da cadeia lateral que lhes confere tamanho, cargas, solubilidade, forma e funcionalidade diferentes (ALBERTS et al; FRANCISCO e FRANCISCO, 2002; STRYER, 1996).

Para compor uma proteína, uma sequência de aminoácidos é formada a partir de uma sequência de nucleotídeos que se estabelece por uma interação química de dois aminoácidos a partir da ligação peptídica. A carboxila (COO–) de um aminoácido e o grupo amino (NH3+) do adjacente se ligam covalentemente formando um dipeptídeo e dessa forma, cadeias polipeptídicas que apresentam quantidades variadas de resíduos de aminoácidos podem ser formadas. Com a perda de uma molécula de água na formação da ligação peptídica, os aminoácidos não se apresentam da mesma forma quando livres, por isso são chamados de resíduos (ALBERTS, 2017).

3.4 SÍNTESE DE PROTEÍNAS

O processamento da informação biológica também é conhecido como síntese de proteínas, onde a sequência nucleotídica do ácido ribonucléico (RNA) é traduzida em um produto biológico funcional. Primeiramente, ocorre na célula a necessidade de uma proteína específica e neste momento o DNA, que não controla o processo de síntese, tem uma fração sua (um gene) transcrita. Essa etapa chamada de transcrição ocorre exatamente no núcleo e gera uma fita de mRNA (ácido ribonucléico mensageiro) complementar que é transportada até o citoplasma. Saindo do núcleo celular, uma segunda etapa chamada de tradução ocorre, onde ribossomos e tRNA (ácido ribonucléico transportador) associam-se a fita de RNAm para gerar uma nova proteína (NELSON; COX, 2014; BERG et al., 2017).

Figura 2 - Etapa de transcrição do DNA na síntese proteica



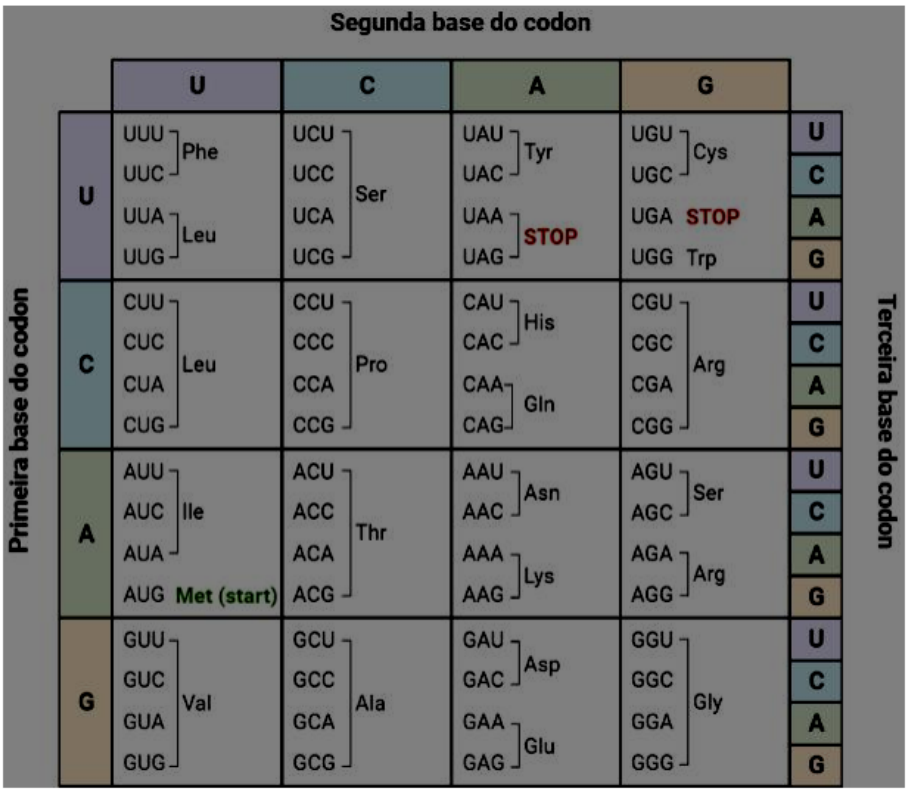
Fonte: Autoria própria, 2023 (Figura criada com o BioRender.com).

A transcrição (Figura 2), inicia-se com a enzima RNA polimerase ao DNA. Ela se liga à região que antecede o gene de interesse que é responsável por guardar a informação biológica que codifica a proteína desejada. Ao percorrer o DNA, essa região específica chamada de promotora, que é reconhecida pelo códon de iniciação A-U-G, indica o início do gene e nesse momento pela ação da enzima, a dupla fita é separada permitindo a utilização da fita de sentido 3'5' como molde para produção de um pré-mRNA (5'3'). Essa nova fita é sintetizada considerando a mesma lógica de pareamento de bases, mas adota um diferencial alterando as ligações Adenina - Timina por Adenina - Uracila. O processo ocorre até que um códon de terminação (U-A-A, U-A-G ou U-G-A) seja compreendido pela enzima, indicando o fim do gene e da etapa transcricional (ALBERTS, 2017; NELSON; COX, 2014).

O pré-mRNA permanece no núcleo para que seja submetido a um tipo de curadoria biológica que se chama *splicing*, onde porções transcritas e não codificantes são retiradas da sequência. Duas modificações importantes ocorrem na nova fita pela adição de grupos químicos em suas extremidades. A cap5', que é uma guanina modificada (G), é adicionada na extremidade 5' para impedir rompimentos por ação de fosfatases e nucleases e mais tarde mediar a ligação do complexo enzimático que codifica a proteína no citoplasma, enquanto uma sequência de aproximadamente 200 adeninas chamada de cauda poli-A, é adicionada na extremidade 3' para proteger o transcrito e conferir maior estabilidade à molécula. Após modificações em ambas as extremidades, partes da fita, chamadas de íntrons, são removidas por não serem segmentos codificantes e as partes restantes que codificam aminoácidos, chamadas de exons, são novamente unidas a fim de obter um mRNA maduro e funcional (ALBERTS, 2017; NELSON; COX, 2014).

Para a produção do pré-mRNA, uma base nitrogenada correspondia a uma outra base seguindo o pareamento diferencial. Para a composição de uma proteína, o maquinário celular considera três nucleotídeos (um códon) para que um aminoácido seja codificado. As relações entre códons e aminoácidos são chamadas de código genético (Figura 3) e resumem quais sequências são necessárias para que um aminoácido seja adicionado à cadeia polipeptídica (ALBERTS, 2017; NELSON; COX, 2014).

Figura 3 - Código genético



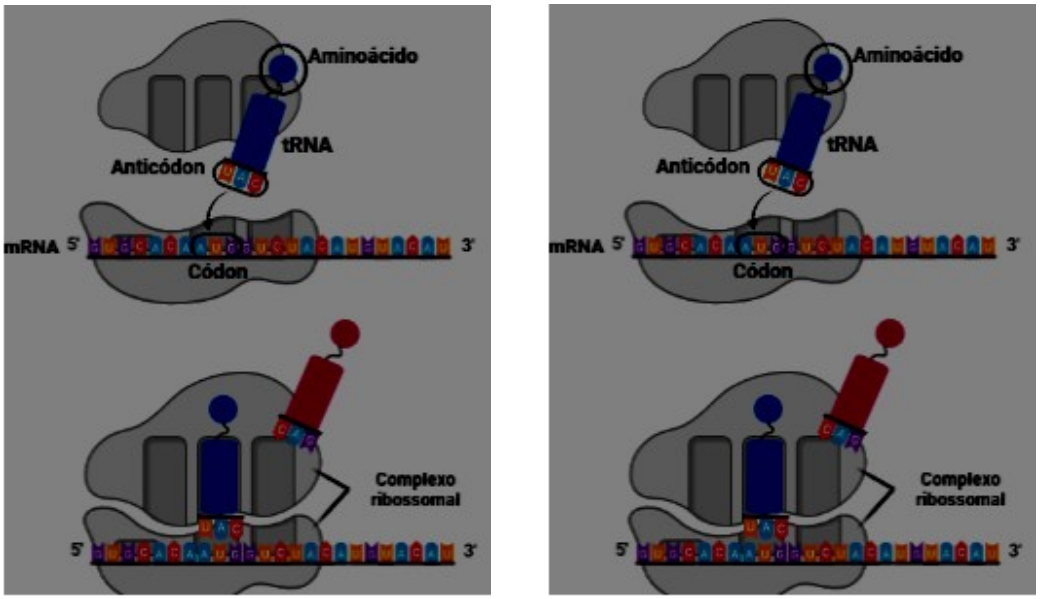
Fonte: Autoria própria, 2023 (Figura criada com o BioRender.com).

Considerando que esse processo ocorre milhões de vezes nas células eucarióticas, durante toda a vida de um indivíduo, a ocorrência de falhas é uma realidade. Acrescer ou remover um único nucleotídeo de um éxon pode acarretar mutações nas proteínas de interesse, tornando-as disfuncionais na maioria das vezes. Para esse trabalho especializado, um grupo de enzimas é encarregado de clivar os íntrons e unir novamente a sequência, garantindo sempre que as etapas descritas sejam efetuadas (NELSON; COX, 2014).

A compreensão do código genético foi imprescindível para o entendimento da etapa de tradução que ocorre pela ligação do ribossomo à fita de mRNA para traduzir os códons. Essa etapa tem uma alta demanda energética, assim como a síntese proteica como um todo. E por isso, ao longo da evolução dos genomas, mutações que reduzem o custo energético do processo de tradução devem ter sido favorecidas. O gasto de ATPs e de outras moléculas desde a adenilação dos aminoácidos, na formação do complexo aminoácido-tRNA e as ligações peptídicas demonstram necessidade de vias especializadas em sintetizar e metabolizar proteínas em um curto período (ALBERTS, 2017; ENCINAS PONCE, 2014).

Quando o mRNA maduro sai do núcleo para o citoplasma e se associa ao ribossomo que inicia a leitura dos códons no sentido 5’3’ mediando a ligação dos anticódons que se encontram na extremidade da molécula de tRNA, que por sua vez desempenha uma função adaptadora ao reconhecer a enzima que se liga ao aminoácido para carregá-lo e conectar o anticódon ao códon do mRNA. Todas as moléculas transportadoras têm muitas características estruturais comuns e devem ser capazes de interagir quase do mesmo modo com ribossomos e fitas de mRNA. Os aminoácidos que se ligam geralmente na extremidade 3’ (braço aceptor) do tRNA são carregados e recebidos em um dos sítios do ribossomo (NELSON; COX, 2014; ALBERTS, 2017**).**

Figura 4 - Etapa de tradução do mRNA na síntese proteica



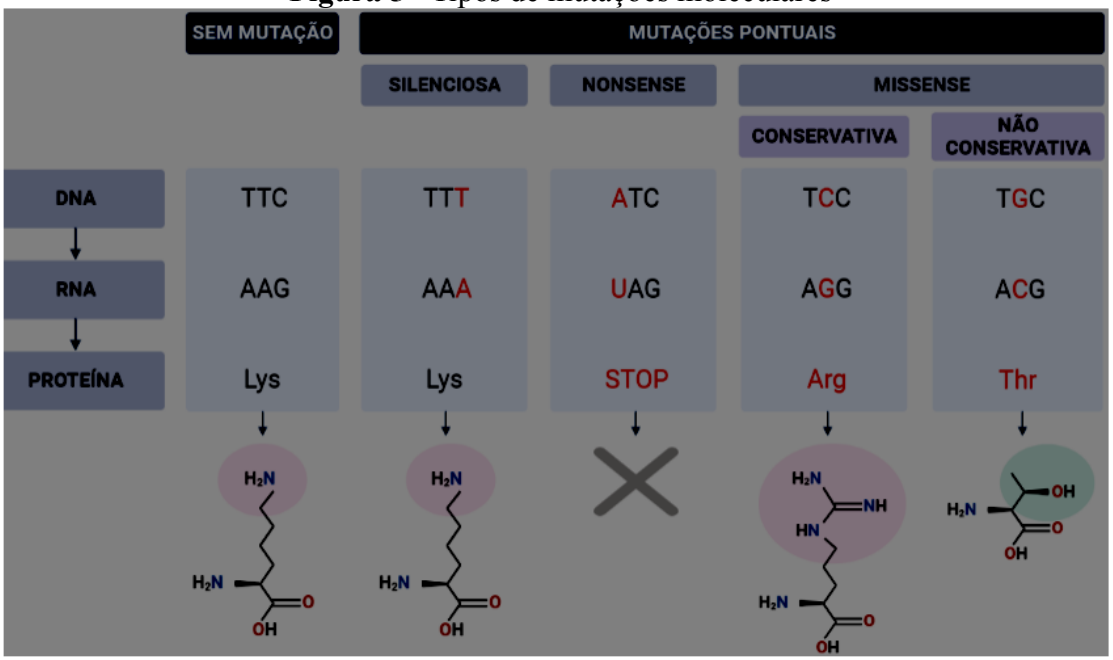
Fonte: Autoria própria, 2023 (Figura criada com o BioRender.com).

A tradução (Figura 4) começa a partir do códon de iniciação, A-U-G, que de acordo com o código genético codifica a metionina. Logoa após, um segundo tRNA é recebido no sítio ribossomal, carregando o próximo aminoácido que será adicionado à sequência. A ligação peptídica é estabelecida entre os dois e o primeiro tRNA é liberado, fazendo o segundo ser transferido para outro sítio e o ribossomo se deslocar na fita para receber o próximo. Dessa forma a fita é percorrida mediante a leitura e o processo continua até que um códon de terminação seja reconhecido pelo complexo para que haja a liberação da proteína, a separação do ribossomo e o fim da etapa de tradução (NELSON; COX, 2014; ALBERTS, 2017).

Assim como a síntese de proteínas, a duplicação do DNA é um processo que ocorre até o fim da vida dos organismos vivos. E por se repetir incontáveis vezes, mutações a nível molecular podem surgir na sequência de nucleotídeos e acarretar consequências às proteínas. Elas ocorrem aleatoriamente e podem ser deletérias, neutras ou benéficas. Resultantes de danos ambientais e endógenas ao DNA, são alterações passíveis de serem herdadas que garantiram toda a variabilidade genética existente (ALBERTS, 2017).

Com relação às proteínas, a alteração de um único nucleotídeo do DNA é chamada de mutação molecular (ALBERTS, 2017). Diferentes efeitos podem ser ocasionados por esse tipo de mudança, que pode ser de caráter silencioso, não silencioso ("*Missense*") ou sem sentido ("*nonsense*") como descrito na Figura 5.

Figura 5 - Tipos de mutações moleculares



Fonte: Autoria própria, 2023 (Figura criada com o BioRender.com).

Em caráter silencioso (Figura 5), há uma alteração de um nucleotídeo no códon, mas que codifica o mesmo aminoácido para compor a proteína. Isso se dá pela característica degenerada do código genético, que permite que um aminoácido seja codificado por mais de um códon (Figura 3). Em caráter *nonsense* (Figura 5), ocorre quando há alteração pontual de um nucleotídeo, de tal forma que o códon resultante será um dos três que indicam o término da sequência codificante. Nesses casos, a alteração impede com que a proteína nascente seja finalizada, interrompendo sua produção prematuramente, podendo ou não, preservar sua

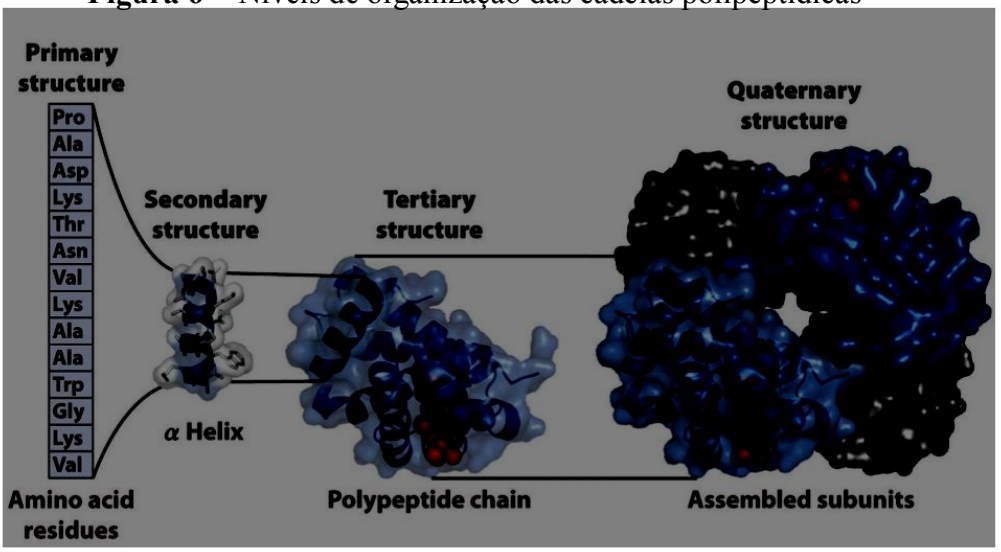
função de acordo com o local em que a mutação foi inserida (DEHGHANPOOR et al., 2018; GERASIMAVICIUS; LIU; MARSH, 2020).

Em caráter *missense* (Figura 5), o códon para um aminoácido é trocado por outro códon que, consequentemente, codifica um aminoácido diferente. Em casos que o aminoácido substituto for quimicamente similar ao substituído, pode-se identificar essa mutação como *missense* de caráter conservativo, já que é menos provável que a alteração afete significativamente a estrutura e função da proteína. Porém, um aminoácido pode ser substituído por outro quimicamente diferente, evidenciando uma mutação *missense* de caráter não conservativo, que pode acarretar sérios danos à estrutura e perda total de suas funções. Assim, esse tipo de mutação têm sido alvo de importantes pesquisas com variantes de um único nucleotídeo, já que essa alteração leva à substituições de aminoácidos no nível proteína e estão associadas a mais da metade de todas as doenças hereditárias conhecidas (DEHGHANPOOR et al., 2018; GERASIMAVICIUS; LIU; MARSH, 2020). As mutações no geral, causam alterações estruturais que podem s

3.5 ESTRUTURAS TRIDIMENSIONAIS DE PROTEÍNAS

Ao serem sintetizadas, as proteínas passam a ser classificadas quanto ao nível conformacional adquirido, uma vez que a conformação se refere às diversas formas que elas podem tomar devido suas ligações, assumindo diferentes estruturas moleculares, variando sua complexidade e consequentemente o desempenho de suas funções específicas.

Figura 6 - Níveis de organização das cadeias polipeptídicas



Legenda: Estrutura primária: representada por uma sequência de resíduos de aminoácido; Estrutura secundária: representada pelo arranjo □- hélice; Estrutura terciária: representada por um conjunto de arranjos secundários; Estrutura quaternária: formada por um conjunto de estruturas terciárias.

Fonte: NELSON; COX, 2014.

A estrutura primária, como pode ser visto na Figura 6 à esquerda, é o nível mais básico de organização estrutural das proteínas, caracterizada por uma sequência de resíduos de aminoácidos que é quimicamente estabilizada pela ligação peptídica, oferece um . Uma modificação substitutiva nos resíduos pode gerar mudanças conformacionais, resultando em prejuízo do desempenho de sua função biológica (NELSON; COX, 2014; VERLI, 2014).

Arranjos regulares, formados pelo enovelamento da cadeia principal da estrutura primária, constituem a estrutura secundária (Figura 6). Dentre os dobramentos mais recorrentes estão as □-hélices, conformações e voltas □, que apresentam-se nas formas helicoidal, pregueada e espiral randômica, respectivamente. Essa classificação se refere à disposição espacial dos resíduos de aminoácidos, e consequentemente, a qualquer segmento da cadeia polipeptídica, sem considerar a posição das cadeias laterais e relações com outras partes. Existem ainda estruturas secundárias que fogem a regra, não apresentando um padrão definido, impossibilitando a descrição adequada dos segmentos por apresentarem aleatoriedade espacial (NELSON; COX, 2014; ALBERTS, 2017; VERLI, 2014).

A estrutura terciária (Figura 6), que é conhecida como forma nativa, considera o arranjo tridimensional total dos átomos da proteína. Os resíduos que estavam distantes na sequência polipeptídica, ou em diferentes estruturas secundárias, com o dobramento da proteína podem interagir covalentemente mantendo suas posições terciárias por diferentes tipos de interações fracas. O arranjo formado por subunidades terciárias constituem as estruturas quaternárias, que são duas ou várias proteínas aglomeradas, por forças eletrostáticas, para formação de um complexo funcional maior, como pode ser visto ainda na Figura 6, à direita. Nesse quarto e último nível de organização, elas podem se dividir em fibrosas, com a cadeia polipeptídica arranjada como longos filamentos ou folheada, ou globulares, apresentando cadeias esféricas ou globulares. Os dois grupos divergem estruturalmente e funcionalmente (NELSON; COX, 2014; ALBERTS, 2017; VERLI, 2014).

As proteínas globulares apresentam maior complexidade estrutural por geralmente apresentarem mais de um tipo de estrutura secundária. Ao longo da cadeia dobrada, padrões estruturais não hierárquicos também podem ser encontrados e identificados em outras

proteínas. Motivo é o nome dado às porções enoveladas que podem envolver apenas dois segmentos secundários sobrepostos, ou pode ser uma estrutura bem elaborada que envolve uma grande quantidade de segmentos (NELSON; COX, 2014).

Outro termo que é utilizado para descrever padrões estruturais mais arranjados em proteínas globulares é o domínio. Não se tratando apenas de segmentos, mas agora de uma parte da cadeia polipeptídica, as estruturas que são compostas por centenas de resíduos de aminoácidos se dobram em um ou múltiplos domínios, cada qual como uma porção globular distinta que pode se movimentar independente do todo. Diferentemente de proteínas pequenas, que normalmente tem somente um domínio, proteínas maiores podem conservar sua estrutura quando separadas do resto. Assim, a comparação de vários domínios recorrentes ou motivos em diversas proteínas, revela que padrões estruturais são conservados ao longo da evolução do genoma, evidenciando que estado nativo pode fornecer muita informação sobre a evolução concomitante do proteoma do que apenas a sequência de aminoácidos (ALBERTS, 2017; VERLI, 2014).

As subestruturas tornaram-se importantes para os estudos de homologia, onde segmentos ultra conservados de uma proteína são utilizados para analisar outra que tem uma semelhança estrutural. Essas relações são possíveis e permitem o agrupamento de proteínas em famílias e superfamílias, baseando-se no grau de semelhança estrutural e funcional, que corroboram com as hipóteses de relações evolutivas.Ainda que muitas estruturas nativas de proteínas sejam resolvidas, as semelhanças evolutivas e os princípios físico-químicos de organização como também as forças que as estabilizam no espaço, não são claramente definidas, mas resultantes importantes como as interações polares e apolares são de grande importância para chegar ao entendimento dos processos necessários para se ter o proteoma atual (PRIVALOV; KHECHINASHVILI, 1974; NELSON; COX, 2014).

Além de todos os padrões já conhecidos e das interações que foram possíveis de mapear, o entendimento do enovelamento de polipeptídeos é um desafio de alta complexidade. As limitações físicas e químicas são vencidas a passos curtos pelo esclarecimento da arquitetura proteica com ajuda de softwares criados para processar os dados biológicos e gerar modelos estruturais que se aproximem das estruturas nativas. As subestruturas (domínios e motivos) têm ajudado na identificação funcional e estrutural das classes que compõem o proteoma, permitindo que semelhanças sejam estabelecidas a partir das regiões conservadas para chegar ao conhecimento de tais fatores evolutivos (PRIVALOV; KHECHINASHVILI, 1974; NELSON; COX, 2014).

3.6 TÉCNICAS DE DETERMINAÇÃO DE ESTRUTURAS TRIDIMENSIONAIS

Diante da necessidade de obter estruturas proteicas em seu estado nativo, a fim de compreender sua funcionalidade, métodos experimentais e computacionais estão sendo desenvolvidos e aperfeiçoados ao passo que uma grande quantidade de dados é gerada e armazenada em bancos.

3.6.1 Técnicas experimentais

Dentre as técnicas experimentais que vêm sendo aplicadas para a obtenção de informações sobre as estruturas proteicas, destacam-se a difração de raios-X, a calorimetria de varredura diferencial (DSC), a ressonância magnética nuclear (RMN) e o dicroísmo circular (CD).

A difração de raios-X permite que a estrutura de proteínas seja determinada em uma escala quase atômica. O ensaio é baseado na geração de cristais contendo proteínas que recebe radiação incidente de um comprimento de onda específico. O raio-X é difratado pelos elétrons que estão distribuídos no cristal e assim, é possível inferir da posição dos próprios núcleos, que também podem ser determinados por difração com feixe de nêutrons (DEVLIN, 2011).

A Calorimetria de varredura diferencial (DSC) é uma técnica usada para caracterizar a estabilidade de proteínas ou outras biomoléculas diretamente em sua forma nativa. Elas são aquecidas a uma taxa de varredura constante, que causa a absorção de calor a partir do desdobramento da estrutura, resultando em um gradiente térmico (ΔT) entre as células. Ainda, os modelos termodinâmicos podem ser ajustados aos dados para obter a energia livre de Gibbs (ΔG), a entalpia calorimétrica (ΔH_{cal}), a entalpia de van't Hoff (ΔH_{vH}), a entropia (ΔS) e a mudança da capacidade de calor (ΔC_p) associada à transição (DEVLIN, 2011).

Ressonância magnética nuclear (RMN) é um ensaio realizado com macromoléculas em solução sob influência de um campo eletromagnético estático potente e baseia-se na liberação de um pulso de energia eletromagnética, em diferentes ângulos, na solução. Parte da energia é absorvida, à medida que o núcleo das moléculas muda do estado de menor energia, que corresponde à orientação paralela do dipolo magnético gerado pelo momento angular do spin nuclear, para o estado de maior energia, com orientação antiparalela ao campo. O espectro resultante apresenta informações sobre a identidade do núcleo e o ambiente químico

das imediações. A RMN torna-se vantajosa por também esclarecer mudanças conformacionais, no enovelamento e interações com outras moléculas (DEVLIN, 2011).

Dicroísmo Circular (CD) é utilizado para mensurar a quantidade e tipo de estruturas secundárias presentes em solução, pois tem se mostrado um ensaio sensível, principalmente à estruturas alfa-hélice e folhas-beta e desordenadas, em virtude dos diferentes espectros gerados em uma faixa de comprimentos de onda. O CD é causado por diferenças na absorção de luz entre os componentes em sentido horário e anti-horário de um feixe de luz polarizada que atravessa uma solução opticamente ativa (DEVLIN, 2011).

3.6.2 Métodos computacionais

Os métodos computacionais para predição de estruturas tridimensionais de proteínas podem ser separados em quatro categorias: Modelagem comparativa por homologia, métodos de reconhecimento de padrões de enovelamento, métodos *ab initio* e *de novo.*

3.6.2.1 Métodos de modelagem comparativa por homologia

Tabela 1 - Principais preditores de de estrutura por modelagem comparativa

PREDITOR	DESCRIÇÃO	LINK
Modeller	Modelagem comparativa de estrutura de proteína por satisfação de restrições espaciais. O usuário fornece um alinhamento de uma sequência a ser modelada com estruturas relacionadas conhecidas e o MODELLER calcula automaticamente um modelo contendo todos os átomos	https://salilab.org/modeller/

PREDITOR	DESCRIÇÃO	LINK
SWISS-MODEL	É um serviço web dedicado à modelagem de homologia de estruturas de proteínas. O servidor constrói modelos a partir da (1) identificação do(s) modelo(s) estrutural(is), do (2) alinhamento da sequência alvo e estrutura(s) do modelo, da (3) construção do modelo e avaliação da qualidade.	https://swissmodel.expasy.org/

Fonte: adaptado de MARQUES, 2021.

3.6.2.2 Métodos de reconhecimento de padrões de enovelamento

Tabela 2 - Principais preditores de estrutura por reconhecimento de padrões

PREDITOR	DESCRIÇÃO	LINK
HHpred	Utiliza sequências ou alinhamento de sequência múltipla como entrada e procura por homólogos remotos em uma variedade de bancos de dados, como PDB, SMART e Pfam.	https://toolkit.tuebingen.mpg.de/tools/hhpred
SPARKS-X:	é um aprimoramento do reconhecimento de dobramento de proteína, empregando correspondência baseada em probabilística entre propriedades estruturais nativas e dos modelos gerados.	https://sparks-lab.org/server/sparks-x/

Fonte: adaptado de MARQUES, 2021.

3.6.2.3 Métodos *de novo*

Tabela 3 - Principais preditores de estrutura por métodos *de novo*

PREDITOR	DESCRIÇÃO	LINK
I-TASSER	É uma abordagem hierárquica para previsão da estrutura de proteínas e anotação de função baseada em estrutura. Ele identifica modelos estruturais do PDB pela abordagem de segmentação múltipla LOMETS, com modelos atômicos completos construídos por simulações interativas de montagem de fragmentos baseadas em modelos.	https://zhanggroup.org/I-TASSER/
ROSETTA interações	É um software utilizado para prever e projetar estruturas de proteínas, mecanismos de dobramento de proteínas e proteína-proteína, a partir de estruturas depositadas em bases de dados.	https://www.rosettacommons.org/software

Fonte: adaptado de MARQUES, 2021.

3.6.2.4 Métodos *ab initio*

Tabela 4 - Principais preditores de estrutura por método *ab initio*

PREDITOR	DESCRIÇÃO	LINK
Amber	é um conjunto de programas de simulação, que utiliza um conjunto de campos de força mecânico molecularar, de domínio público, para a simulação de biomoléculas.	https://ambermd.org/
CHARMM	É um programa de simulação molecular que visa, principalmente, sistemas biológicos. Disponibiliza um conjunto abrangente de	https://www.charmm.org/

	funções de energia, uma variedade de métodos de amostragem aprimorados, para análise e construção de modelos.	
GROMOS	É um pacote desenvolvido para a modelagem dinâmica de biomoléculas usando os métodos de dinâmica molecular, dinâmica estocástica e minimização de energia.	https://www.gromos.net/

Fonte: adaptado de MARQUES, 2021.

3.6.2.5 AlphaFold

Quando em 2018, os resultados da Avaliação Crítica de Predição de Estruturas (CASP) foram anunciados, o DeepMind, grupo de pesquisa de aprendizado de máquina do *Google*, recebeu o primeiro lugar, pelo desenvolvimento do *AlphaFold*, um algoritmo baseado em redes neurais profundas, no qual modelos estruturais de proteínas foram gerados por meio do uso de previsões de distância ou contato entre pares de resíduos de aminoácidos (MARQUES, 2021).

O AlphaFold, foi treinado a partir de um conjunto de dados públicos de proteínas com estruturas tridimensionais conhecidas e um banco de dados de sequências sem estruturas conhecidas. Como os modelos gerados apresentaram alta precisão, considerou-se um grande passo para solucionar o problema de entendimento sobre o enovelamento de proteínas (SENIOR et al, 2019).

3.7 DINÂMICA DE ENOVELAMENTO E DESNATURAÇÃO

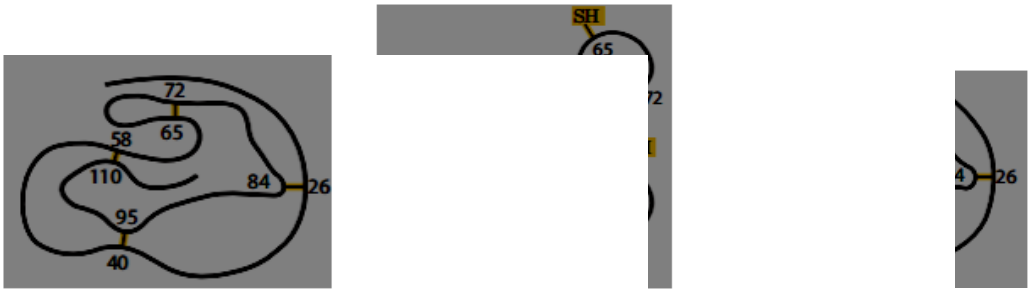
Proteínas têm um tempo de vida hábil muito curto e constantemente precisam de manutenções. Com a evolução do proteoma humano, as estruturas proteicas adaptaram-se para trabalhar no ambiente celular e não toleram mudanças abruptas. Diante de condições diferentes, elas podem desnaturar e por isso, vias de verificação formadas por centenas de enzimas e outras proteínas especializadas, estão presentes no interior da célula e contribuem

para o equilíbrio entre a síntese e a degradação de proteínas que estejam mal dobradas, desnaturadas e mutadas (NELSON; COX, 2014).

O estado desnaturado é caracterizado pelo desdobramento da estrutura tridimensional, de forma parcial ou completa, e consequentemente, perda de sua função. A desnaturação pode ser induzida em diversas condições e pela ação de agentes como calor, mudanças de pH e solventes orgânicos, que suscitam a exposição das regiões mais hidrofóbicas da proteína, permitindo a formação de agregados (NELSON; COX, 2014).

Algumas proteínas globulares desnaturadas podem reassumir sua conformação nativa e função se forem postas novamente em condições ideais. Esse processo, chamado de renaturação, foi demonstrado por (ANFINSEN, 1973), enquanto estudava o enovelamento da ribonuclease A, que possui uma única cadeia principal e quatro pontes dissulfeto, propondo que a estrutura tridimensional de uma proteína seria consequência direta de sua estrutura primária.

Figura 7 - Experimento de Anfinsen: Renaturação da ribonuclease desnaturada e desenovelada



(1) (2) (3)

Legenda: RNase, PDB: 1KF5. A ureia desnatura a ribonuclease A, e o beta-mercaptoetanol a reduz, rompendo as ligações dissulfeto. A renaturação acontece com a retirada dos reagentes e a volta da conformação nativa.

Fonte: adaptado de NELSON; COX, 2014.

Por ser uma enzima, o estado dobrado da RNase A pode ser atestado pela mensuração de sua atividade enzimática. E partindo do princípio posposto, Anfinsen preparou amostras enzimáticas usando combinações de dois reagentes diferentes, uréia 8M ($CO(NH_2)_2$) e beta-mercaptoetanol (βME). O βME promoveu a quebra das quatro pontes dissulfeto, resultando em oito resíduos de cys, enquanto a uréia induziu o desenovelamento da enzima por interagir com o esqueleto peptídico rompendo as ligações hidrofóbicas de estabilização.

Anfinsen atestou a perda total de atividade catalítica e então removeu a ureia e o βME, por diálise, e adicionou concentrações de um agente oxidante para catalisar a volta das ligações dissulfeto (ANFINSEN, 1973; HUA et al., 2008).

Quando a ureia e o βME foram removidos, a RNase A desnaturada, voltou à estrutura nativa correta de modo que as ligações dissulfeto se estabelecem nos mesmos lugares e consequentemente, a atividade catalítica foi restaurada. Dessa forma, ele concluiu que a informação necessária para enovelar uma proteína está contida em sua sequência de aminoácidos. Estudos posteriores, demonstraram que apenas um grupo de proteínas pequenas e estáveis, enovelam-se espontaneamente (ANFINSEN, 1973; TANOUYE, 2017; NELSON; COX, 2014).

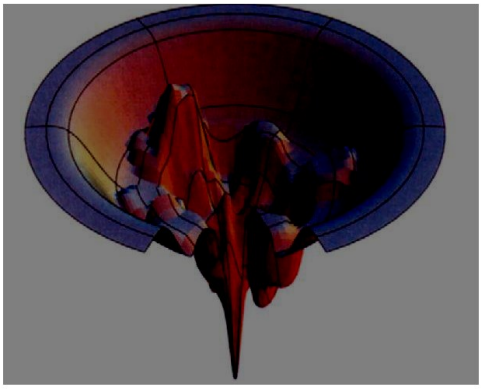
O enovelamento de proteínas sendo determinado pela sequência de aminoácidos, como já descrito, é tido como dogma para a maioria das proteínas. Mas, entender como ocorre o dobramento da estrutura em pouco tempo, levou Cyrus Levinthal a postular que as milhares conformações possíveis de dobramento de uma estrutura proteica não são testadas aleatoriamente até que se ache a melhor estabilidade (TANOUYE, 2017). O paradoxo de Levinthal considera a sequência temporal dos eventos intermediários que ocorrem entre os estados desenovelado e nativo. O pressuposto de que aleatoriamente as possíveis conformações vão sendo testadas à medida que a proteína emerge do ribossomo, foi contraposto com o tempo necessário para validar todas essas posições como não nativas. Considerando que a cada $\sim 10^{-13}$ segundos (tempo de uma vibração molecular) uma conformação fosse assumida, para uma proteína formada por 100 resíduos de aminoácidos, seria necessário, aproximadamente, 10^{77} anos para que todas as posições fossem testadas (NELSON; COX, 2014). Levinthal então sugeriu que o enovelamento das estruturas devia ser acelerado e guiado pela rápida formação de interações locais de curto e longo alcance (TANOUYE, 2017).

As interações de curto e longo alcance submetem a estrutura à vias de enovelamento que limitam o acesso da proteína a determinadas conformações não favoráveis, consequentemente direcionando o dobramento da cadeia para as posições mais favoráveis e de menor energia livre. As □-hélices e conformações □ locais são formadas primeiro devido a uma série de restrições que norteiam seus surgimentos. Em seguida, interações iônicas de grupos carregados e ligações de longo alcance são estabelecidas, além das interações hidrofóbicas que promovem a agregação das partes apolares dos resíduos, conferindo uma estabilidade entrópica a formas enoveladas intermediárias. O processo continua até que os domínios sejam completamente formados e estabilizados. A formação de estados

intermediários não nativos com alto grau de entropia conformacional e energia livre relativamente alta ocorre nas vias de enovelamento. Um intermediário bem descrito é o glóbulo fundido. Nesse estado, a estrutura se apresenta condensada, contendo muitas estruturas secundárias favoráveis à conformação nativa, mas ainda muitas interações desfavoráveis à estrutura terciária (NELSON; COX, 2014; TANOUYE, 2017; HUA et al., 2008).

Anos de trabalho experimental e teórico forneceram uma compreensão mecanicista detalhada das forças que governam o dobramento das proteínas. A dinâmica de enovelamento pode ser visualizada com um funil (Figura 8) que descreve a tendência termodinâmica da estrutura de assumir uma conformação de menor variação energética, que é expressa em kcal/mol (variação da energia livre de Gibbs, $\Delta\Delta G$). O início do dobramento é representado pela área superior do funil onde a energia livre se encontra em maior grau e com a formação dos estados intermediários semi estáveis, depressões ao longo das paredes do funil representam as variações conformacionais e energéticas que ocorrem durante todo o processo. Convergindo para o fundo do funil, o ponto de menor variação de energia livre conformacional é encontrado e o conjunto de intermediários é reduzido a uma única conformação nativa estável (GERSHENSON et al., 2020; NELSON; COX, 2014).

Figura 8 - Paisagem termodinâmica de energia livre em forma de funil



Legenda: paisagem de energia em forma de funil onde as estruturas nativas em seu mínimo global guiam cada molécula de um conjunto heterogêneo de cadeias polipeptídicas desdobradas de alta energia através de diferentes vias de enovelamento até o fundo do funil, onde a estrutura estará condensada em uma forma que apresente a menor variação de energia livre de Gibbs

Fonte: Adaptado de (GERSHENSON et al., 2020).

O estudo da estabilidade do estado nativo e desnaturado das proteínas são temas de grande interesse biotecnológico, onde o controle desses estados configuram uma importante estratégia para o desenvolvimento de novos fármacos, por exemplo. E à medida que o

conhecimento avança, a necessidade de novas metodologias biofísicas e de softwares cada vez mais específicos também cresce, visando a predição de estruturas tridimensionais a partir dos métodos de determinação e análise do impacto de mudanças termodinâmicas e cinéticas de proteínas mutantes (MCGUINNESS et al., 2018; PÁL; PAPP; LERCHER, 2006).

A termoestabilidade tem sido um alvo importante para a biotecnologia, pois sua mensuração por meio das técnicas experimentais, tem sido base para a seleção de mutantes que permaneçam funcionais em ambientes não nativos (MCGUINNESS et al., 2018; PÁL; PAPP; LERCHER, 2006). Ela pode ser definida pela temperatura na qual metade da população de proteínas está em um estado dobrado, ou seja, temperatura de fusão (Tm). A análise da variação dessa temperatura de fusão é um objetivo importante em biotecnologia ao projetar proteínas, pois mutações *missenses* podem favorecer o aumento da estabilidade térmica, reduzir agregações indesejadas e permitir a manutenção das funções em diferentes condições. Dessa forma, tem havido grande interesse no desenvolvimento e aplicação de métodos computacionais, gerando dados termodinâmicos para ajudar a orientar experimentalmente o processo de seleção.

3.7.1 Preditores de efeito de mutações

Ferramentas computacionais para predição de efeitos causados por mutações na estrutura e estabilidade de proteínas, estão periodicamente sendo desenvolvidas ou aperfeiçoadas. Várias destas alcançaram altas taxas de previsão e precisão na faixa de 70 a 80% ao gerar modelos que se aproximam das estruturas nativas identificadas a partir das técnicas experimentais. (DEHGHANPOOR et al., 2018; GERASIMAVICIUS; LIU; MARSH, 2020) Os principais preditores estão descritos na Tabela 1.

Tabela 5 - Principais preditores de estabilidade de proteínas

PREDITOR	DESCRIÇÃO	LINK
DynaMut	Preditor que utiliza os resultados de três	https://biosig.lab.uq.edu.au

–	outros programas: Bio3D, ENCoM e DUET, para avaliar o impacto das mutações na estabilidade das proteínas. Devido à sua natureza, o preditor utiliza várias metodologias, como a análise de modo normal e potenciais estatísticos.	/dynamut/
ENCoM	Um método de previsão baseado na análise que relaciona as alterações na entropia vibracional após a mutação com as alterações na estabilidade da proteína. Utiliza representações proteicas de granularidade que levam em conta as propriedades dos resíduos.	http://biophys.umontreal.ca/nrg/resources.html
DUET	Um preditor baseado em aprendizado de máquina que aproveita os resultados dos preditores SDM e mCSM integrados, utilizando máquinas de vetores como suporte.	https://biosig.unimelb.edu.au/duet/stability
SDM	Um potencial servidor baseado no conhecimento derivado da utilização de propensões evolutivas de substituição de resíduos específicos.	Não está mais disponível como um servidor autônomo
FoldX	Consiste em interação baseada na física e termos entrópicos, parametrizados em dados de treino empírico. Permite executar facilmente previsões em conjuntos de várias cadeias.	https://foldxsuite.crg.eu/
Rosetta	Conjunto de software de modelagem macromolecular Rosetta, que inclui algoritmos para a previsão de impacto na	https://www.rosettacommons.org/home

	estabilidade de biomoléculas. Acionado por uma função de pontuação que é uma combinação linear de termos energéticos, estatísticos e empíricos.	
INPS3D	Baseia-se na sequência e na conservação físico-química, empregando características derivadas da estrutura, como a acessibilidade do solvente e as diferenças de energia locais. O preditor é treinado através de regressão com vetores de suporte.	https://inpsmd.biocomp.unibo.it/inpsSuite/default/index3D
mCSM	Uma abordagem de aprendizagem automática que avalia as alterações de assinatura estrutural conferidas por mutações. Deriva uma representação gráfica das características físico-químicas e geométricas do ambiente dos resíduos	https://biosig.unimelb.edu.au/mcsm/stability
SDM2	Versão atualizada do SDM, é baseado no conhecimento de tabelas de substituição de resíduos específicos, informações sobre a conformação suas interações, bem como a densidade de empacotamento e a profundidade dos resíduos, para avaliar as alterações de estabilidade das proteínas.	https://marid.bioc.cam.ac.uk/sdm2/prediction
CUPSAT	Método de previsão que utiliza os ângulos de torção de resíduos e pares de átomos específicos do ambiente para avaliar as	https://cupsat.tu-bs.de/

	alterações de estabilidade.	
PoPMuSiC resíduo	Consiste em 13 termos estatísticos que consideram a diferença de volume entre os resíduos do tipo selvagem e mutante, bem como a acessibilidade do solvente do original para diferenciar as substituições do núcleo e da superfície.	https://soft.dezyme.com/query/create/pop
MAESTRO	Combina 3 funções de pontuação estatística de exposição ao solvente e distâncias de pares de resíduos, bem como 6 propriedades proteicas, num quadro de aprendizagem automática para obter uma previsão consensual do impacto na estabilidade	https://pbwww.che.sbg.ac.at/maestro/web
I-Mutant 3.0 _ de	Um método derivado de aprendizagem automática que considera o ambiente espacial do resíduo mutado em termos de tipos resíduos circundantes e acessibilidade de superfície.	https://gpcr2.biocomp.unibo.it/cgi/predictors/I-Mutant3.0/I-Mutant3.0.cgi

Fonte: adaptado de (GERASIMAVICIUS; LIU; MARSH, 2020).

3.8 THERMOMUTDB E A CURADORIA DE BASES BIOLÓGICOS

Desde as primeiras postulações sobre a hereditariedade e a importância dos ácidos nucléicos e seus derivados, a biologia molecular ascendeu na escala de geração de dados relacionados a processos ou sistemas biológicos, que podem ser chamados de dados biológicos. As abordagens informatizadas que tentaram explicar processos bióticos, por meio

das emergentes tecnologias, em meados de 1970, aliadas ao desenvolvimento das tecnologias de sequenciamento de DNA, apontaram para o surgimento de um ramo interdisciplinar de estudos, que envolvia a intensa utilização de recursos computacionais para solucionar os problemas evidenciados pelos dados. Dessa forma, a Bioinformática, surgia como uma subárea que deveria coletar, organizar, armazenar e analisar os dados biológicos, mediando sua interpretação (MELGAÇO,2021; NELSON; COX, 2014).

Com o constante aumento do poder tecnológico, desde o início das abordagens computacionais, a bioinformática tornou-se complexa e abrangente, deixando de ser apenas um suporte, tornou-se um braço imprescindível para a ciência, combinando técnicas de computação e teoria da informação aplicada à biologia, para não só gerenciar dados, mas tratar, interpretar e disponibilizar suas demandas. A análise desses dados a partir do desenvolvimento de sistemas computacionais, converge na implementação de bases de dados biológicos. São enormes bibliotecas virtuais que reúne dados relacionados e categorizados que possuem significado implícito e finalidade específica requerida por usuários reais (LIBÓRIO; RESENDE, 2021).

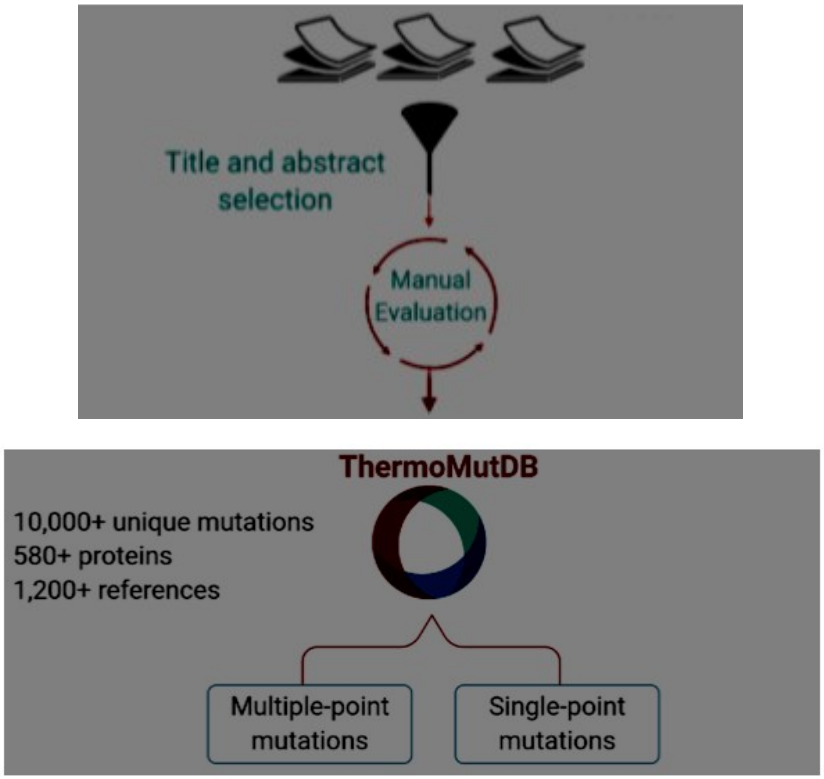
A bioinformática se vale de bancos primários para gerir os dados recuperados dos ensaios e métodos experimentais que podem ser depositados em bancos especializados como o PDB (Protein Data Bank), que armazena milhares de estruturas tridimensionais de proteínas e ácidos nucléicos, ou seguir para bancos também especializados, mas que recebem os dados a partir dos níveis e métodos de curadoria que precisam ser submetidos (LIBÓRIO; RESENDE, 2021), como o ThermoMutDB que se propõe a solucionar o problema de recuperação e curadoria de dados termodinâmicos de mutações missenses de proteínas (XAVIER et al., 2021).

No que diz respeito a curadoria de dados biológicos, grandes problemas podem ser encontrados quando se trabalha com dados experimentais brutos. Etapas de tratamento estatístico são recorrentes antes de quaisquer publicações científicas serem submetidas a repositórios, como o PubMed, por exemplo. Ainda assim, os dados presentes no imenso volume de conteúdo precisam ser sujeitos a uma curadoria manual, para a qual, o trabalho humano se faz necessário. Geralmente, pessoas especializadas agrupam publicações de interesse e extraem as informações relevantes para a base de dados aplicando técnicas de padronização e garantindo que toda a informação seja preservada e de fácil recuperação (Melgaço 2021; ALLOT et al., 2021; XAVIER et al., 2021).

A busca por referências que contenham os dados desejados para alimentar as bases também configura um importante problema a ser solucionado. Pois, o tempo gasto na

pesquisa, identificação, checagem e curadoria é um inconveniente que pode tornar as atualizações e depósito de novos dados infrequentes. Para superar isso, a ThermoMutDB conta com duas frentes para aquisição dos dados. A primeira estratégia é a disponibilização de um espaço de contribuição para os usuários submeterem as suas próprias contribuições referentes à publicação, a proteína, protocolo experimental e informações termodinâmicas para uma ou mais mutações. Ao serem enviadas, as informações são selecionadas manualmente e incorporadas na base. A segunda estratégia de aquisição é pela recuperação de referências depositadas em repositórios. Esta é a principal via de obtenção de dados termodinâmicos para a base e também a forma mais árdua, pois ainda utiliza de recursos humanos para tal. O alinhamento dessas duas estratégias garantem a melhoria significativa na quantidade e na qualidade dos dados, permitindo não apenas o desenvolvimento de uma nova geração de métodos, mas também uma avaliação imparcial dos métodos propostos anteriormente(Melgaço 2021; XAVIER et al., 2021).

Figura 9 - Fluxo de Aquisição e processamento de dados para desenvolvimento do ThermoMutDB



Fonte: Adaptado de (XAVIER et al., 2021).

Relacionado a curadoria manual de dados biológicos, o ThermoMutDB é uma base de dados pública desenvolvida para armazenar dados termodinâmicos, incluindo o DDG de proteínas após mutações missense, além de conter informações da proteína, informações mutacionais, métodos e condições experimentais (IQBAL et al., 2021; XAVIER et al., 2021). E com >14.669 dados experimentais de parâmetros termodinâmicos para proteínas de tipo selvagem e mutantes, essa base pública é uma das maiores no âmbito proposto.

3.8.1 Processamento de linguagem natural

O Processamento de Linguagem Natural (PLN) compreende o desenvolvimento de modelos computacionais dedicados a analisar e criar uma linguagem para os seres humanos, de forma que seja possível a comunicação entre pessoa e máquina. O processamento da linguagem torna-se possível, a partir da extração de atributos, que são variáveis presentes no conjunto de dados e são fundamentais no reconhecimento de padrões (H. A. MALTA; ANTONIO R. L. KUROIVA, 2019).

3.8.2 Aprendizado de máquina e mineração de textos e dados

O aprendizado de máquina vem sendo utilizado para diversas finalidades, por compreender estudos estatísticos e aplicações de algoritmos computacionais para aquisição de novos saberes e modos diferenciais de gerir conhecimentos já existentes. De acordo com os tipos de dados recebidos, retornados e pelo problemas propostos, pode-se dividir o aprendizado em aprendizado supervisionado/semi-supervisionado, não supervisionado e por reforço (MARQUES, 2021; DALL'AGNO, 2012).

O aprendizado supervisionado, compreende algoritmos que geram um modelo matemático a partir de dados recebidos, fornecidos em grupos distintos, o conjunto de treino e conjunto de teste. Esse tipo de aprendizado é indicado para finalidade de predição e classificação, assim como o aprendizado semi-supervisionado. Porém, uma diferença entre esses dois tipos de aprendizado, é a rotulagem (saída), uma vez que para o semi-supervisionado, nem todos os dados possuem rótulos (MARQUES, 2021; DALL'AGNO, 2012)

O aprendizado não supervisionado utiliza dados de de entrada sem rótulos. A ideia desse tipo de aprendizado é encontrar um agrupamento (cluster) de dados que expressem características comuns ((MARQUES, 2021; DALL'AGNO, 2012)).

O aprendizado por reforço, é uma abordagem que implica na realização de ações por um aprendizado que é recompensado. A tomada de decisão, que gera ações, objetivam uma maior recompensa, alcançada pela exploração do ambiente em que se encontra. Esse tipo de

aprendizado é comumente usado quando um computador aprende um jogo (MARQUES, 2021; DALL'AGNO, 2012).

O *data mining* ou mineração de dados, tem sido comumente utilizada nas áreas educacionais, financeiras e medicina, por permitir a extração de informações que antes da sua aplicação não era possível. A partir das técnicas de aprendizado de máquina, que é uma área da inteligência artificial voltada para o desenvolvimento de algoritmos e técnicas computacionais que possibilitam que computadores aprendam, a mineração se tornou possível e está relacionada à criação de modelos que sejam capazes de conectar um conjunto de entrada a um resultado específico (MARQUES, 2021; DALL'AGNO, 2012).

Dentre as tarefas mais comuns desempenhadas por esses modelos de aprendizado de máquina; classificação, estimação, predição, agrupamento e associação são as mais destacadas e funcionalmente utilizadas. O agrupamento e a associação de dados são considerados como aprendizagem não supervisionada. A classificação, estimação e predição estão associadas ao método de aprendizagem supervisionado (MARQUES, 2021; DALL'AGNO, 2012).

A mineração de textos, que é uma busca de informações relevantes a partir de um grande volume de textos, escritos na linguagem natural. Semelhante a mineração de textos, algoritmos de aprendizagem de máquina são utilizados a fim de promover a coleta de dados, pré-processamento, indexação, mineração e análise dos resultados contidos nos textos. Enquanto a mineração dos dados a informação extraída está implícita, a mineração de texto capta informações explícitas não estruturas (ALLOT et al., 2021; LIBÓRIO; RESENDE, 2021). Para melhoria quantitativa dos dados minerados em textos e otimização do tempo gasto com as etapas de busca e curadoria, sistemas web de recomendação e de curadoria de literatura, foram propostos como uma alternativa de automatizar a recomendação de literatura.

3.8.3 LitSuggest

O LitSuggest surge como uma alternativa que alia técnicas avançadas de aprendizado de máquina para gerenciamento e recomendação de literatura do repositório PubMed. A ferramenta, de modo geral, treina modelos de aprendizagem de máquina com base num conjunto de artigos fornecidos pelo usuário. O modelo depois de treinado é então utilizado para ordenar e classificar novas publicações que seguem para etapa de curadoria. A interface de curadoria atribui um score probabilístico para cada referência recuperada a partir do conjunto de treinamento e dispõe as informações bibliográficas de cada uma permitindo ao

usuário classificá-las como referências positivas (de interesse e que possivelmente tenha os dados buscados) e negativas (não são de interesse e possivelmente não tenham os dados buscados). A automatização dessa busca pode ser feita na própria ferramenta, de modo que semanalmente novas referências são adicionadas ao repertório do usuário, tornando-se uma potencial estratégia a ser utilizada para reduzir o tempo de busca de literatura que contenha dados biológicos e assim otimizar o tempo de atualização das bases de dados (ALLOT et al., 2021; LIBÓRIO; RESENDE, 2021).

4 MATERIAS E MÉTODO

4.1 MÉTODO

A busca por referências foi automatizada na plataforma do LitSuggest, a partir de um modelo, treinado previamente com um conjunto de PMIDs (identificador PubMed) positivos, e um conjunto de PMIDs negativos. Ambos os conjuntos, foram resultantes da metodologia aplicada para criação do ThermoMutDB. O modelo recuperou novos PMIDs e listou todos para serem submetidos a curadoria manual. Em seguida, os selecionados manualmente, tiveram seus dados minerados e preparados para submissão na base de dados. O fluxograma de aquisição dos dados está descrito na Figura 10.

Entre 24 de maio de 2021 e 11 de setembro de 2022, a ferramenta sugeriu semanalmente novas referências recuperadas do repositório PubMed, baseada no aprendizado oferecido anteriormente. Um total de 466 referências foram classificadas e receberam o score.

Dentro da base, são captadas informações termodinâmicas, condições experimentais e citações da literatura. Também ocorre a padronização das medições e cálculos nas entradas de dados, incluindo temperatura em Kelvin, energia em kcal/mol e energia livre de Gibbs (ΔΔG) como na Equação 1:

$$\Delta\Delta G = \Delta G\,(\text{ selvagem }) - \Delta G\,(\text{ mutante }) \qquad \text{Equação 1}$$

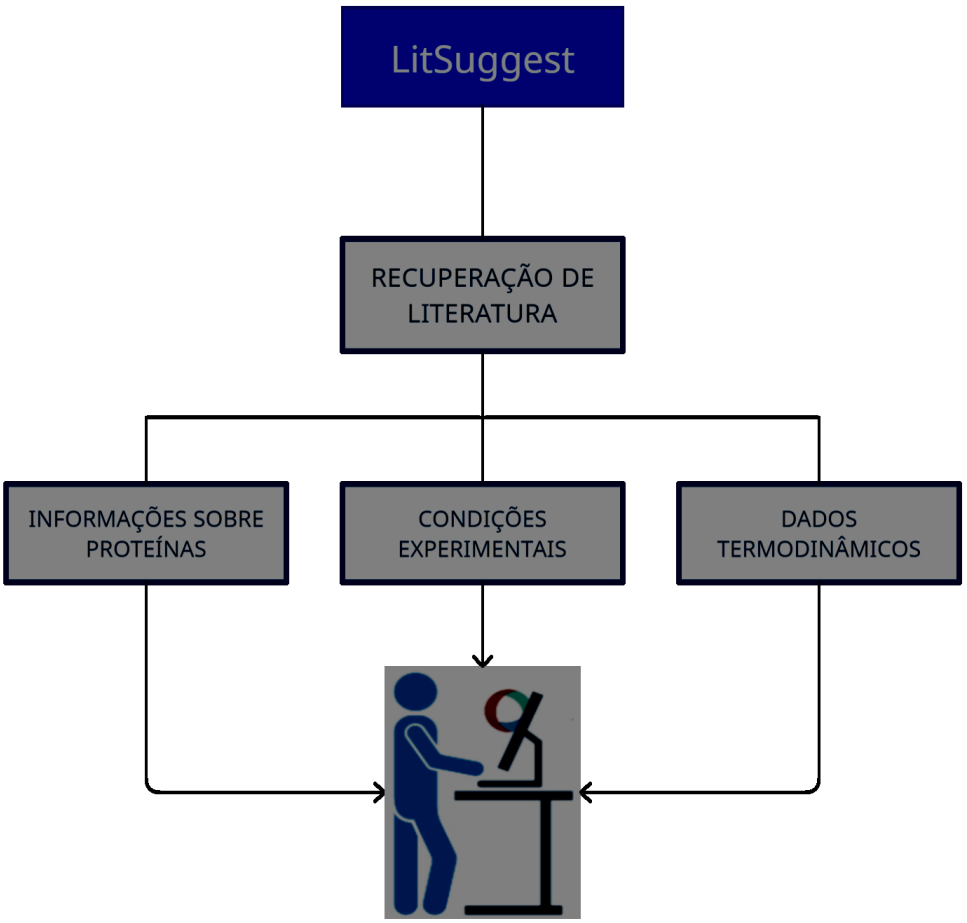
Na qual, valores ΔΔG negativos indicam que a mutação desestabilizou a proteína e valores ΔΔG positivos indicam que a proteína mutante é mais estável.

Durante a curadoria, as referências tiveram as sessões de resumo, introdução, metodologia e resultados explorados. As que não continham dados sobre mutações missenses foram retiradas do repertório, enquanto as que continham, tiveram seus dados minerados. As correções de ΔΔG foram efetuadas de acordo com a equação 1 e temperaturas convertidas

para valores em Kelvin, quando os dados se apresentavam fora da padronização estabelecida pela base de dados.

Para que uma referência fosse classificada e selecionada para submissão no ThermoMutDB, foram utilizados como requisitos a descrição sobre a proteína selvagem e as proteínas mutantes e valores de $\Delta\Delta G$ ou ΔTm, bem como as condições experimentais estabelecidas.

Figura 10 - Fluxograma de etapas de aquisição de referências por IA e processamento de dados para o ThermoMutDB



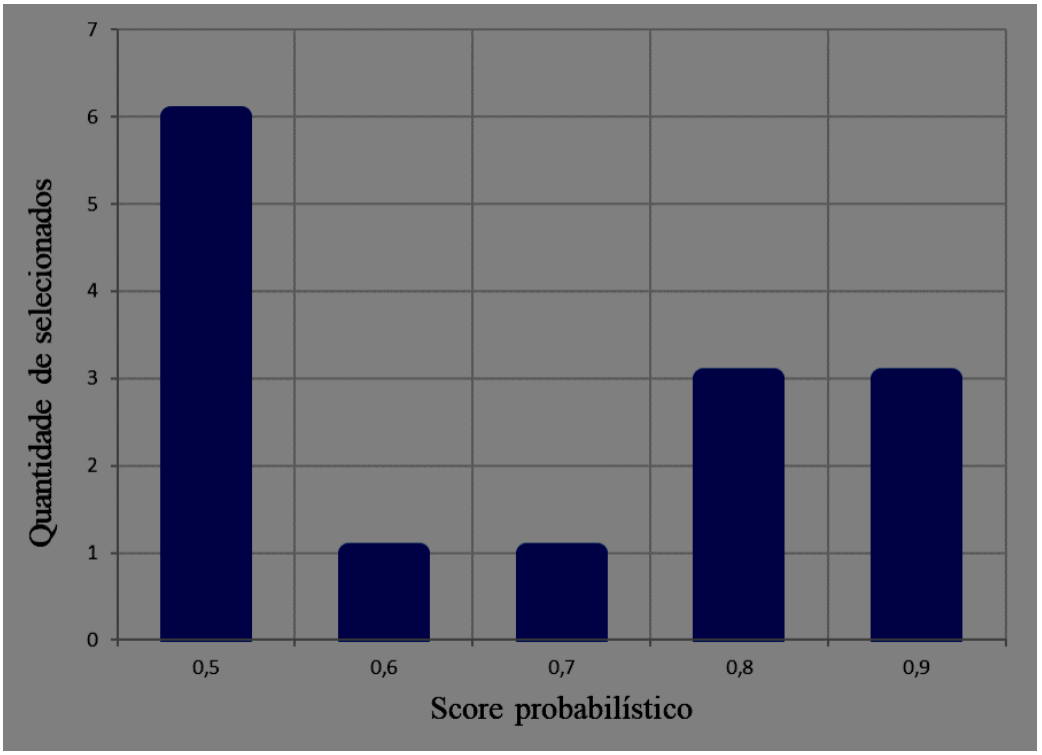
Fonte: Autoria própria, 2023.

5 RESULTADOS E DISCUSSÃO

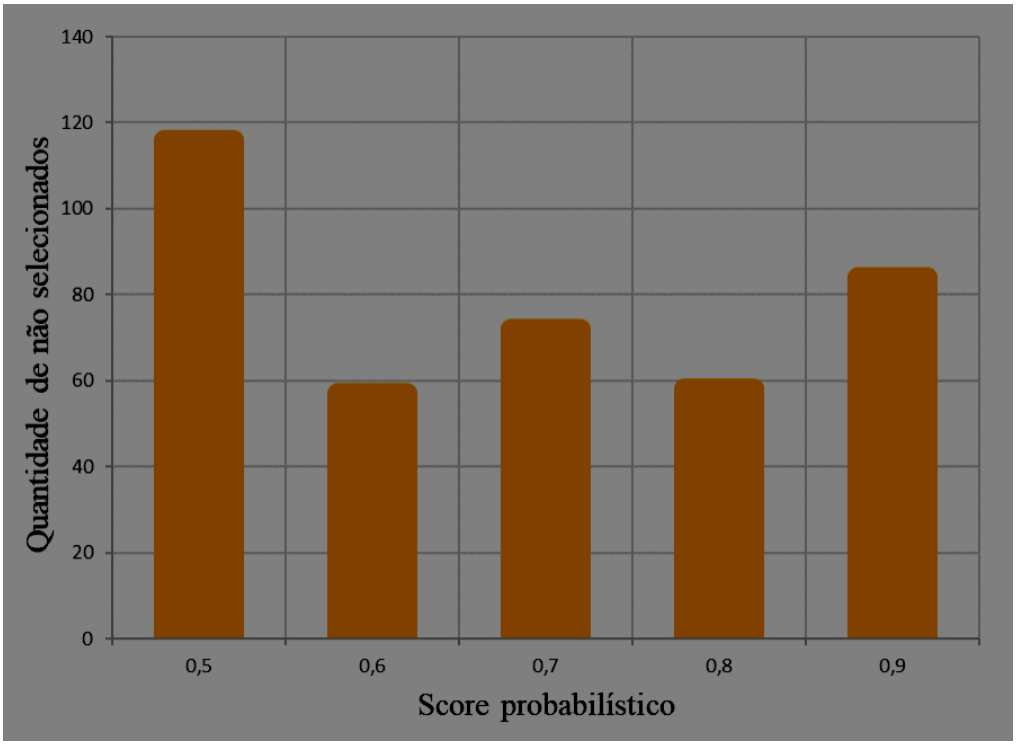
Um total de 466 referências foram recuperadas e classificadas pelo LitSuggest com score > 0,5. Dentre as classificadas, ao fim da curadoria manual, 452 foram descartadas (não selecionadas) por não apresentarem os dados requeridos e 14 foram selecionadas por apresentarem os dados. No gráfico 1-A, estão apresentadas as quantidades de referências selecionadas por faixa de score atribuído pelo LitSuggest, assim como no gráfico 1-B, estão apresentadas as quantidades de referências não selecionadas para cada faixa.

Gráfico 1 - Quantidade de referências selecionadas e não selecionadas por faixa de score

A



B



Fonte: Autoria própria, 2023.

O comportamento descrito pelos gráficos mostra que não há relação entre os classificados e seus respectivos scores. Esperava-se que o comportamento fosse crescente, de modo que em faixas mais altas tivessem com as maiores quantidades de referências selecionadas. E de forma inversa, às faixas mais baixas deveriam apresentar maiores quantidades de referências não selecionadas.

A utilização do score para seleção das referências é de suma importância para redução do tempo gasto no processo de busca. E para que esse critério possa ser aplicado, uma análise do perfil dos dados que foram minerados nessa pesquisa pode ser suficiente para identificar termos-chaves, tags e características que sendo adicionadas na etapa de treinamento do modelo de ML, podem contribuir para que os classificados e selecionados sejam restritos a faixas mais altas de score.

Explorando o perfil dos dados minerados a partir do Gráfico 2, que apresenta as mutações *missenses* separadas filogeneticamente, é importante notar que a observância das tendências e demandas científicas, são uma variável imprescindível para a adaptação do modelo e atualização dos termos de busca, pois das mutações mineradas pela ferramenta, cerca de 23% partem de uma proteína selvagem de origem animal, com destaque para *Homo sapiens* como organismo mais frequente, porém estão presentes também várias descrições de

mutações pontuais de origem viral, que correspondem a 5,5% do total. Considerando o período pandêmico vivenciado a pouco tempo pela humanidade, as referências selecionadas que descrevem mutações de origem viral, são todas relacionadas a proteína Spike do SARS-CoV-2, enquanto parte das mutantes animais, são relacionadas ao receptor humano ACE2, que é uma proteína de membrana com domínio catalítico que se liga a Spike.

Gráfico 2 - Distribuição taxonômica dos organismos de origem das proteínas selvagens e mutantes mineradas pela ferramenta LitSuggest

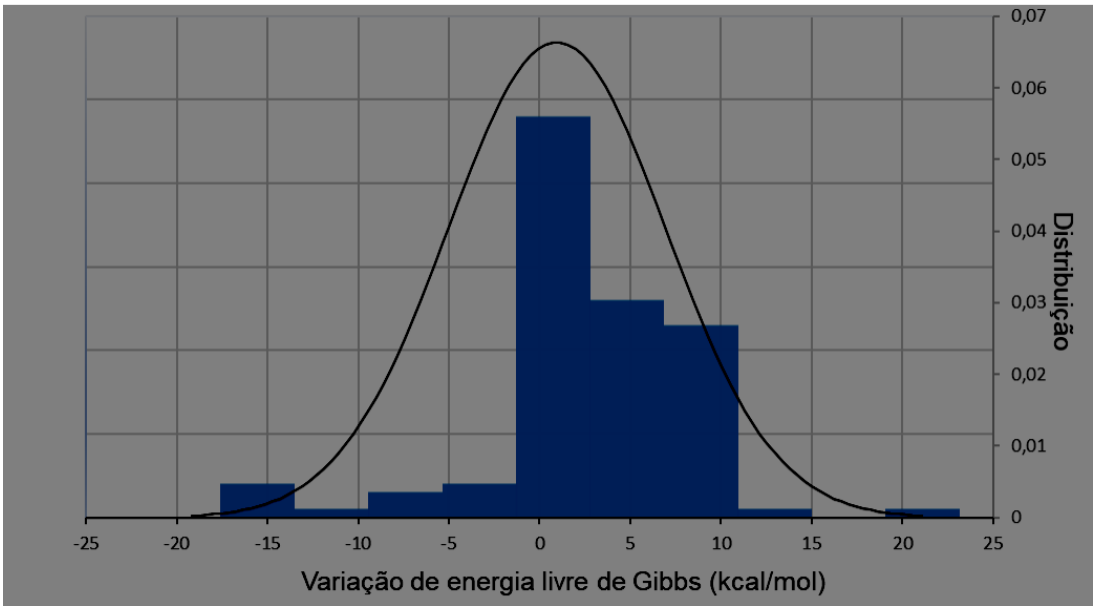
Fonte: Autoria própria, 2023.

Portanto, é necessário que considerações sejam feitas para que os filtros que dão suporte a seleção de referências sejam regularmente atualizados, para que a recuperação corresponda também às tendências do momento.

A outra parcela das mutantes de origem animal e também as mutantes de origem sintética (8,8%) apresentadas no gráfico 2, se relacionam por estarem atreladas a pesquisas que buscam o desenvolvimento de terapias gênicas, principalmente para doenças contemporâneas resultantes de alterações nos genes e retardo de consequências morfofuncionais. As bactérias que estão representadas pela *E. coli*, correspondem a maior parcela das mutantes mineradas (50,7%) e também se relacionam com a parcela de mutantes de origem animal e sintéticos, uma vez que ela é utilizada como *cell factory* (organismo utilizado para produzir proteínas) para expressar proteínas humanas, como a insulina. E uma outra parcela de 12% do total representa os protozoários e reúnem descrições sobre mutações em proteínas específicas utilizadas em pesquisas sobre microgravidade espacial.

Com relação aos dados de estabilidade termodinâmica que foram minerados, a fim de identificar assimetrias e descontinuidades, histogramas foram plotados separadamente para os conjuntos ΔΔG e ΔTm, resultando nas distribuições apresentadas nos Gráficos 3 e 4, respectivamente. Ambas as distribuições são unimodais, apresentando apenas um pico proeminente. Isso indicou que o comportamento dos valores plotados se assemelha com a distribuição normal, por isso uma função de densidade foi utilizada junto ao histograma para gerar uma curva de distribuição para análise das medidas de centro. A média aritmética dos valores de ΔΔG é igual a 0,95 e dos valores de ΔTm é igual a -1,05.

Gráfico 3 - Distribuição dos efeitos termodinâmicos em mutantes pela variação da Energia Livre de Gibbs (ΔΔG)

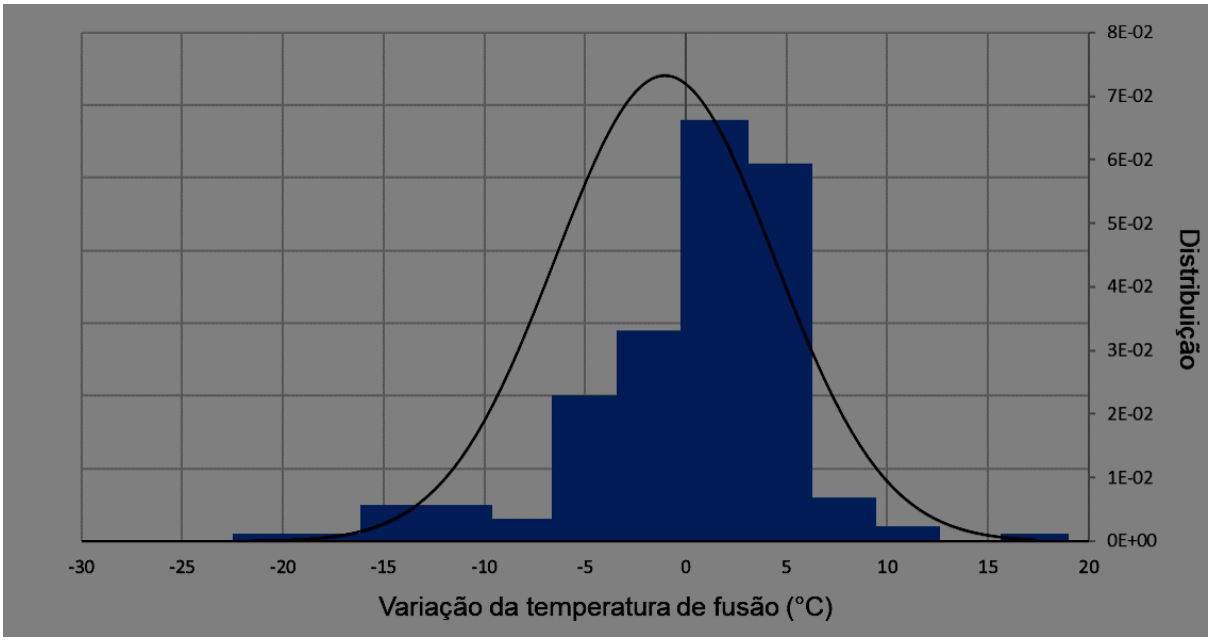


Fonte: Autoria própria, 2023.

A dispersão de 111 valores de ΔΔG plotados no Gráfico 3, demonstra uma assimetria com pico entre -1 e 2, tendo a maior frequência de dados descritos entre o intervalo -1 e 11. Esse tipo de distribuição mostra que valores positivos são descritos e melhor representados, uma vez que eles significam a quantidade de proteínas mutantes estáveis. E já que a maior parte das referências mineradas estão relacionados à busca de alterações favoráveis a aplicações na saúde, as mutações desestabilizadoras (ΔΔG < 0) estão descritas em menor quantidade (valores abaixo da média), pois a baixa estabilidade resulta em moléculas desdobradas e pouco funcionais e consequentemente, de menor interesse, enquanto a alta

estabilidade resulta em moléculas com estruturas mais dobradas, rígidas e de maior resistência.

Gráfico 4 - Distribuição dos efeitos termodinâmicos em mutantes pela variação da temperatura de fusão (ΔTm)



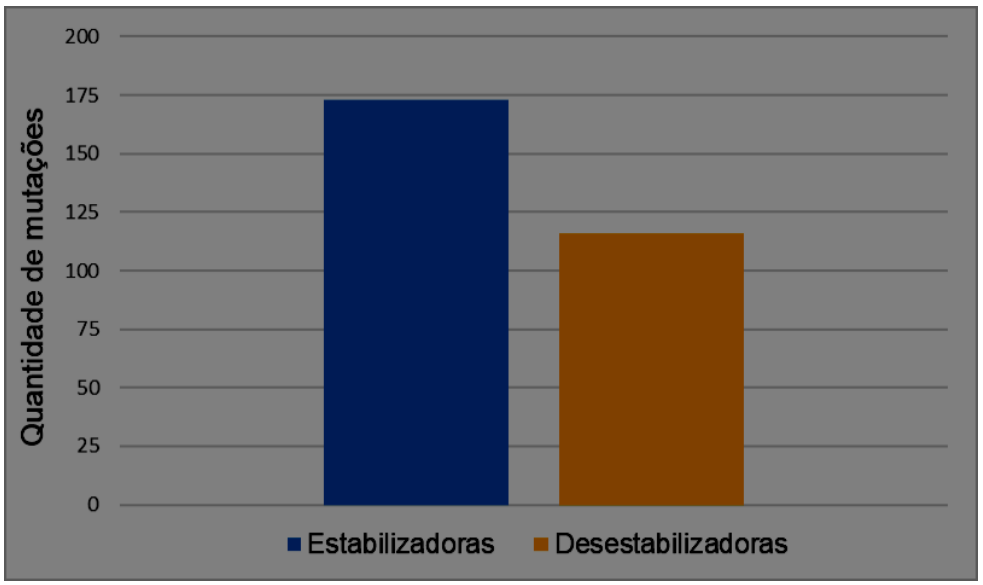
Fonte: Autoria própria, 2023.

Os 185 valores de ΔTm plotados no Gráfico 4, apresentam uma distribuição de comportamento assimétrico com relação à normalidade, com pico proeminente entre -1,6 e 4,5. É possível observar que há uma diferença na distribuição dos valores de variação de temperatura com relação aos valores de variação de energia, enquanto o relato da estabilidade dos mutantes por ΔΔG é mais atraente para as aplicações que relacionam biotecnologia e saúde, a descrição de mutantes a partir da termoestabilidade tem oferecido melhores evidências para seleção de proteínas mais estáveis. E apesar dessas mutações mineradas também estarem associadas a doenças e patogenicidades, a estabilização térmica de mutantes para aplicação em processos industriais é um dos grandes interesses da biotecnologia.

Uma análise quantitativa do efeito das mutações mineradas (Gráfico 5) é de suma importância para métodos computacionais construídos com base nessas informações, que predizem efeitos de mutações estabilizadoras, as quais correspondem a 60%, nesta pesquisa. No Gráfico 5 também verifica-se um percentual de 40% de mutações desestabilizadoras que foram mineradas e que evidenciam a necessidade de flexibilidade ao ajustar o modelo de ML, adicionando mais referências ao conjunto de treino positivo, que considerem a pluralidade

mutações missenses, uma vez que valores de estabilidade termodinâmica podem envolver assuntos aparentemente opostos e ocasionar a atribuição de um score probabilístico não condizente.

Gráfico 5 - Efeitos das mutações mineradas pela ferramenta LitSuggest



Fonte: Autoria própria, 2023.

Assim, foi possível recuperar 283 novas descrições de mutações *missenses* que correspondem a um aumento de 1,94% no número de mutações que estarão disponíveis na base de dados e em termos unitários, foram um total 2.901 novos dados minerados, correspondendo a um aumento de 19,8%.

6 CONCLUSÕES

A análise do perfil dos dados, permite inferir que a ferramenta LitSuggest pode ser utilizada como meio de recuperação de literatura que contenha dados termodinâmicos de proteínas. Também, mostra a necessidade de calibração do modelo de ML que foi treinado, considerando que os conjuntos de treino devem ser aprimorados.

Com a recuperação de referências selecionadas correspondendo a 3,14% de todas as classificadas, outra curadoria deve ser aplicada após a calibração do modelo, para que a automatização da busca seja confirmada.

Apesar de um baixa recuperação de referências pelo atual modelo, as 283 novas mutações que foram mineradas, serão submetidas ao ThermoMutDB, promovem um aumento no número de mutações *missenses,* dados experimentais e literários. Estes podem ser acessados atualmente pelo link: https://biosig.lab.uq.edu.au/thermomutdb/

6.1 TRABALHOS FUTUROS

Dentre as possibilidades de estudo que podem ser realizados futuramente, temos:

1. Após a validação da ferramenta para selecionar as referências que contenha os dados termodinâmicos, uma nova IA pode ser desenvolvida para processamento de linguagem natural, com o intuito de automatizar a mineração dos dados e avaliando a possibilidade de automatizar o processo de mineração e curadoria de dados termodinâmicos;
2. Com a base de dados ThemoMutDB sendo atualizada e se configurando a maior do seu nicho, o aproveitamento das informações para desenvolvimento de novos preditores mais assertivos seria intuitivo, para os próximos passos.

REFERÊNCIAS

ALBERT, B. et al. (Orgs.). **Biologia Molecular da Célula** - 6° Ed. 2017, Ed. Artes Médicas, Porto Alegre.

ALLOT, A. et al. LitSuggest: a web-based system for literature recommendation and curation using machine learning. **Nucleic Acids Research**, v. 49, n. W1, p. W352–W358, 2 Jul. 2021.

ALVES, E. A.; SOUZA, D. S. **Biologia molecular**. *In*: MOLINARO, Etelcia Moraes; CAPUTO, Luzia Fátima Gonçalves; AMENDOEIRA, Maria Regina Reis (Org.). Conceitos e métodos para a formação de profissionais em laboratórios de saúde. v.3. Rio de Janeiro: EPSJV, 2013. p. 134-185.

ANFINSEN, C. B. Principles that govern the folding of protein chains. **Science**, v. 181, n. 4096, p. 223–230, 20 Jul. 1973.

CECCATTO, V. M. **Biologia Molecular**. 1. ed. Fortaleza: Editora da UAB, 2010. v. 1. 124p

DALL'AGNO, K.C. M. **Um estudo sobre a predição da estrutura 3D aproximada de proteínas utilizando o método CReF com refinamento.** 2012. Dissertação de Mestrado. Pontifícia Universidade Católica do Rio Grande do Sul.

DEVLIN, T.M. Manual de Bioquímica com Correlações Clínicas, 7ª ed., Ed. Blucher, 2011.

DEHGHANPOOR, R. et al. Predicting the effect of single and multiple mutations on protein structural stability. Molecules (Basel, Switzerland), v. 23, n. 2, 27 Jan. 2018.

DOMINGUES, M. L. C. S. Mineração de dados utilizando aprendizado não-supervisionado: um estudo de caso para bancos de dados da saúde. 2003.

ENCINAS PONCE, L. F. Determinantes e forças seletivas na evolução das proteínas. 2014.

FRANCISCO J. W. E.; RANCISCO, W. Proteínas: Hidrólise, Precipitação e um Tema para o Ensino de Química. Química Nova na Escola , v. 24, p. 12-16, 2006.

FREITAS, E. K. H. DE. Aplicações de ensemble learning para o estudo do efeito de mutações pontuais em estruturas tridimensionais de proteínas. 2020.

OLIVEIRA, G. T. H.; SANTOS, N. F. DOS; BELTRAMINI, L. M. O DNA: uma sinopse histórica. Revista de Ensino de Bioquímica, v. 2, n. 1, p. 1, 20 Dec. 2004.

GERASIMA VICIUS, L.; LIU, X.; MARSH, J. A. Identification of pathogenic missense mutations using protein stability predictors. Scientific Reports, v. 10, n. 1, p. 15387, 21 Sep.

2020.

GERSHENSON, A. et al. Successes and challenges in simulating the folding of large proteins. The Journal of Biological Chemistry, v. 295, n. 1, p. 15–33, 3 Jan. 2020.

GUIDO, R. V. C.; ANDRICOPULO, A. D.; OLIVA, G. Planejamento de fármacos, biotecnologia e química medicinal: aplicações em doenças infecciosas. Estudos Avançados, v. 24, n. 70, p. 81–98, 2010.

HUA, L. et al. Urea denaturation by stronger dispersion interactions with proteins than water implies a 2-stage unfolding. Proceedings of the National Academy of Sciences of the United States of America, v. 105, n. 44, p. 16928–16933, 4 Nov. 2008.

IQBAL, S. et al. Assessing the performance of computational predictors for estimating protein stability changes upon missense mutations. Briefings in Bioinformatics, v. 22, n. 6, 5 Nov. 2021.

LIBÓRIO, L.; RESENDE, V. H. Introdução aos bancos de dados biológicos. In: MARIANO, **D. et al. (Eds.). . BIOINFO - Revista Brasileira de Bioinformática e Biologia Computacional.** Alfahelix, 2021.

MALTA, L. H. A.; KUROIVA, M. Antonio R. L.. Aprendizado de máquina e processamento de linguagem natural aplicados à identificação de discurso de ódio. 2019. Universidade de Brasília, Brasília, 2019.

Marques, F. B. Predição da estrutura tridimensional de proteínas utilizando o método CReF com informações de contato. Dissertação (Mestrado) - Programa de Pós-Graduação em Ciências da Computação, PUCRS. Porto Alegre. 2021.

MARTINS, L. Bateson e o programa de pesquisa mendeliano. Episteme, v. 14, n. 1, p. 27–55, 2002.

MELGAÇO, L. S. V. Desenvolvimento de sistema de publicação e rotulagem de dados para o ThermoMutDB. 2021. Trabalho de Conclusão de Curso (Graduação em Engenharia de Computação) - Universidade Federal de Itajubá.

MCGUINNESS, K. N. et al. Role of simple descriptors and applicability domain in predicting change in protein thermostability. Plos One, v. 13, n. 9, p. e0203819, 7 Sep. 2018.

NELSON, D.L.; COX, M.M. Princípios de Bioquímica de Lehninger. 6. ed. São Paulo: Artmed, 2014.

PÁL, C.; PAPP, B.; LERCHER, M. J. An integrated view of protein evolution. Nature Reviews. Genetics, v. 7, n. 5, p. 337–348, May 2006.

PRIVALOV, P. L.; KHECHINASHVILI, N. N. A thermodynamic approach to the problem of stabilization of globular protein structure: a calorimetric study. Journal of Molecular Biology, v. 86, n. 3, p. 665–684, 5 Jul. 1974.

RIBEIRO, M. C. M. GENÉTICA MOLECULAR. 2. ed. Florianópolis: BIOLOGIA/EAD/UFSC, 2014.

SENIOR, A. W.; EVAN, R.; JUMPER, J.; Kirkpatrick, J.; SIFRE, L.; GREEN, T.; QIN, C.; ŽíDEK, A.; NELSON, A. W.; Bridgland, A. Improved protein structure prediction using potentials from deep learning. Nature, vol. 577, pp. 706–710. jan. 2020.

Waizbort, R. & Solha, G. C. F. Os genes interrompidos: o impacto da descoberta dos íntros sobre a definição de gene molecular clássico. Revista da Sociedade Brasileira de História da Ciência, 5, p. 63-82, 2007.

TRYER, L. Bioquímica. 4 ed. Rio de Janeiro: Guanabara-Koogan, 1996. VOET, D.; VOET, J. G.; PRATT, C. W. Fundamentos de Bioquímica. Porto Alegre, Editora Artmed, 2000. TANOUYE, F. T. Enovelamento de proteínas e ligações de hidrogênio - estudo de modelos mínimos. 22 Sep. 2017.

TORRES-FREIRE, C.; GOLGHER, D.; CALLIL, V. Biotecnologia em saúde humana no Brasil: produção científica e pesquisa e desenvolvimento. Novos Estudos - CEBRAP, n. 98, p. 69–93, Mar. 2014.

VERLI, H. (Org). Bioinformática : da biologia à flexibilidade molecular. Sociedade Brasileira de Bioquímica e Biologia Molecular, 2014.

XAVIER, J. S. et al. ThermoMutDB: a thermodynamic database for missense mutations. Nucleic Acids Research, v. 49, n. D1, p. D475–D479, 8 Jan. 2021.

ANEXO

Anexo contendo as referências recuperadas pelo LitSuggest : ANEXO.01

Anexo contendo os dados minerados nas referências selecionadas: