PUCRS

ESCOLA POLITÉCNICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO
MESTRADO EM CIÊNCIA DA COMPUTAÇÃO

FLAVIELLE BLANCO MARQUES

PREDIÇÃO DA ESTRUTURA TRIDIMENSIONAL DE PROTEÍNAS UTILIZANDO O MÉTODO CReF COM INFORMAÇÕES DE CONTATO

Porto Alegre
2021

PÓS-GRADUAÇÃO - *STRICTO SENSU*

Pontifícia Universidade Católica
do Rio Grande do Sul

PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL
ESCOLA POLITÉCNICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

PREDIÇÃO DA ESTRUTURA TRIDIMENSIONAL DE PROTEÍNAS UTILIZANDO O MÉTODO CReF COM INFORMAÇÕES DE CONTATO

FLAVIELLE BLANCO MARQUES

Dissertação apresentada como requisito parcial à obtenção do grau de Mestra em Ciência da Computação na Pontifícia Universidade Católica do Rio Grande do Sul.

Orientador: Prof. Dr. Rafael Heitor Bordini
Co-Orientador: Prof. Dr. José Fernando Ruggiero Bachega

Porto Alegre
2021

Ficha Catalográfica

M357p Marques, Flavielle Blanco

Predição da estrutura tridimensional de proteínas utilizando o método CReF com informações de contato / Flavielle Blanco Marques. – 2021.

103 f.

Dissertação (Mestrado) – Programa de Pós-Graduação em Ciência da Computação, PUCRS.

Orientador: Prof. Dr. Rafael Heitor Bordini.

Co-orientador: Prof. Dr. José Fernando Ruggiero Bachega.

1. Bioinformática. 2. Predição de Estruturas de Proteínas. 3. Contato. I. Bordini, Rafael Heitor. II. Bachega, José Fernando Ruggiero. III. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da PUCRS com os dados fornecidos pelo(a) autor(a).

Bibliotecária responsável: Clarissa Jesinska Selbach CRB-10/2051

FLAVIELLE BLANCO MARQUES

PREDIÇÃO DA ESTRUTURA TRIDIMENSIONAL DE PROTEÍNAS UTILIZANDO O MÉTODO CReF COM INFORMAÇÕES DE CONTATO

Dissertação apresentada como requisito parcial para obtenção do grau de Mestra em Ciência da Computação do Programa de Pós-Graduação em Ciência da Computação, Escola Politécnica da Pontifícia Universidade Católica do Rio Grande do Sul.

Aprovado(a) em 30 de Abril de 2021.

BANCA EXAMINADORA:

Prof. Dr. Luis Fernando Saraiva Macedo Timmers (Univates)

Prof. Dr. Duncan Dubugras Alcoba Ruiz (PPGCC/PUCRS)

Prof. Dr. José Fernando Ruggiero Bachega (UFCSPA- Co-Orientador)

Prof. Dr. Rafael Heitor Bordini (PPGCC/PUCRS - Orientador)

DEDICATÓRIA

Dedico este trabalho aos meus pais, Nara e Flavio, a meu irmão, Allan e a minha cachorrinha, Kelly (*in memoriam*).

“I was taught that the way of progress was neither swift nor easy.”
(Marie Curie)

AGRADECIMENTOS

Primeiramente, gostaria de agradecer a todos aqueles que contribuíram de algum modo para este trabalho. Agradeço pelo suporte intelectual e emocional que eu tive em um dos momentos de maior instabilidade e crescimento, pessoal e profissional, da minha vida.

Ao professor Dr. Osmar Norberto de Souza, pelo privilégio de suas aulas, orientações, conversas, discussões e convivência por quase todo o mestrado. Entendo que ciclos se fecham e agradeço imensamente pelos ensinamentos e pela confiança depositada em mim para com o desafiador CReF. Obrigada por ter acreditado e visto potencial, o qual muitas vezes eu mesma não o vi.

Ao professor Dr. Rafael Heitor Bordini, por ter me acolhido tão abertamente em seu grupo de pesquisa e por estar sempre disponível para ajudar e orientar.

Aqui, guardo um parágrafo dedicado de coração ao professor Dr. José Fernando Ruggiero Bachega, meu coorientador. Se hoje eu sou uma pesquisadora melhor, que acredito que sou, foi graças às nossas longas conversas, críticas e discussões. A ciência precisa de mais pessoas como você. Obrigada por ter trazido luz e otimismo em tempos sombrios, obrigada por entender as minhas ideias, propostas, alucinações e ainda guiar-me com maestria ao longo deste trabalho.

Aos colegas e amigos do LABIO, especialmente, Gabriel Fernandes Leal, Lucas Santos Chitolina e Vanessa Stangherlin Machado Paixão-Cortes. Vocês foram parte fundamental deste trabalho. Obrigada por estarem ao meu lado.

Aos colegas e amigos que tive a honra de conhecer e ter em minha vida graças ao PPGCC. Obrigada por todas as conversas, risadas e auxílios. Não os citarei aqui com receio de esquecer alguém, mas saibam que os agradeço profundamente.

Aos poucos, mas verdadeiros amigos que tenho desde a infância, adolescência ou graduação. Obrigada por entenderem minha ausência e por sempre torcerem por mim.

Aos professores e membros da secretaria do PPGCC que ajudaram de algum modo à minha formação, meu muito obrigada.

Aos professores e membros da banca, que me acompanharam ao longo deste trabalho, obrigada pelas críticas e sugestões que tanto contribuíram para tornar esta dissertação melhor.

A Kelly, minha cachorrinha, que partiu durante este trabalho. Obrigada por ter estado presente na minha vida por 14 anos. Obrigada por ter me ensinado o significado de cuidado, amor e morte.

Por fim, e não menos importante, a minha família, que desde sempre priorizaram os meus estudos. Se estou aqui hoje é graças a vocês. Serei eternamente grata.

PREDIÇÃO DA ESTRUTURA TRIDIMENSIONAL DE PROTEÍNAS UTILIZANDO O MÉTODO CReF COM INFORMAÇÕES DE CONTATO

RESUMO

O modo como uma proteína se enovela é um problema de rápida solução na natureza. Entretanto, mapear todas as possibilidades conformacionais que uma sequência de resíduos de aminoácidos pode assumir não é uma tarefa trivial. Desta maneira, o problema da predição de estrutura apresenta diferentes métodos computacionais e soluções aproximadas. Entre esses, o CReF (*Central Residue Fragment-based method*). Nos últimos anos, abordagens que utilizam informações de contato entre resíduos apresentaram resultados promissores para a predição da estrutura 3D. Assim, nesta dissertação, a incorporação das informações de contato ao método CReF foi realizada. Para tanto, alterações funcionais e metodológicas foram necessárias. Adicionalmente, questões relacionadas à avaliação e a amostragem das conformações foram cruciais para que as informações de contatos pudessem ser utilizadas. Para o primeiro, um modelo de $RMSD_{predito}$ baseado em um potencial atômico dependente de distância, contatos de curto, médio e longo alcance foi desenvolvido. Para o segundo, um módulo de simulação molecular baseado em *simulated annealing* foi acoplado. Ademais, uma função baseada nos mesmos termos do modelo de predição de RMSD foi proposta com o objetivo de mimetizar uma função de energia. Um conjunto de proteínas foi avaliado, o qual mostrou que ao se ter as informações de contato, é possível chegar em conformações de baixa energia e baixo RMSD. Assim, apresentamos uma nova versão para o método CReF com informações de contato, uma função para calcular a energia do sistema e um método de simulação molecular.

Palavras-Chave: Bioinformática, Predição de Estruturas de Proteínas, Contato.

PROTEIN STRUCTURE PREDICTION USING CReF METHOD WITH CONTACT INFORMATION

ABSTRACT

Protein folding is a problem with a quick solution in nature. However, look for all conformations that a sequence of amino acid residues can assume is not a trivial task. In this way, several predictors are trying to create solutions for protein structure prediction. One of these is the CReF (*Central Residue Fragment-based method*). In the last years, approaches using the inter-residue contact information had promissory results to protein structure prediction. Thus, we incorporated contact information into the CReF method. However, functionals and methodological changes were necessary. Questions about conformation evaluation and sampling were central to contact information uses were effective. In the first, we created an $RMSD_{predict}$ model. It uses distance-dependent atomic potential and short, medium, and long-range inter-residue contacts. In the second, a molecular simulation module based on simulated annealing was coupling. Additionally, we proposed a function based on the same terms in the model. It behaves such as function energy. Furthermore, we simulated and evaluated a protein sample. If the inter-residue contact information is available, it is possible to find protein conformations with low energy and RMSD. Hence, we created a CReF new version with contact information, a function to calculate the energy system, and a simulation method.

Keywords: Bioinformatics, Protein Structure Prediction, Contact.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

3D – Tridimensional

AA – Aminoácido

AM – Aprendizado de Máquina

BE – Bioinformática Estrutural

BLASTp – *Protein-Protein Basic Local Alignment Search Tool Protein*

CASP – *Critical Assessment of Structure Prediction*

CF – Campo de Força

CMA – *Correlated Mutation Analysis*

CPC – Condições Periódicas de Contorno

CReF – *Central Residue Fragment-based method*

DNA – Ácido Desoxirribonucleico

DM – Dinâmica Molecular

DSSP – *Dictionary of Protein Secondary Structure*

EM – *Expectation Maximization*

FC – *Fully Connected*

FM – *Free Modeling*

GDT – *Global Distance Test*

GPU – *Graphics Processing Unit*

KCAL – Quilocaloria

LABIO – Laboratório de Bioinformática, Modelagem e Simulação de Biossistemas

MAE – *Mean Absolute Error*

MC – Monte Carlo

MMC – Método de Monte Carlo

mRNA – Ácido Ribonucleico Mensageiro

MSA – *Multiple Sequence Alignment*

MSE – *Mean-Squared Error*

RMSD – *Root Mean Square Deviations*

RNM – Ressonância Nuclear Magnética

PAM – *Point Accepted Mutation*

PDB – *Protein Data Bank*

pH – Potencial Hidrogeniônico

PSI-BLAST – *Position-Specific Iterative Basic Local Alignment Search Tool*

PSP – *Protein Structure Prediction*

ReLU – Unidade Linear Retificada

RNA – Ácido Ribonucleico

RR – Resíduo-Resíduo

RW – *Random-Walk*

TBM – *Template-Based Modeling*

TPU – *Tensor Processing Unit*

LISTA DE ABREVIATURAS

Est. – Estrutura

Frag. – Fragmento

Id. – Identidade

ID. – Identificador

Pos. – Posição

Res. – Resíduo

Sec. – Secundária

Seq. – Sequência

LISTA DE SÍMBOLOS

SUMÁRIO

1. INTRODUÇÃO

As proteínas são macromoléculas biológicas responsáveis pela maior parte da estrutura e atividade dos organismos. Elas são abundantes e participam dos mais diversos processos bioquímicos (Lesk, 2008). Determinar a estrutura tridimensional das proteínas e as conformações que podem assumir são a chave para entender seus papéis nos inúmeros processos biológicos, médicos ou farmacêuticos (Ma, 2015).

Uma proteína ou polipeptídeo é composta por uma sequência de resíduos de aminoácidos conhecida por estrutura primária. Ao longo do tempo, milhões de proteínas tiveram sua sequência determinada e depositada no *GenBank*[1] (cerca de 351 milhões em 05 de fevereiro de 2021) considerando apenas proteínas não redundantes, ou seja, aquelas que possuem apenas uma entrada no banco de dados (Benson et al., 2005). Porém, no mesmo dia, apenas 174.507 proteínas tinham a estrutura tridimensional[2] (3D) ou terciária determinada experimentalmente e depositada no *Protein Data Bank* (PDB)[3] (Berman et al., 2000). As proteínas depositadas no PDB são identificadas através de um código de quatro caracteres alfanuméricos (*e.g.* 1ZDD) em que para cada chave de acesso é possível encontrar informações sobre a estrutura, os artigos associados e dados experimentais.

Apesar do desenvolvimento e avanços de técnicas experimentais como difração por raio-X (Rupp, 2009), ressonância nuclear magnética (RNM) (Wüthrich, 1986) e microscopia eletrônica (Baumeister e Steven, 2000), muitas estruturas terciárias não são determinadas por tais técnicas (Maggio e Ramnarayan, 2001). O tempo de desenvolvimento, a complexidade e a limitação das técnicas e o alto custo são fatores que impossibilitam que todas as estruturas primárias tenham a estrutura terciária correspondente determinada (Jana et al., 2018; Sousa et al., 2006). Sendo assim, a diferença entre estruturas primárias e terciárias depositadas nos bancos de dados indica o quão necessário é encontrar métodos que possam determinar a estrutura 3D das proteínas.

A fim de reduzir a lacuna entre estruturas primárias e terciárias conhecidas, métodos computacionais estão sendo cada vez mais usados para auxiliar na predição das estruturas 3D (Källberg et al., 2014; Rohl et al., 2004; Roy et al., 2010; Venkatesan et al., 2013; Jumper et al., 2020). O processo, os fatores e as informações que possibilitam a formação da estrutura 3D, a partir da estrutura primária, constituem um dos maiores desafios da Bioinformática Estrutural e uma das questões centrais da Biologia Molecular. O problema da predição de estrutura de proteínas (PSP – do inglês, *Protein Structure Prediction*) visa entender como a partir da sequência de resíduos de aminoácidos é obtida a estrutura tridimensional de uma proteína (Lesk, 2001; Zhang, 2008).

[1] https://www.ncbi.nlm.nih.gov/genbank/ acesso em: 05 de fevereiro de 2021.
[2] Estrutura terciária, estrutura 3D e estrutura tridimensional são consideradas sinônimas.
[3] https://www.rcsb.org/ acesso em: 05 de fevereiro de 2021.

Diversos métodos e algoritmos foram propostos desde o surgimento do PSP. Os métodos podem ser separados em dois grupos. O primeiro grupo engloba métodos de modelagem comparativa (Martí-Renom et al., 2000) e *fold recognition* (Bowie et al., 1991; Jones et al., 1992). O segundo grupo engloba métodos *ab initio* (Osguthorpe, 2000) e *de novo* (Bowers et al., 2000; Srinivasan e Rose, 1995), os quais permitem que novos padrões de enovelamento possam ser encontrados.

Entre os métodos, *ab initio* utiliza somente a sequência de resíduos de aminoácidos e não necessita de informações contidas nas estruturas já conhecidas e depositadas nos bancos de dados. Por outro lado, os métodos baseados em modelagem comparativa utilizam estruturas já conhecidas, fornecendo resultados mais precisos e predições mais confiáveis (Baker e Sali, 2001; Fiser et al., 2002; Tramontano, 1998).

Adicionalmente, os métodos possuem limitações tanto no entendimento dos processos biológicos envolvidos como na complexidade, pois mesmo uma pequena molécula torna o problema da predição inacessível computacionalmente (Paradoxo de Levinthal) (Levinthal, 1968). Além disso, as diferentes posições que os resíduos de aminoácidos podem ocupar no espaço conformacional leva o problema do PSP a explosão combinatória, sendo o problema pertencente à classe NP-completo (Berger e Leighton, 1998; Crescenzi et al., 1998). Entretanto, o grande avanço computacional permitiu que soluções cada vez mais aproximadas fossem alcançadas.

Em virtude da complexidade e a fim de estabelecer o atual estado da arte na predição de estruturas proteicas, o *Critical Assessment of Structure Prediction* (CASP) foi criado. O CASP identifica os progressos e destaca onde os esforços futuros podem ser concentrados, a fim de resultar em predições cada vez mais próximas às experimentais (Kryshtafovych et al., 2019). O CASP ocorre a cada dois anos desde 1994 e o número de alvos disponíveis, grupos engajados, métodos de avaliação e categorias aumentaram significativamente até agora, assim como, a robustez dos preditores (Kryshtafovych et al., 2019). É importante destacar o resultado mais surpreendente e satisfatório obtido no CASP14 ocorrido em 2020, pelo *AlphaFold*, o qual atingiu predições com *Global Distance Test* (GDT) no valor de 90 (sendo 100 a correspondência exata da predição com a estrutura 3D nativa) o qual, informalmente, é considerada competitiva com os resultados obtidos experimentalmente (AlphaFold, 2020; Jumper et al., 2020).

Com a relevância do problema e com o intuito de construir um novo preditor, híbrido, que compartilhasse os bons resultados dos métodos baseados em modelagem comparativa, mas que permitisse encontrar novos padrões de enovelamento como nos métodos de predição *de novo*, o *Central Residue Fragment-based method* (CReF) foi proposto por Dorn e de Souza (2008), no Laboratório de Bioinformática, Modelagem e Simulação de Biossistemas (LABIO). O método consiste na predição tridimensional aproximada de uma proteína, utilizando como dados de entrada somente a estrutura primária. A versão inicial

do CReF envolvia diversas etapas manuais, atribuindo ao usuário toda a responsabilidade pelo funcionamento do método e que foram automatizadas por da Motta Dall'Agno (2012).

Apesar dos resultados promissores, problemas no enovelamento das proteínas ocorreram pela não formação de ligações de hidrogênio e organização inadequada da estrutura no espaço 3D. Essa desorganização ocorreu por dificuldades na modelagem de estruturas irregulares das proteínas como voltas e alças (Fiser et al., 2000). Na busca pela melhoria na predição das regiões irregulares e das estruturas em geral, um protocolo de refinamento, utilizando simulação por Dinâmica Molecular (DM) foi desenvolvido (Dall'Agno e Norberto de Souza, 2013; Soletti, 2015). As estruturas preditas pelo CReF foram utilizadas como dados de entrada para as simulações e apesar de resultados instigantes, as predições realizadas pelo CReF, necessitavam de maior precisão, para que as estruturas refinadas pela DM, pudessem apresentar estruturas mais próximos às experimentais.

No entanto, para simplificar computacionalmente o problema do PSP, o método de Monte Carlo (MMC) surgiu como alternativa. Assim, com MMC há uma busca pelo mínimo global através do valor de energia de uma conformação dentre uma gama de conformações existentes (Heilmann et al., 2020). Além disso, podendo ser utilizado para diferentes objetivos como refinamento (Lorenzen e Zhang, 2007), predição de cadeia lateral (Bhowmick e Head-Gordon, 2015) ou como método de busca pela conformação mais próxima da nativa (Heilmann et al., 2020; Okamoto, 2019; Zhang et al., 2007).

Adicionalmente, com o avanço dos estudos na área de predição de estrutura 3D de proteínas, o uso de informações relacionadas a contato entre resíduos de aminoácidos se tornou cada vez mais comum (Moult et al., 2018). A predição de contatos é uma área de estudo que avançou muito, na qual é possível predizer o quão próximo estão os resíduos na estrutura terciária dado a sequência de resíduos de aminoácidos, visto que aqueles que estão distantes na estrutura primária, podem estar próximos quando a proteína é enovelada (Amala e Emerson, 2019).

As predições de contato surgiram na década de 90 (Vendruscolo e Domany, 2000; Mirny e Domany, 1996), porém a taxa de falsos-positivos inviabilizavam o uso desse tipo de informação. Contudo, nos últimos anos o uso dos contatos apresentaram melhora significativa na predição da estrutura proteica em abordagens *de novo* (de Oliveira et al., 2017). Com isso, o aumento da confiabilidade das predições de contato e a possibilidade de utilizá-las, acelerando o processo de predição da estrutura 3D, tornou-se possível (Feng et al., 2020).

Desta maneira e a fim de melhorar as predições realizadas pelo CReF, identificamos a possibilidade de incorporar ao método informações relacionadas a contato entre os resíduos de aminoácidos unido a simulações por Monte Carlo. Assim, nesta dissertação, apresenta-se a nova versão do método CReF, a exploração das informações pré-existentes, a incorporação das informações de contato, uma função de seleção de conformação e um método de simulação molecular buscando manter as principais características do método original, mas alterando e aprimorando determinadas etapas.

1.1 Organização

Esta dissertação está organizada em 10 capítulos, seguidos de 4 apêndices:

- O primeiro capítulo introduz o problema de pesquisa e uma visão geral do que foi proposto neste trabalho.
- O segundo capítulo apresenta a fundamentação teórica e os conceitos relevantes para a compreensão do mesmo. Desse modo, é aprofundado o conceito de proteínas, os problemas associados à predição de suas estruturas, métodos computacionais para tratá-los, o CASP, o conceito de contatos, métodos para determiná-los, além dos métodos de simulação molecular e da métrica de avaliação mais conhecidos para proteínas.
- A motivação do trabalho, o objetivo geral e os objetivos específicos são apresentados no capítulo 3.
- A versão inicial do método CReF é representada no capítulo 4.
- No capítulo 5 é apresentada a metodologia associada à incorporação dos contatos ao método, ao modelo de $RMSD_{predito}$ e ao módulo de simulação molecular.
- O capítulo 6 mostra o estado inicial que o método CReF se encontrava, a investigação por problemas pré-existentes e os principais novos arquivos de saída do método.
- O capítulo 7 apresenta o desenvolvimento de um modelo de avaliação da qualidade que considera as informações de contato.
- No próximo capítulo, de número 8, é apresentada a proposta para resolver o problema de amostragem do método CReF com simulação molecular.
- A versão final do método CReF é mostrada no capítulo 9.
- As conclusões constam no capítulo 10. Essas apresentam as principais contribuições, com uma visão geral do que foi realizado, limitações e perspectivas.
- No final, localiza-se as referências necessárias para o desenvolvimento desta dissertação, seguida por 4 apêndices.

2. FUNDAMENTAÇÃO TEÓRICA

A bioinformática é uma área interdisciplinar que conecta a ciência da computação e as ciências biológicas (Xiong, 2006). Ela aplica técnicas derivadas da matemática, da ciência da computação e da estatística para entender e organizar informações relacionadas a macromoléculas biológicas como ácido desoxirribonucleico (DNA), ácido ribonucleico (RNA) e proteínas (Luscombe et al., 2001).

As aplicações da bioinformática podem ser divididas em duas classes. A primeira engloba aplicações relacionadas ao dogma central da biologia molecular, no qual o DNA é transcrito em RNA mensageiro (mRNA), que é traduzido em proteína e essa se enovela formando a estrutura tridimensional (Crick, 1970). A segunda classe é baseada em métodos científicos que criam hipóteses sobre a atividade biológica e as testam com diferentes experimentos (Gu e Bourne, 2009).

A bioinformática trabalha com uma grande quantidade de dados, muitas vezes complexos, que inviabilizam a realização de análises manuais (Nagaraj et al., 2018). Entre estes dados estão as estruturas proteicas depositadas em bases de dados como o PDB. Desta maneira, e considerando o número exacerbado de dados, a bioinformática estrutural (BE) surgiu como subdisciplina da bioinformática (Gu e Bourne, 2009). A BE é a área que estuda moléculas biológicas que apresentam estrutura, como por exemplo, as proteínas. Sendo assim, uma aplicação relacionada ao dogma central da biologia molecular e um dos grandes desafios da BE, ainda sem solução, mas com soluções aproximadas, é a predição tridimensional das proteínas (PSP).

2.1 Proteínas e sua Composição

As proteínas são polímeros lineares, sintetizados a partir de unidades monoméricas chamadas de aminoácidos (Voet e Voet, 2013). Do ponto de vista químico, as proteínas são as moléculas estruturalmente mais complexas e sofisticadas que conhecemos (Alberts et al., 2017). As proteínas estão relacionadas aos mais diversos processos biológicos e estão presentes em todas as células.

Um aminoácido (AA) é formado por um átomo de carbono quiral (Cα), o qual faz ligações com um grupo amino (-NH_2), um grupo carboxílico (-COOH) e um hidrogênio (H), que compõem a cadeia principal de uma proteína, e um grupo R também conhecido como cadeia lateral (Figura 2.1). Há 20 aminoácidos diferentes e um iminoácido (prolina) com diferentes propriedades químicas que são codificadas por meio da tradução do mRNA (Voet e Voet, 2013).

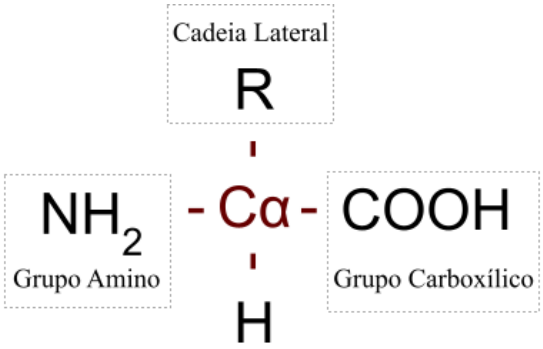


Figura 2.1 – Fórmula estrutural geral dos aminoácidos. Cada aminoácido é formado por um átomo de carbono (Cα) ligado a quatro diferentes grupamentos químicos: um grupamento amino (NH_2), um grupamento carboxila (COOH), um átomo de hidrogênio (H) e uma cadeia lateral (R). Fonte: Adaptada de Voet e Voet (2013).

A cadeia lateral é o que diferencia cada aminoácido e o que caracteriza as propriedades físicas e individuais deles (Xiong, 2006). Os aminoácidos são ligados covalentemente por meio de ligações peptídicas que são formadas pelo grupo amino de um aminoácido e o grupo carboxílico do aminoácido seguinte, liberando H_2O (Alberts et al., 2017). A Figura 2.2 representa a formação da ligação peptídica entre dois resíduos de aminoácidos.[4]

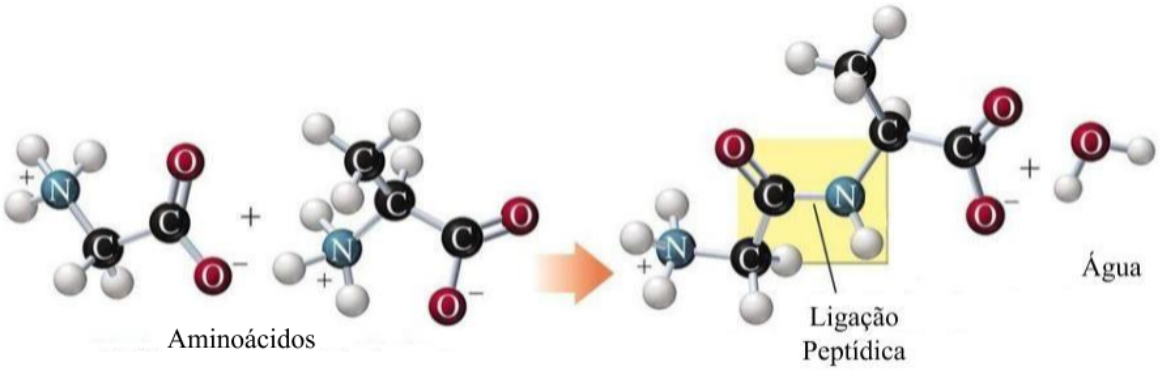


Figura 2.2 – Representação esquemática da ligação peptídica entre o grupo amino de um aminoácido e o grupo carboxílico do aminoácido seguinte liberando água e formando um dipeptídeo. Fonte: Adaptada de Timberlake (2015).

A diversidade que a cadeia lateral proporciona aos aminoácidos é importante para a conformação das proteínas e, consequentemente, à função (Lodish et al., 2008; Marzzoco e Torres Baptista, 2015). Além disso, auxilia na classificação desses em diferentes grupos (Marzzoco e Torres Baptista, 2015). Uma das propriedades que determina essa classificação é a polaridade, a qual varia do apolar (hidrofóbico) ao polar (hidrofílico) que podem ser subdivididos em carregados positivamente (básicos), carregados negativamente (ácidos) e não carregados (Verli, 2014; Marzzoco e Torres Baptista, 2015). Apesar disso,

[4] Resíduo de aminoácido: devido aos aminoácidos perderem uma parte de sua composição, OH ou H para a formação da ligação peptídica, o que resta do aminoácido é chamado de resíduo de aminoácido.

alguns aminoácidos são difíceis de classificar ou não se enquadram perfeitamente em um grupo, principalmente glicina, histidina e cisteína. Assim, suas atribuições são resultantes de avaliações parciais e não absolutas (Nelson e Cox, 2014).

Ademais, como as proteínas estão em meio aquoso, os aminoácidos apolares têm cadeias laterais que não interagem com a água e por isso, frequentemente, estão voltadas para o interior da molécula. Os aminoácidos polares têm nas cadeias laterais grupos que possibilitam a interação com a água, pois esses apresentam cargas residuais. Assim tendem a estar dispostos na superfície da molécula. Esse padrão é conhecido como efeito hidrofóbico. Outro fator importante são as interações eletrostáticas, na qual as cargas dos resíduos influenciam na atração ou repulsão desses com os outros (Lodish et al., 2008; Marzzoco e Torres Baptista, 2015).

Assim, conforme os princípios de bioquímica de Lehninger (Nelson e Cox, 2014), os aminoácidos podem ser separados em cinco classes: i) Grupos R apolares, alifáticos; ii) Grupos R polares, não carregados; iii) Grupos R carregados positivamente (básicos); iv) Grupos R carregados negativamente (ácidos) e v) Grupos R aromáticos. Com isso, as estruturas dos 20 aminoácidos e do iminoácido, seus nomes, suas representações em códigos de três e uma letra e suas classificações são ilustradas na Figura 2.3.

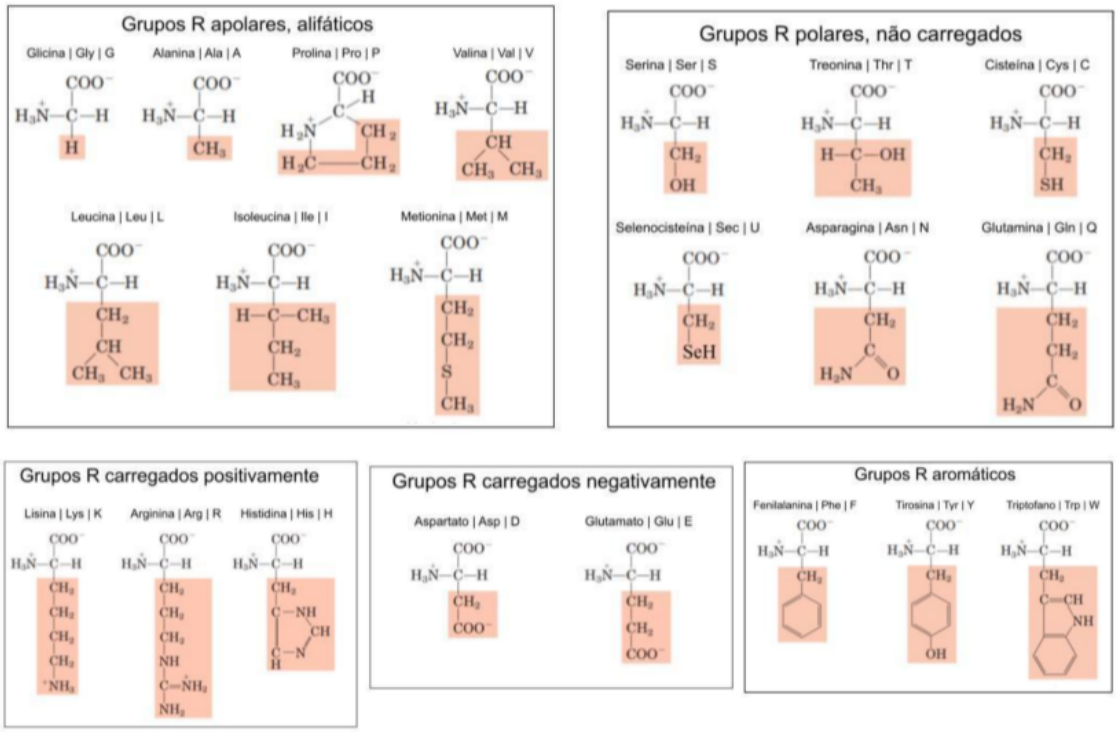


Figura 2.3 – Os 20 aminoácidos e 1 iminoácido. As regiões não sombreadas são aquelas comuns a todos os aminoácidos e as regiões sombreadas são os grupos R ou cadeia lateral. Apesar da histidina aparecer sem carga, uma pequena porção é positivamente carregada em potencial hidrogeniônico (pH) 7,0. Fonte: Adaptada de Nelson e Cox (2014).

2.2 Hierarquia Estrutural

A estrutura das proteínas é estudada e organizada em quatro níveis hierárquicos: estrutura primária - a sequência de resíduos de aminoácidos (Sanger, 1949); estrutura secundária - o dobramento de segmentos curtos (3 a 30 resíduos) de modo regular ou irregular (Pauling e Corey, 1951); estrutura terciária - a união das estruturas secundárias em unidades funcionais maiores (Kendrew et al., 1958; Perutz et al., 1960, 1968); e estrutura quaternária - arranjo de duas ou mais estruturas terciárias (Svedberg e Fåhraeus, 1926) (Figura 2.4).

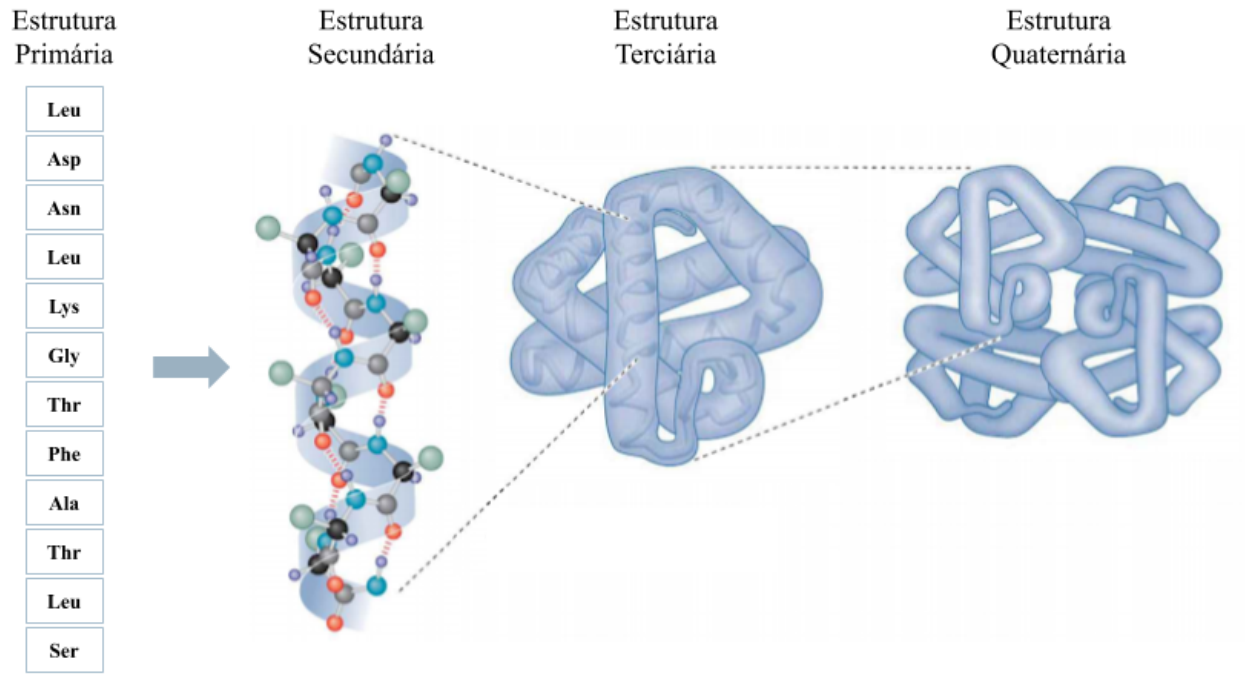


Figura 2.4 – Níveis hierárquicos da estrutura proteica ilustrada pela hemoglobina. A estrutura primária consiste na sequência de resíduos de aminoácidos e aqui representada pelo código de três letras. A estrutura secundária representa os padrões de interações entre os resíduos, sendo aqui representada pela estrutura regular hélice-α. A estrutura terciária refere-se ao enovelamento das estruturas secundárias, representando a proteína ou parte de um domínio (quando a proteína possui múltiplos domínios). A estrutura quaternária é o arranjo de múltiplas subunidades proteicas formando um complexo. Fonte: Adaptada de Watson et al. (2015).

A sequência linear de resíduos de aminoácidos forma a estrutura primária (Voet e Voet, 2013). Os resíduos ligados formam a cadeia principal e apresentam mobilidade com restrições nos ângulos diédricos. Os ângulos são conhecidos por phi (ϕ), o ângulo de ligação entre N-Cα, psi (ψ) , ângulo de ligação entre Cα-C' e ômega (ω), ângulo de ligação entre C'-N (Figura 2.5).

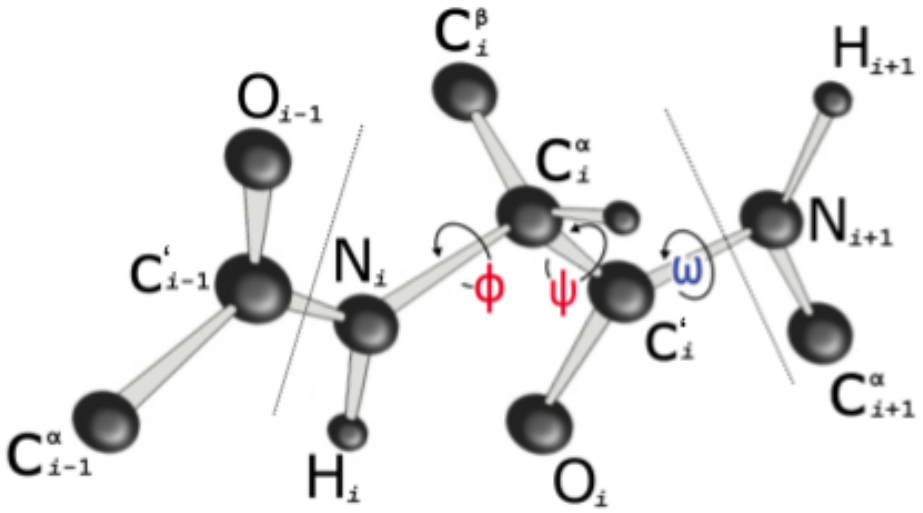


Figura 2.5 – Representação esquemática de um peptídeo identificando os ângulos de torção da cadeia principal ϕ, ψ e ω. Os ângulos ϕ, ψ estão em torno das ligações de N-Cα e Cα-C', enquanto que o ângulo ω está em torno das ligações peptídicas. Fonte: Adaptada de Lesk (2008).

Além dos ângulos da cadeia principal, os ângulos qui (χ) representam os ângulos da cadeia lateral. Esses podem influenciar nas conformações adotadas e na estabilização da molécula. Diferente dos ângulos da cadeia principal, os ângulos χ dependem do tipo de resíduo de aminoácido. Os átomos da cadeia lateral dos resíduos são nomeados sequencialmente de acordo com o alfabeto grego (podendo ser adaptado ao alfabeto latino). Por exemplo, a cadeia lateral da lisina é composta por um carbono β (CB), um carbono γ (CG), um carbono δ (CD), um carbono ϵ (CE) e um nitrogênio Z (NZ) (Guex e Peitsch, 2006).

A partir da estrutura primária, padrões de interação entre os resíduos e os resíduos com o solvente formam a estrutura secundária. Como dito anteriormente, os aminoácidos com características hidrofóbicas tendem a ser interiorizados e os aminoácidos com características hidrofílicas expostos à superfície. Assim, os resíduos se organizam em estruturas regulares no formato de hélices e folhas β (Pauling e Corey, 1951).

As hélices são formadas a partir de três padrões: i) hélice-α com 3,6 resíduos de aminoácidos por volta, estrutura mais comum e que corresponde, aproximadamente, 97% de todas as hélices; ii) hélice-3_{10} com 3,0 resíduos de aminoácidos por volta representando, aproximadamente, 3% de todas as hélices e; iii) hélice-π com 4,4 resíduos de aminoácidos por volta e que raramente ocorrem (Pevsner, 2015). As folhas β são constituídas de 2-15 resíduos (geralmente 5-10 resíduos) dispostos em orientação paralela, antiparalela e mista, que ocorrem quando existe pelo menos duas fitas próximas, lado a lado (Pevsner, 2015).

Além das estruturas regulares, há estruturas irregulares como voltas e alças. As voltas apresentam entre dois e quatro resíduos de aminoácidos e as alças cinco ou mais resíduos. Essas estruturas são fundamentais para conectar sucessivas estruturas secun-

dárias regulares (Voet e Voet, 2013). As estruturas secundárias regulares e irregulares são representadas na Figura 2.6.

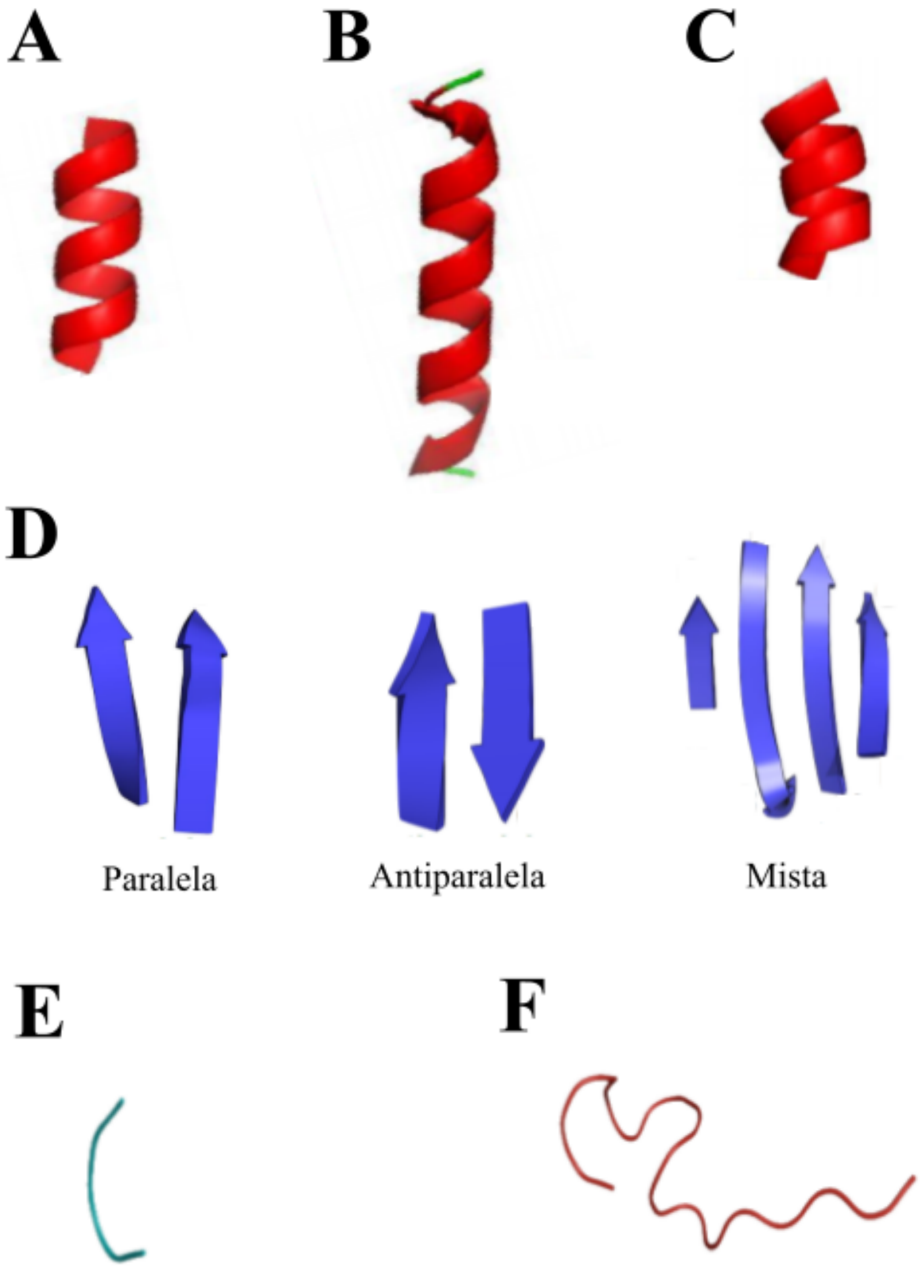


Figura 2.6 – Representação das estruturas secundárias regulares e irregulares. (A) Estrutura secundária regular do tipo hélice-α com 3,6 resíduos de aminoácidos por volta. (B) Estrutura secundária regular do tipo hélice-3_{10} com 3,0 resíduos de aminoácidos por volta. (C) Estrutura secundária regular do tipo hélice-π com 4,4 resíduos de aminoácidos por volta. (D) Estrutura secundária regular do tipo folha β, paralela, antiparalela e mista. (E) Estrutura secundária irregular do tipo volta. (F) Estrutura secundária irregular do tipo alça. Fonte: Autora.

Além disso, para cada aminoácido a atribuição da estrutura secundária correspondente é atribuída por meio do *Dictionary of Protein Secondary Structure* (DSSP). O DSSP é um banco de dados padronizado de atribuições de estrutura secundária para todas as entradas de proteínas disponíveis no PDB (Kabsch e Sander, 1983; Touw et al., 2015). Cada estrutura secundária é representada por uma letra conforme a Tabela 2.1.

Tabela 2.1 – Representação dos tipos de estrutura secundária definidas pelo DSSP. Fonte: Magnan e Baldi (2014).

Símbolo	Classe
H	Hélice - α
G	Hélice-3_{10}
I	Hélice - π
E	Folhas β
B	Resíduo em ponte β isolada
T	Volta
S	Dobra
C	Outros

Adicionalmente, sabendo os valores dos ângulos ϕ e ψ adotados por cada resíduo de aminoácido em uma proteína, é possível representá-los através de um gráfico ϕ versus ψ denominado mapa de Sasisekharan-Ramakrishnan-Ramachandran, mais conhecido por mapa de Ramachandran (Ramachandran e Sasisekharan, 1968). Assim, segundo as definições de Efimov (1993), dependendo da região do mapa em que os ângulos estão dispostos a estrutura secundária mais provável é indicada.

Os eixos do mapa variam de -180º a +180º e apresentam as regiões mais favoráveis, regiões adicionalmente permitidas, regiões generosamente permitidas e regiões não permitidas para os resíduos. As regiões demarcadas no mapa são regiões não permitidas para os aminoácidos, exceto para a glicina em que sua cadeia lateral se restringe a um átomo de hidrogênio, o que possibilita maior liberdade para os ângulos de torção. Além disso, cada aminoácido possui um padrão de distribuição no mapa de Ramachandran, mas que quando combinados, forma um mapa único para cada proteína. Assim, a Figura 2.7 mostra o mapa referente a proteína de código PDB:1ZDD gerada pelo PROCHECK - PDBsum [5] (Laskowski et al., 1993).

[5] http://www.ebi.ac.uk/thornton-srv/databases/cgi-bin/pdbsum/GetPage.pl?pdbcode=index.html acesso em: 05 de fevereiro de 2021.

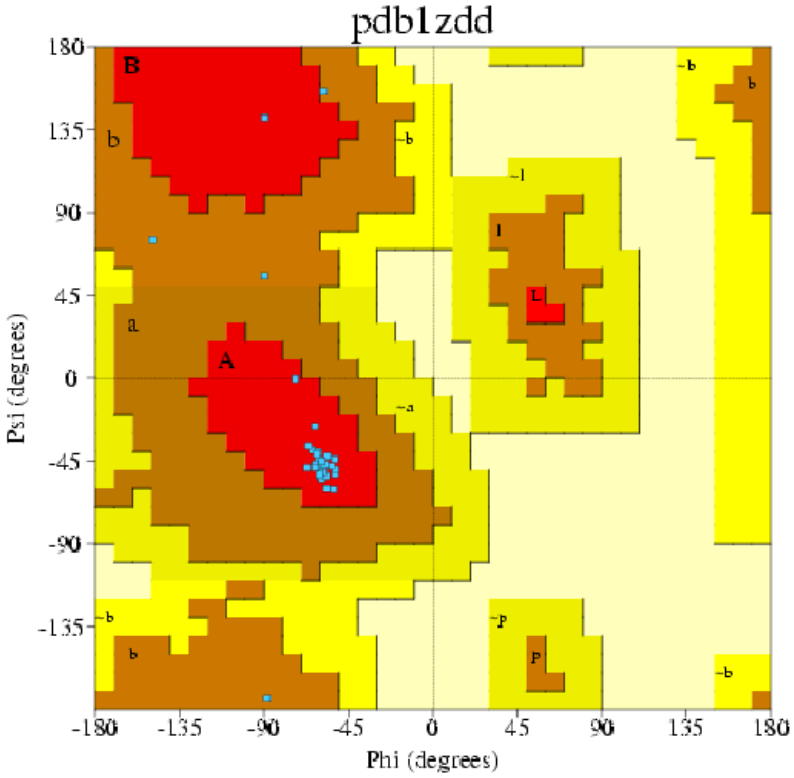


Figura 2.7 – Mapa de Ramachandran da proteína 1ZDD gerada pelo PROCHECK - PDBsum (Laskowski et al., 2018). O eixo x e y representam, respectivamente, os ângulos ϕ e ψ. As regiões mais favoráveis são representadas em vermelho, as regiões adicionalmente permitidas em marrom, as regiões generosamente permitidas em amarelo e as regiões não permitidas em amarelo claro. Os pontos azuis representam os resíduos de aminoácidos com seus ângulos correspondentes em x e y. Fonte: Autora.

Com a disponibilidade do mapa, os estudos de Efimov (1993) mostraram que as estruturas do tipo hélice, geralmente, estão dispostas no 1º e 3º quadrante, as folhas β no 2º quadrante e as estruturas irregulares podem ocupar qualquer região, inclusive as regiões de hélices e folhas (Figura 2.8). Sendo assim, predizer estruturas irregulares por métodos computacionais não é uma tarefa trivial.

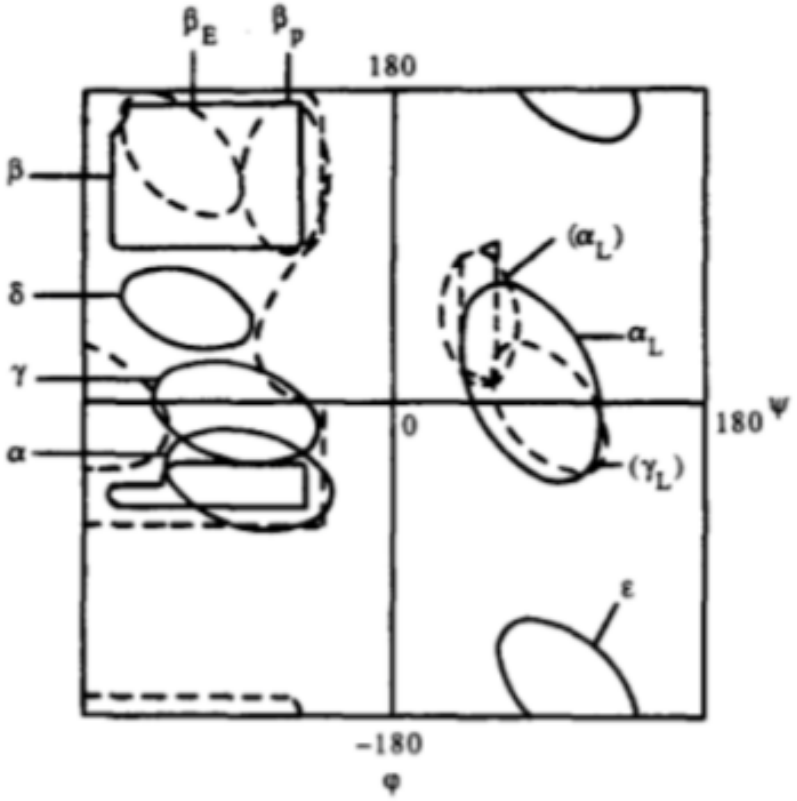


Figura 2.8 – Representação dos tipos de estrutura secundária mais encontrados nas regiões do mapa de Ramachandran conforme Efimov. A nomenclatura indica regiões do mapa para as conformações: β - folhas β; α - hélices α; α_L- hélices α à esquerda; γ - volta γ; ϵ - volta ϵ; δ - volta δ. Fonte: Efimov (1993).

Assim, o modo como essas estruturas interagem permitem que uma proteína se enovele formando a estrutura terciária. Este processo é influenciado por diversos fatores tais como as ligações de hidrogênio, interações de van der Waals, ângulos da cadeia principal, interações eletrostáticas da cadeia lateral e interações hidrofóbicas que evitam sobreposições físicas que não existem (Pevsner, 2015). O arranjo de duas ou mais estruturas terciárias forma um complexo multiproteico denominada estrutura quaternária. Este complexo compõem uma proteína funcional mantida por ligações não-covalente entre as estruturas (Marzzoco e Torres Baptista, 2015).

Além disso, a estrutura terciária, experimentalmente, é determinada por meio de técnicas de difração por raio-X (Rupp, 2009), RMN (Wüthrich, 1986) ou microscopia eletrônica (Baumeister e Steven, 2000). Porém, computacionalmente, as estruturas podem ser preditas por diferentes métodos que buscam solucionar o problema da predição da estrutura tridimensional de proteínas (PSP).

2.3 Problema: Enovelamento e Predição da Estrutura Tridimensional de Proteínas

O modo como uma proteína se enovela é um problema de rápida solução na natureza. Entretanto, mapear todas as possibilidades conformacionais a partir de uma amostragem aleatória que uma sequência de resíduos de aminoácidos pode assumir necessita de um período de tempo maior que a idade do universo. Com isso, Levinthal (1968) mostrou que a natureza busca um caminho não aleatório para encontrar o estado conformacional mais apropriado (mais tarde conhecido como Paradoxo de Levinthal).

Apesar disso, determinar a estrutura primária de uma proteína é uma tarefa relativamente fácil, mas o mesmo não ocorre com a estrutura 3D. Assim, o problema do enovelamento de proteínas está relacionado a algumas questões das quais destaco a seguir.

- Qual código ou informação contida na sequência de resíduos de aminoácidos dita a estrutura nativa de uma proteína?
- Como uma proteína consegue se enovelar tão rápido na natureza?
- A partir de uma sequência de resíduos de aminoácidos é possível prever a estrutura 3D correspondente?

Em busca de responder tais questões, principalmente a última destacada na Figura 2.9, o PSP surge com diferentes propostas, métodos computacionais e soluções aproximadas. Além disso, tais técnicas computacionais reduzem alguns problemas existentes nas técnicas experimentais como o alto custo, a complexidade e a impossibilidade de uso da técnica para determinadas proteínas.

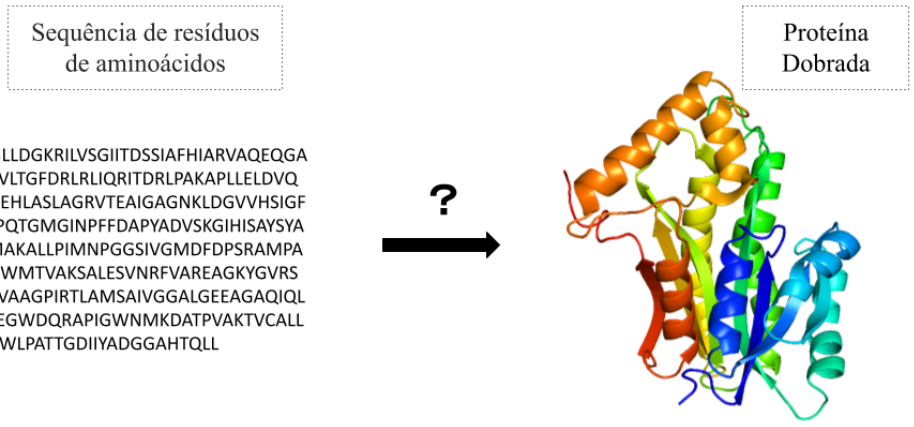


Figura 2.9 – Problema da predição da estrutura tridimensional de proteínas. A partir de uma sequência de resíduos de aminoácidos é possível conhecer a estrutura 3D correspondente? A figura indica a sequência de resíduos de aminoácidos da InhA e a partir dela sua estrutura terciária - Código PDB: 1ENY. Fonte: Adaptada de Machado (2016).

2.4 Métodos Computacionais para a Predição Tridimensional de Proteínas

Os métodos computacionais para a predição tridimensional de proteínas podem ser separados em quatro categorias (Floudas et al., 2006).

1. Modelagem Comparativa (Martí-Renom et al., 2000);
2. Reconhecimento de Padrões de Enovelamento ou *Fold Recognition* (Bowie et al., 1991; Jones et al., 1992);
3. Métodos *de novo* (Bowers et al., 2000; Srinivasan e Rose, 1995);
4. Métodos *ab initio* (Osguthorpe, 2000).

2.4.1 Modelagem Comparativa por Homologia

A modelagem comparativa é baseada no princípio de que se duas sequências estão relacionadas evolutivamente, as suas estruturas tridimensionais são similares, pois pequenas mudanças sequenciais têm pequenas alterações estruturais (Chothia e Lesk, 1986; Xiang, 2006). Este tipo de abordagem garante alta precisão nos modelos gerados quando as sequências apresentam tal evidência. Apesar disso, métodos baseados em modelagem comparativa dependem das estruturas já determinadas experimentalmente e não permitem que novos padrões de enovelamento sejam encontrados, já que conseguem apenas determinar padrões encontrados nas estruturas das quais se baseiam. Os principais preditores conhecidos são o Modeller (Sali e Blundell, 1993; Webb e Sali, 2016) e o SWISS-MODEL (Arnold et al., 2006).

2.4.2 Reconhecimento de Padrões de Enovelamento ou *Fold Recognition*

Os métodos de *fold recognition* consideram que a estrutura 3D de uma proteína é evolutivamente mais conservada que a estrutura primária correspondente. Assim, se uma sequência de resíduos não apresenta alta similaridade com uma estrutura já conhecida, ainda assim, é possível que proteínas conhecidas tenham estrutura similar àquela estudada. Essa identificação ocorre através de um conjunto de enovelamentos candidatos (Bowie et al., 1991). Se essa etapa for bem sucedida, a etapa do alinhamento estrutural ocorre como na modelagem comparativa. Caso contrário, a técnica de alinhavamento é utilizada (Jones et al., 1992). Dentre os métodos mais conhecidos desse grupo estão HHpred (Söding, 2005), SPARKS-X (Yang et al., 2011) e Phyre (Kelley et al., 2015).

2.4.3 Métodos *de novo*

Os métodos *de novo* não se baseiam em estruturas depositadas em bases de dados, porém extraem informações a partir delas. As informações extraídas podem ser utilizadas para construir estruturas tridimensionais desde o princípio, não sendo limitado aos padrões conhecidos, já que as comparações são realizadas por fragmentos e não com estruturas inteiras (Baker e Sali, 2001). Além disso, é possível utilizar informações baseadas na hipótese de Anfinsen (1973).

A hipótese assume que a energia global livre de uma proteína é determinada por meio da energia potencial, na qual descreve a energia interna e as interações desta com o meio, sendo a menor energia a que representaria o estado nativo da proteína (Figura 2.10)(Anfinsen, 1973; Jana et al., 2018). Além disso, algumas predições intermediárias podem ser utilizadas para compor o conjunto de informações relacionadas à estrutura 3D, como preditores de estrutura secundária e preditores de contato.

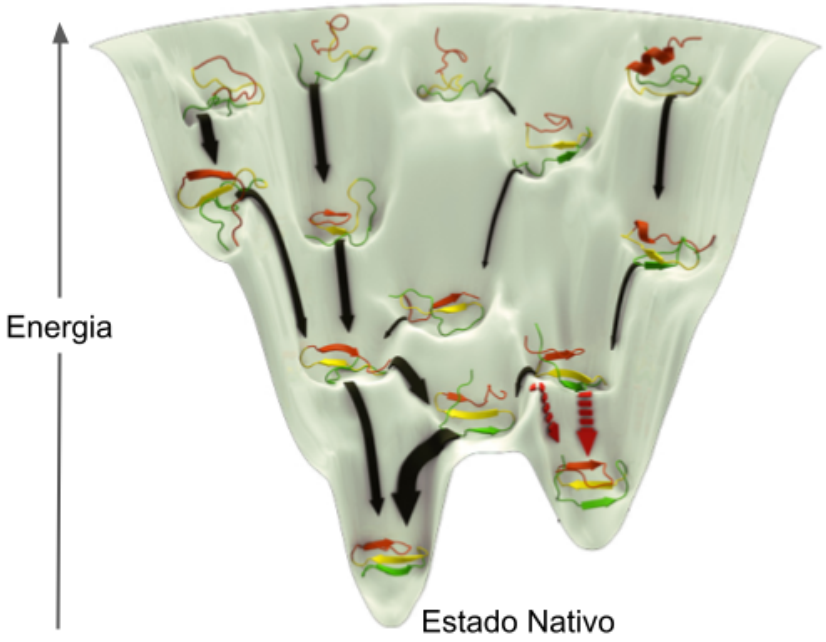


Figura 2.10 – Funil Energético. O estado nativo de uma proteína apresenta o menor valor energético em relação a outras conformações. Fonte: Adaptada de Siow (2018).

Dentre os métodos que se encaixam nesta categoria de predição, os mais conhecidos são I-TASSER (Roy et al., 2010) e ROSETTA (Rohl et al., 2004).

2.4.4 Métodos *ab initio*

Os métodos *ab initio* são aqueles que, para encontrar a estrutura enovelada de uma proteína, a partir da sequência de resíduos de aminoácido, são baseados, exclusivamente, nos princípios físicos (Osguthorpe, 2000). As informações utilizadas nesse tipo de método estão relacionadas à parametrização dos campos de força que são incorporados em diversas abordagens. Os campos de força mais utilizados são o AMBER (Cornell et al., 1995), CHARMM (Brooks et al., 1983) e GROMOS (Christen et al., 2005).

As abordagens deste tipo permitem que novos padrões de enovelamento sejam preditos, porém os graus de liberdade que a cadeia polipeptídica pode assumir e o número de conformações que devem ser considerados influenciam na complexidade do problema, sendo o PSP considerado um problema NP-Completo (Berger e Leighton, 1998; Crescenzi et al., 1998).

2.5 CASP: Critical Assessment of Structure Prediction

Independe do método computacional escolhido para a predição 3D de proteínas, no início da década de 90, a comunidade científica reconheceu que esses aumentariam significativamente. Assim, em 1994 surgiu a ideia de avaliar os preditores por meio de experimentos às cegas, no qual estruturas tridimensionais de algumas proteínas determinadas experimentalmente e ainda não disponíveis ao público fossem utilizadas (Gu e Bourne, 2009; Moult et al., 2018).

Desse modo, um encontro em formato de competição chamado *Critical Assessment of Structure Prediction* (CASP) ocorre a cada dois anos com o objetivo de determinar e avançar o estado da arte na predição de estrutura de proteínas. As sequências de resíduos de aminoácidos são disponibilizadas e cada grupo utiliza seu método de predição. Com isso, o CASP consegue avaliar os preditores e compará-los. Contudo, ao longo dos anos, os métodos de avaliação e as categorias de participação tiveram alterações.

Na última edição ocorrida em 2020 (CASP14), as categorias foram: (i) *free modeling* (FM), (ii) *template-based modeling* (TBM), (iii) *contact prediction*, (iv) *help structural biologists*, (v) *refinement* e (vi) *data-assisted modeling*. Além disso, quase 100 grupos de todo o mundo enviaram mais de 67.000 modelos para 90 alvos proteicos [6]. Com isso, os resultados alcançados nos últimos CASP, principalmente no CASP14, causaram grandes entusiasmos para a comunidade, destacando o *AlphaFold* (Jumper et al., 2020).

[6] Informação disponível em: https://predictioncenter.org/index.cgi

2.5.1 AlphaFold

Em 2018, o *DeepMind* - grupo de pesquisa de aprendizado de máquina (AM) do *Google* - participou do CASP pela primeira vez. Eles criaram um algoritmo baseado em redes neurais profundas conhecido por *AlphaFold*, no qual os modelos foram gerados através do uso de previsões de distância ou contato entre os pares de resíduos de aminoácidos. Com isso, os modelos apresentaram alta precisão quando comparados a outros métodos (Senior et al., 2019).

Assim, em 2020, uma nova versão do *AlphaFold* foi apresentada no CASP14, mas o código e os detalhes de funcionamento não foram disponibilizados até o presente momento. No entanto, das informações já conhecidas, é importante destacar a utilização de um conjunto de dados públicos de 170.000 proteínas com estruturas 3D conhecidas e um banco de dados de sequências sem estrutura 3D conhecidas, o emprego de técnicas de alinhamento múltiplo de sequências (MSA - do inglês, *multiple sequence alignment*) e a permanência da previsão dos pares de resíduos de aminoácidos em contato (Senior et al., 2020). Além da necessidade de algumas semanas de processamento em 16 TPUv3s (equivalente a aproximadamente 100 - 200 GPUs).

Conforme John Moult, co-fundador do CASP, os resultados atingidos pelo *AlphaFold* no CASP14 podem ser comparados àqueles obtidos por técnicas experimentais e com isso, a ciência e a comunidade de predição de estrutura 3D de proteínas estão passando por um momento especial [7]. Apesar do grande passo, ainda há muitas questões relacionadas à problemática em aberto, tais como a dependência da qualidade na previsão dos contatos, o entendimento do cenário energético e os estados de transição da proteína para o processo de enovelamento, afinidade com ligantes e compreensão de doenças relacionadas a problemas no enovelamento.

2.6 Contato entre Resíduos de Aminoácidos

Os resíduos de aminoácidos distantes na sequência linear, normalmente, não apresentam efeito uns sobre os outros, a menos que estejam próximos no espaço 3D quando a proteína é enovela. Desse modo, ter a informação de quais resíduos estão próximos, ou seja, em contato, pode ser valiosa para prever a estrutura terciária de uma proteína e identificar os principais resíduos responsáveis (Amala e Emerson, 2019). Assim, quando a distância euclidiana entre as coordenadas do Cβ (Cα para glicina) de dois resíduos de

[7] Informação disponível em: https://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology

aminoácidos (par de resíduos) é menor que 8 Å [8] defini-se que os resíduos estão em contato [9]. (Di Lena et al., 2012). A Equação 2.1 indica se o par está em contato, onde $d(a_i, a_j)$ representa a distância entre dois resíduos de aminoácidos a_i e a_j e $|r_i - r_j|$ a distância das coordenadas dos Cβ (Cα para glicina).

$$d(a_i, a_j) = |r_i - r_j| \tag{2.1}$$

Além disso, os contatos existentes entre os pares de resíduos de uma proteína são separados em interações de curto, médio e longo alcance. O alcance é definido considerando a posição dos resíduos na estrutura primária. Assim, um par de resíduos está em contato de curto alcance quando o valor do módulo da distância do par na estrutura primária é entre 6 e 11 resíduos. Para o par de resíduos estar em contato de médio alcance, o valor do módulo da distância do par na estrutura primária precisa ser entre 12 e 23 resíduos. Logo, para o par de resíduos estar em contato de longo alcance, o valor do módulo da distância do par na estrutura primária é de pelo menos 24 resíduos, conforme a Tabela 2.2. (Jones et al., 2012). Aqueles pares de resíduos em que a distância é menor que 6, geralmente, não são considerados por estarem sequencialmente próximos e provavelmente em contato. Com isso, nem todos os contatos preditos são usados.

Tabela 2.2 – Separação dos contatos entre os pares de resíduos de aminoácidos. Os pares de resíduos podem apresentar contatos de curto, médio ou longo alcance definido pela posição dos resíduos na estrutura primária. Assim, um par de contato é dito de curto alcance quando o módulo da distância na estrutura primária dos resíduos do par é entre 6 e 11 resíduos, médio alcance quando o módulo da distância dos resíduos do par é entre 12 e 23 resíduos e longo alcance quando o módulo da distância dos resíduos do par é de pelo menos 24 resíduos. Fonte: Jones et al. (2012).

Distância dos resíduos de aminoácidos na estrutura primária	Tipo de Contato (alcance)
entre 6 e 11	Curto
entre 12 e 23	Médio
pelo menos 24	Longo

A partir dos contatos é possível representá-los por meio de um mapa. O mapa gerado é espelhado e a diagonal principal representa o contato do resíduo com ele mesmo. Por isso, o mapa não apresenta essa informação. A Figura 2.11 mostra uma proteína em

[8] Unidade de medida de comprimento usada para distâncias atômicas na qual 1Å equivale a 10^{-10} m.

[9] Outros valores de distância já foram e podem ser utilizados, menor que 8 Å é o valor mais aceito pela comunidade científica, inclusive o CASP, atualmente.

que sua estrutura 3D apresenta hélice-α e o mapa de contatos correspondente. Os resíduos que formam as estruturas secundárias regulares do tipo hélice-α são representados, principalmente, em retas adjacentes a diagonal principal e interrupções nessas diagonais, geralmente, representam estruturas irregulares. Assim, os resíduos em contato que formam as hélice-α dessa proteína são destacadas pelas letras A e B e as regiões irregulares destacada pela letra C.

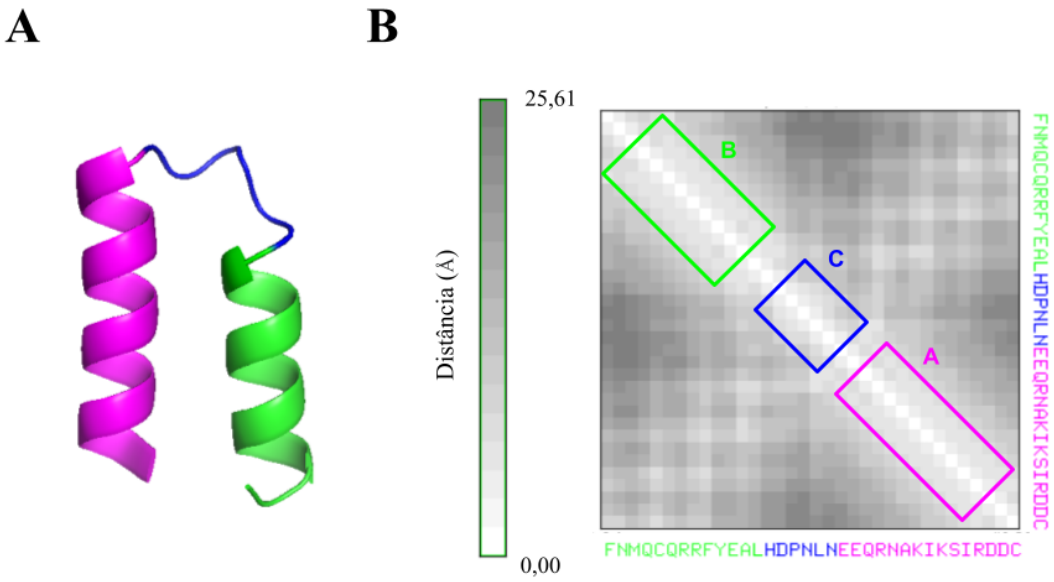


Figura 2.11 – Representação do mapa de contatos da proteína de código PDB: 1ZDD. **(A)** Estrutura 3D da proteína de código PDB: 1ZDD. **(B)** Mapa de contato da proteína 1ZDD. As hélices-α formam retas adjacentes a diagonal principal e são destacadas em rosa (letra A) e verde (letra B). A interrupção dos contatos das hélices representam as regiões irregulares da proteína que são destacadas em azul (letra C). Fonte: Autora.

As proteínas com estruturas secundárias regulares do tipo folhas-β apresentam outros padrões nos mapas como indicado na Figura 2.12. Há ainda diferença nas representações quando as fitas formam folhas-β paralelas ou antiparalelas. Assim, folhas-β paralelas são representadas no mapa por retas paralelas a diagonal principal e as folhas-β antiparalelas por retas perpendiculares a diagonal principal (Amala e Emerson, 2019). Essa no mapa não é marcada como em mapas de proteínas que formam hélices.

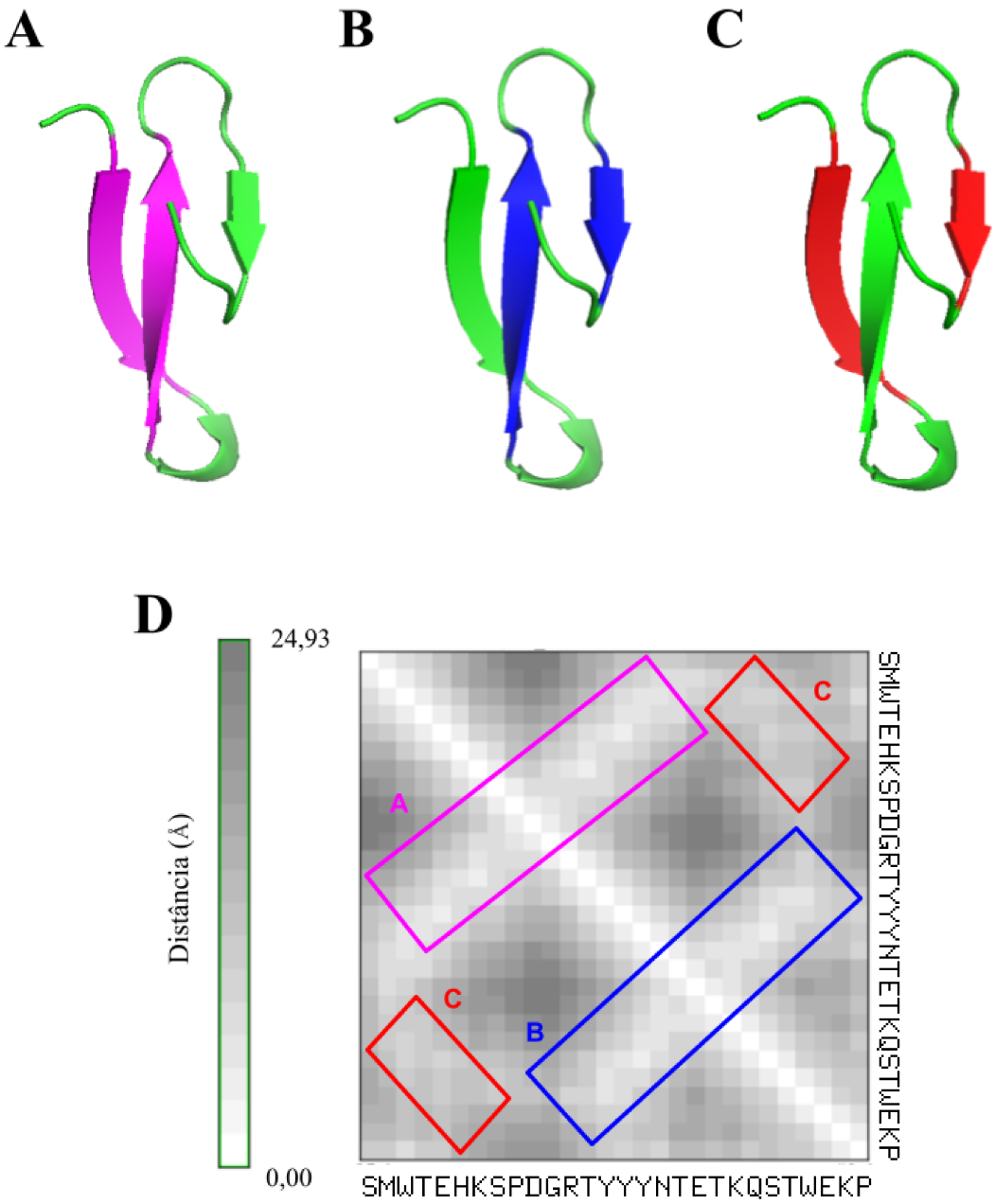


Figura 2.12 – Representação do mapa de contatos da proteína de código PDB: 1YWJ. **(A)** Estrutura 3D da proteína 1YWJ, destacando as fitas que formam a primeira folha-β antiparalela. **(B)** Estrutura 3D da proteína 1YWJ, destacando as fitas que formam a segunda folha-β antiparalela. **(C)** Estrutura 3D da proteína 1YWJ, destacando as fitas que formam a folha-β paralela. **(D)** Mapa de contato da proteína 1YWJ. As folhas-β antiparalelas aparecem, perpendicularmente, a diagonal principal (letras A e B), já a folha-β paralela aparece, paralelamente, à diagonal principal em vermelho (letra C). Fonte: Autora

Assim, os mapas de contato são um tipo de representação muito utilizada. Porém, quando a estrutura 3D da proteína não é conhecida, os contatos também não são. Com isso, técnicas de predição de contatos são desenvolvidas com o objetivo de prever os contatos, para que esses possam descrever a estrutura 3D.

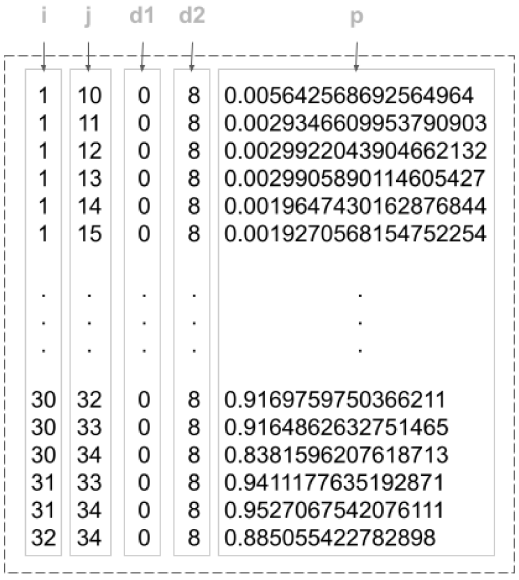
2.7 Métodos de Predição de Contato

As predições de contato surgiram na década de 90 (Vendruscolo e Domany, 2000; Mirny e Domany, 1996), porém a taxa de falsos-positivos inviabilizavam o uso desse tipo de informação. Contudo, nos últimos anos o uso dos contatos apresentaram melhora significativa na predição da estrutura proteica em abordagens *de novo* (de Oliveira et al., 2017). Com isso, o aumento da confiabilidade das predições de contato e a possibilidade de utilizá-las acelerando o processo de predição da estrutura 3D se tornou possível (Feng et al., 2020).

O uso dos contatos pode auxiliar nas restrições espaciais em algoritmos que buscam a melhor conformação (Miller e Eisenberg, 2008; Wang et al., 2011). Em geral, um termo de energia relacionado aos contatos de resíduos é projetado e adicionado a função de energia, para que a otimização do processo evite uma grande busca no espaço conformacional, principalmente, com conformações não consistentes com os contatos previstos (Adhikari e Cheng, 2016; Feng et al., 2020).

Assim sendo, os métodos para predições de contato podem ser classificados em diferentes categorias, as quais são modificadas constantemente (Björkholm et al., 2009; Di Lena et al., 2012; Schneider e Brock, 2014; Feng et al., 2020). Apesar disso, há dois grandes grupos de métodos de predição de contato que são i) baseado em coevolução e ii) baseado em aprendizado de máquina. Contudo, com o avanço dos métodos, alguns preditores utilizam as duas abordagens, sendo difícil classificá-los em uma única categoria (Adhikari e Cheng, 2016).

No entanto, independente do método utilizado para a predição dos contatos, o CASP adotou um formato de arquivo para a submissão dos resultados gerados por esses com o intuito de conseguir compará-los. Esse formato de arquivo resíduo-resíduo (RR) possui cinco colunas e em cada linha do arquivo apresenta um par de resíduos (as posições desses na estrutura primária) às distâncias minímas e máximas entre Cβ (Cα para glicina) dos dois resíduos e a probabilidade desses estarem em contato. Assim, os pares que apresentam maiores valores de probabilidade, são mais prováveis de estarem em contato de fato (Jones et al., 2015). O formato do arquivo pode ser visualizado na Figura 2.13.



i	j	d1	d2	p
1	10	0	8	0.005642568692564964
1	11	0	8	0.0029346609953790903
1	12	0	8	0.0029922043904662132
1	13	0	8	0.0029905890114605427
1	14	0	8	0.0019647430162876844
1	15	0	8	0.0019270568154752254
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
30	32	0	8	0.9169759750366211
30	33	0	8	0.9164862632751465
30	34	0	8	0.8381596207618713
31	33	0	8	0.9411177635192871
31	34	0	8	0.9527067542076111
32	34	0	8	0.885055422782898

Figura 2.13 – Formato RR adotado pelo CASP para submissão dos resultados dos métodos de predição de contato. A lista dos contatos segue um formato de cinco colunas aqui representadas por: i j d1 d2 p. Os índices i e j representam a posição na sequência dos dois resíduos em contato de modo que i < j, fornecendo apenas metade do mapa. As colunas d1 e d2 são as distâncias entre Cβ (Cα para glicina) de dois resíduos de aminoácidos, geralmente, d1 = 0 e d2 = 8. Esses parâmetros, atualmente, são utilizados apenas para consistência do formato. A coluna p representa a probabilidade dos dois resíduos estarem em contato, com valores entre 0 e 1. Valores maiores que 0,5 identificam pares de resíduos previstos mais prováveis de estarem em contato. Fonte: Autora.

2.7.1 Métodos baseados em coevolução

Os métodos baseados em coevolução são aqueles que, se um par de resíduos está em contato e é crítico para manter a estrutura 3D da proteína dobrada, os resíduos do par apresentam mutações correlacionada ou coevolução (Shindyalov et al., 1994). Ou seja, se um dos resíduos do par sofrer uma mutação que afeta a estrutura 3D, o outro resíduo tem maiores chances de sofrer uma mutação que garanta que a estrutura 3D e a função da proteína não sejam alteradas. Assim, as primeiras abordagens que utilizaram esse princípio extraíam os pares de resíduos através do alinhamento múltiplo de sequências (MSA - do inglês, *Multiple Sequence Alignment*) e esse processo ficou conhecido por análise de mutação correlacionada (CMA - do inglês, *Correlated Mutation Analysis*).

O MSA consiste em alinhar três ou mais sequências que são parcialmente ou completamente alinhadas, pois acredita-se que essas tenham um ancestral em comum (Pevsner, 2015). Quanto mais semelhantes essas sequências de resíduos de aminoácidos, mais provável é que essas proteínas tenham um propósito semelhante para os organismos

nos quais estão presentes, o que significa que é mais provável que compartilhem uma estrutura semelhante. Algumas ferramentas que realizam esses alinhamentos são o PSI-BLAST (Altschul et al., 1997), HHblits (Remmert et al., 2012) e Jackhmmer (Johnson et al., 2010). Deste modo, a predição de contatos depende das sequências homólogas à proteína e da qualidade do alinhamento (Feng et al., 2020).

Por esse motivo, as primeiras abordagens tiveram um desempenho ruim, já que o número de sequências disponíveis era baixo e ruídos de correlação eram frequentes. Por exemplo, se o resíduo A e o resíduo B eram previstos em contato com o resíduo C (correlação direta), era provável que os resíduos A e B fossem previstos em contato (correlação indireta), porém não necessariamente esse contato ocorria. Assim, para resolver esse problema, a análise de correlação direta surgiu em busca de separar a correlação direta da indireta. Há diversos métodos que visam solucionar esse problema e os mais conhecidos são PSICOV (Jones et al., 2012), GREMLIN (Kamisetty et al., 2013), CCMpred (Seemayer et al., 2014) e FreeContact (Kaján et al., 2014).

2.7.2 Métodos baseados em aprendizado de máquina

Há uma grande variedade de métodos de predição de contatos que utilizam aprendizado de máquina (AM) com diferentes abordagens como máquina de vetores de suporte (Cheng e Baldi, 2007), algoritmo genético (Chen e Li, 2010), e mais recente redes neurais artificiais. Dessas, se destacam *DeepContact* (Liu et al., 2018), RaptorX-Contact (Wang et al., 2018), DNCON2 (Adhikari et al., 2018), RESPRE (Li et al., 2019) e *Deep*MetaPSICOV (Kandathil et al., 2019).

Muitos dos métodos são disponibilizados online como servidores ou para download. Além disso, entre as abordagens mais recentes, é possível encontrar aquelas com ou sem informações de coevolução, estrutura secundária, acessibilidade ao solvente, informações do tipo de resíduo como polaridade e informações de pares de resíduos em contato.

2.8 Simulação Molecular

A simulação computacional surgiu como alternativa às formas experimentais e teóricas da ciência. Assim, os métodos de simulação computacional buscam resolver modelos teóricos com o uso de computadores (Rino e Costa, 2013). Desse modo, os dois métodos computacionais de simulação molecular mais conhecidos e utilizados são a Dinâmica Molecular (DM) e o Monte Carlo (MC). Os dois se diferenciam, respectivamente, pelo caráter determinístico e probabilístico.

A DM é fundamentada nos princípios da mecânica clássica, o qual confere informações relacionadas ao comportamento dinâmico microscópico dos átomos individuais ao longo do tempo (Höltje et al., 2008). Assim, sua função é gerar um sistema de N partículas que interagem de acordo com regras usadas no campo de força (CF) (Rino e Costa, 2013; Höltje et al., 2008). Consequentemente, a partir da estrutura 3D do sistema é possível calcular a energia potencial total desse. Um campo de força é composto por termos harmônicos (comprimento e ângulos de ligação). Além disso, em simulações por DM, o uso de condições periódicas de contorno (CPC), em que os átomos do sistema são colocados em uma caixa e essa é replicada em todas as direções do espaço, auxilia na conservação do sistema (Frenkel e Smit, 2001).

Em contraste à DM, o método de MC gera uma série de estados microscópicos sob uma lei estocástica, independente das equações de movimento das partículas. Assim, permite que esses sejam feitos em qualquer direção e atribui um valor de probabilidade para cada movimento (Satoh, 2010). A partir disso, como o método é mais simples se comparado a DM, esse surgiu como alternativa a ser acoplada a métodos de predição de estrutura 3D de proteínas. Adicionalmente, independente do método de simulação, ambos buscam através do cálculo da energia do sistema encontrar a conformação de menor valor e que representaria a conformação nativa adotada por uma proteína, conforme a hipótese de Anfinsen.

2.9 Métricas de avaliação

A fim de avaliar a qualidade das estruturas preditas por diferentes métodos computacionais e das conformações geradas por métodos de simulação molecular, há diversas métricas que podem ser aplicadas. Dependendo da métrica, essa requer ou não uma estrutura de referência para ser calculada. Quando há uma estrutura de referência, as métricas mais conhecidas são o desvio quadrático médio (RMSD, do inglês *Root-Mean-Square Deviation*) e o teste de distância global (GDT, do inglês *Global Distance Test*). Assim, a métrica aplicada neste trabalho é explicada a seguir.

2.9.1 RMSD

O RMSD é a medida da distância média entre os átomos das proteínas sobrepostas. A Equação 2.2 indica como o cálculo é realizado.

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\delta_i^2} \tag{2.2}$$

Onde δ é a distância entre o átomo *i* e a estrutura de referência ou a distância entre *N* pares de átomos. Geralmente, esses pares são referentes àqueles pertencentes a cadeia principal (C, N, O e Cα) e o átomo apenas o Cα. Além disso, rotações e translações em uma das proteínas são realizadas com o intuito de encontrar a melhor sobreposição e, consequentemente, o menor RMSD. Assim, dois conjuntos *v* e *w* de *n* pontos, o RMSD é definido pela Equação 2.3 e o valor é dado em Å.

$$\begin{aligned} RMSD(v,w) &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}||v_i - w_i||^2} \\ &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}((v_{ix}-w_{ix})^2 + (v_{iy}-w_{iy})^2 + (v_{iz}-w_{iz})^2)} \end{aligned} \tag{2.3}$$

3. MOTIVAÇÃO E OBJETIVOS

3.1 Motivação

Embora as técnicas experimentais para determinar a estrutura tridimensional das proteínas tenham evoluído, utilizá-las para identificar a estrutura 3D de todas as proteínas se torna inviável. Desta forma, abordagens computacionais são, provavelmente, a única forma de preencher a lacuna entre o número de sequências depositadas e o número de estruturas tridimensionais conhecidas. Além disso, o uso de recursos computacionais apresenta opções menos custosas em condições financeiras e temporais. O problema da PSP apareceu na década de 60 e apenas nos últimos anos apresentou soluções aproximadas aos resultados obtidos em métodos de determinação experimental. Assim sendo, as tendências observadas e analisadas nos últimos CASP apresentaram avanços na predição da estrutura 3D, utilizando a informação de contato entre resíduos de aminoácidos (Moult et al., 2018). Portanto, incorporar as informações de contato em métodos de predição de estruturas tridimensionais pode resultar em predições mais próximas as estruturas determinadas experimentalmente. Ademais, há um aumento de interessados no problema, na solução e nas possibilidades de novos estudos que esses podem gerar assim representados pelo crescimento do CASP.

3.1.1 Objetivo Geral

O objetivo geral deste trabalho foi incorporar informações sobre os contatos entre os resíduos de aminoácidos ao método CReF a fim de melhorar a qualidade de suas predições.

3.1.2 Objetivos Específicos

Para que o objetivo geral fosse alcançado, os seguintes objetivos específicos foram definidos. Os objetivos 1 até 3 são contemplados no Capítulo 6, 4 no Capítulo 7 e 5 e 6 no Capítulo 8.

1. Identificar possíveis problemas pré-existentes nas etapas do CReF;
2. Investigar os preditores de contato;
3. Definir a etapa para incorporar as informações de contato ao método;

4. Criar um modelo e uma função para selecionar a melhor conformação com as informações de contato;
5. Realizar simulações e analisar as estruturas com o uso das informações de contatos;
6. Comparar os resultados com aqueles obtidos inicialmente pelo CReF.

4. PREDIÇÃO DA ESTRUTURA 3D DE PROTEÍNAS COM O MÉTODO CReF - VERSÃO INICIAL

Entre os diversos métodos de predição *in silico* o *Central Residue Fragment-based method* (CReF) foi proposto. A primeira versão desse englobava princípios dos métodos *de novo* para encontrar novos padrões de enovelamento e dos métodos baseados em modelagem comparativa para obter resultados mais acurados. A fim de reduzir o número de variáveis a serem manipuladas e a sua complexidade, o CReF representava a cadeia polipeptídica por meio dos ângulos de torção da cadeia principal ϕ e ψ (Dorn, 2008). Sendo assim, o método CReF era composto por oito etapas representadas na Figura 4.1.

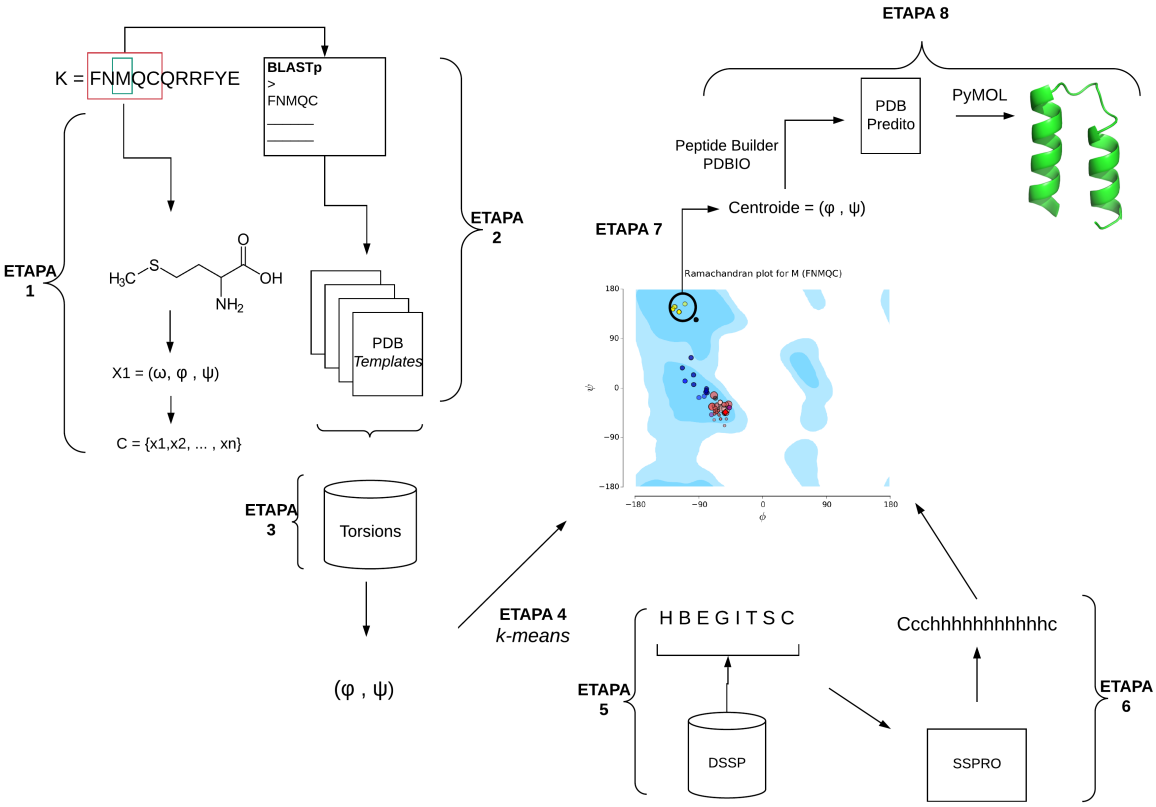


Figura 4.1 – Representação geral do CReF. O método era divido em oito etapas. A sequência de resíduos de aminoácidos era lida em fragmentos de 5 resíduos. O fragmento era alinhado por meio do Blastp e os templates que tinham relação evolutiva com aquele fragmento eram excluídos, utilizando no máximo 100 templates para obter os ângulos ϕ e ψ do resíduo central de cada fragmento. A predição da estrutura secundária da sequência alvo era realizada pelo SSPRO, utilizando a classificação do DSSP. Assim, as informações dos ângulos de cada template eram agrupadas por meio do *k-means* no mapa de Ramachandran. Após isso, um grupo era selecionado através dos rótulos definidos pelo DSSP e o centroide do grupo era utilizado para representar os ângulos de torção do resíduo central daquele fragmento. A última etapa consistia em construir a estrutura tridimensional convertendo as informações obtidas em coordenadas x, y, z no formato PDB. Fonte: Autora.

Desse modo, na ETAPA 1, a sequência alvo era dividida em fragmentos de tamanho *l*. Em geral, o tamanho do fragmento apresentado era de 5 resíduos. Para cada fragmento obtido, apenas os ângulos de torção do resíduo central eram considerados e armazenados, o que indicava que quanto maior o fragmento, maior seria o número de resíduos nas extremidades que não seriam preditos.

Na ETAPA 2, cada fragmento obtido (alvo) na etapa anterior era alinhado com as estruturas disponíveis no PDB (template) através do BLASTp (*Basic Local Alignment Search Tool Protein*) (Altschul et al., 1990). O BLASTp utilizava a matriz de substituição PAM30 (*Point Accepted Mutation*) (Dayhoff et al., 1978). Uma matriz de substituição avalia a qualidade do alinhamento. Assim, ela pontuava o alinhamento entre os pares de resíduos da sequência alvo e do template representando as taxas relativas de substituições evolutivas. Além disso, apenas os templates com tamanho *l* igual ao fragmento e sem relação evolutiva com a sequência de interesse eram considerados.

Adicionalmente, os templates relacionados evolutivamente, ou seja, com pelo menos 30% de similaridade, eram descartados por meio de uma lista de exclusão. Além disso, o número de templates, em formato PDB, a ser utilizado para cada fragmento era de 100 templates e para cada um dos arquivos PDBs de cada fragmento obtido na ETAPA 2, os ângulos de torção do resíduo central eram calculados. Os cálculos dos ângulos eram realizados por meio do software *Torsions* (Grupo do Dr. Andrew C. R. Martim) e armazenados em uma tupla $t_i = (\phi, \psi)$, sendo esta a ETAPA 3.

O conjunto de tuplas de um fragmento (tuplas calculadas para no máximo 100 templates de cada fragmento) e a informação da estrutura secundária classificada pelo DSSP eram usadas para agrupar cada tupla por meio do algoritmo de agrupamento *k-means* (ETAPA 4), buscando identificar as regiões onde cada tupla se concentrava no mapa de Ramachandran, mesmo provenientes de diferentes estruturas secundárias.

Apesar do DSSP não prever a estrutura secundária, ele era utilizado para rotular e classificar os agrupamentos (ETAPA 5). Sendo assim, a predição da estrutura secundária da sequência alvo era predita pelo SSPRO e executado localmente (ETAPA 6) (Magnan e Baldi, 2014). Após o agrupamento nas etapas anteriores, um grupo era selecionado para representar os ângulos de torção do resíduo central do fragmento (ETAPA 7). Assim, a seleção era dada pela informação da estrutura secundária definidos na Tabela 2.1 e os ângulos ϕ e ψ. O valor do centroide de cada grupo era utilizado para definir os valores dos ângulos do resíduo central daquele fragmento.

A partir disso, a ETAPA 8 consistia na construção da estrutura tridimensional aproximada da sequência alvo por meio da biblioteca python *PeptideBuilder* (Tien et al., 2013) e da criação do arquivo PDB da estrutura predita pela biblioteca PDBIO (Hamelryck e Manderick, 2003). Por fim, a estrutura predita era visualizada por meio do software PyMOL (Schrödinger, LLC, 2015).

5. METODOLOGIA

A metodologia aplicada nesta dissertação teve como base os métodos de predição *de novo*. Assim, para que os contatos fossem incorporados ao método CReF, questões relacionadas à avaliação e a amostragem das conformações foram levantadas. Para a avaliação, um modelo de $RMSD_{predito}$ foi desenvolvido. Para a amostragem, um módulo de simulação foi criado. Além disso, uma função de energia foi utilizada para selecionar a melhor conformação com termos baseados em um potencial atômico dependente de distância, contatos de curto, médio e longo alcance. Desse modo, a seguir, é apresentado como este trabalho trata das questões relacionadas aos contatos, a representação conformacional para a avaliação da qualidade e o problema de amostragem através de um método de simulação molecular.

5.1 Representação de Contato

Para que as informações de contato entre os resíduos de aminoácidos fossem incorporadas ao método, a seleção dos preditores de contato a serem analisados foi realizada. Esses são representados na Tabela 5.1.

Tabela 5.1 – Preditores de contato. Fonte: Autora.

Preditor	Disponível em:	Referência
CCMpred	https://github.com/soedinglab/ccmpred	(Seemayer et al., 2014)
Deepcontact	https://github.com/largelymfs/deepcontact	(Liu et al., 2018)
DeepCov	https://github.com/psipred/DeepCov	(Jones e Kandathil, 2018)
DeepMetaPSICOV	http://bioinf.cs.ucl.ac.uk/psipred/	(Kandathil et al., 2019)
DNCON2	http://sysbio.rnet.missouri.edu/dncon2/	(Adhikari et al., 2018)
FreeContact	https://rostlab.org/owiki/index.php/FreeContact	(Kaján et al., 2014)
GREMLIN	http://gremlin.bakerlab.org/	(Ovchinnikov et al., 2014)
PconsC3	https://pconsc3.bioinfo.se/	(Skwark et al., 2014)
RaptorX-Contact	http://raptorx.uchicago.edu/ContactMap/	(Ma, 2015)
ResPRE	https://github.com/leeyang/ResPRE	(Li et al., 2019)
SPOT-contact	https://sparks-lab.org/server/spot-contact/	(Hanson et al., 2018)

Inicialmente, haveria a seleção de um único preditor de contatos a ser utilizado pelo CReF, seguindo os seguintes critérios:

- **Entrada de dados:** O preditor de contato deveria usar como dado de entrada apenas a sequência de resíduos de aminoácidos, sem depender que as informações de MSA fossem fornecidas previamente.
- **Método de predição:** Os métodos de predição de contato deveriam utilizar informações evolutivas junto a técnicas de aprendizado de máquina.

- **Arquivos de saída:** Os preditores deveriam gerar o arquivo de contatos no formato RR.
- **Execução:** Os métodos de predição não deveriam precisar ser rodados localmente. Assim, sendo disponibilizados via internet.

Contudo, ao selecioná-los, percebeu-se a inviabilidade da predição de contato ser realizada durante a execução do CReF. Assim, optou-se por utilizar as informações de contato no formato RR como entrada de dados ao método, viabilizando o uso das informações de contato e possibilitando o uso de diferentes métodos de predição de contato, dependendo apenas que o formato de arquivo fosse compatível. Assim, o preditor selecionado para ser utilizado como exemplo de entrada foi selecionado levando em consideração:

- **Disponibilidade do preditor:** O preditor precisa estar disponível online.
- **Disponibilidade do arquivo de contatos:** O arquivo de contato no formato RR precisava ser gerado.
- **Acesso aos resultados:** Os resultados deveriam estar disponíveis para consulta posterior.

Assim, a avaliação dos preditores de contato é descrita na Tabela 5.2. Os preditores *DeepMeta*PSICOV, RaptorX-*Contact* e SPOT-*contact* atenderam a todos os critérios. Assim, pelo tempo de resposta dos preditores, o *DeepMeta*PSICOV foi selecionado. O RaptorX-*Contact* realiza a predição da estrutura 3D junto a predição dos contatos e o SPOT-*contact* disponibiliza dados relacionados aos MSA e ao perfil da família proteica. Assim, o tempo de envio dos resultados desses é consideravelmente maior.

Tabela 5.2 – Avaliação dos preditores de contato considerando os critérios de disponibilidade do preditor, disponibilidade do arquivo de contato e acesso aos resultados posteriormente. Fonte: Autora.

Preditor	Disponibilidade	Arquivo de Contato	Acesso Posterior
CCMpred			
Deepcontact			
DeepCov			
DeepMetaPSICOV	X	X	X
DNCON2	X		
FreeContact			
GREMLIN	X		
PconsC3	X		
RaptorX-Contact	X	X	X
ResPRE			
SPOT-contact	X	X	X

5.1.1 Incorporação dos contatos ao CReF

Com o preditor de contato selecionado, a incorporação das informações de contato ao método CReF foi realizado conforme a Figura 5.1. Assim, o arquivo de contato no formato RR gerado pelo preditor é passado como arquivo de entrada ao CReF. Após isso, o arquivo é lido e dado um *cutoff* para a probabilidade do par de resíduos estar em contato, o par é separado. Adicionalmente, a distância dos resíduos do par na sequência é calculada e se essa for maior que 5, o par de resíduos é separado em contato de curto, médio ou longo alcance como descrito na tabela 2.2. Finalmente, essa informação é armazenada no formato de contato criado e que será explicado junto aos arquivos de saída no capítulo 6.

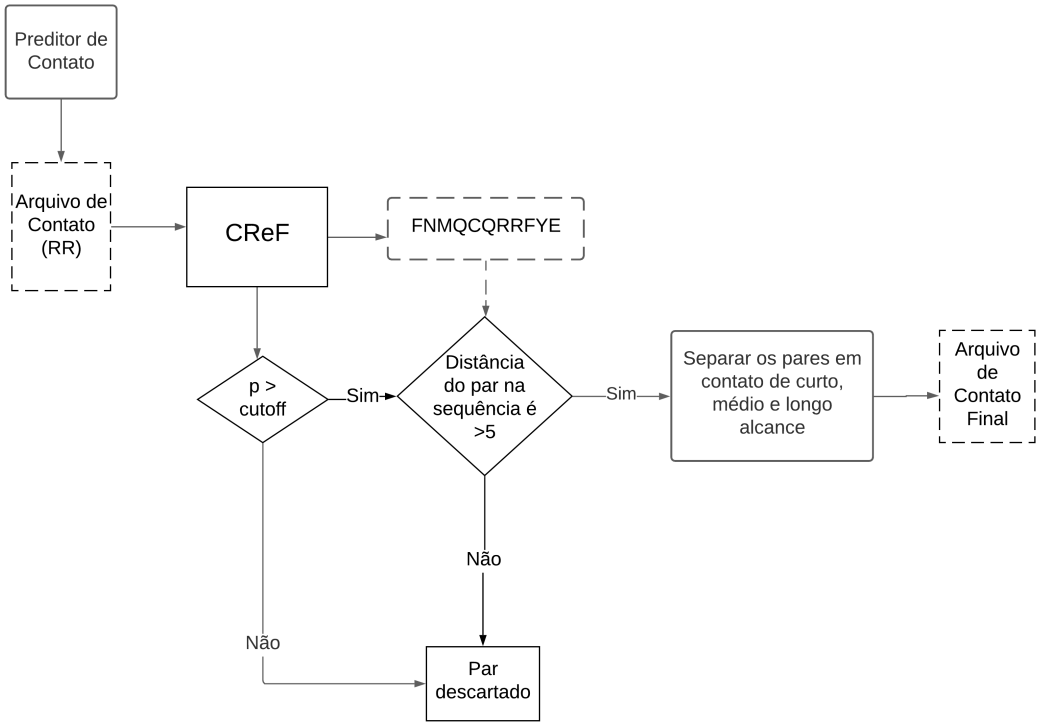


Figura 5.1 – Esquema de incorporação das informações de contato ao método CReF. Fonte: Autora.

Assim, é importante enfatizar que consideramos que dois resíduos de aminoácidos fazem algum tipo de contato se, e somente se:

1. A probabilidade de contato entre os dois resíduos for maior que o *cutoff*;
2. A distância sequencial entre os resíduos de aminoácidos for maior que 5 resíduos;

5.2 Representação Conformacional

Quando a estrutura nativa é conhecida, selecionar a melhor conformação, dentro de um conjunto de conformações que a descrevem, não parece ser uma tarefa difícil. No entanto, quando a estrutura nativa não está disponível ou não é conhecida, o cenário é outro. Assim, uma das dificuldades em predições de estrutura baseada em primeiros princípios é a dificuldade em projetar campos de força precisos que reconheçam que a estrutura nativa é aquela com valor de energia mais baixo (Anfinsen, 1973).

Para a construção de um campo de força, um conjunto de estruturas de proteínas não nativas (chamada de *decoys* de estrutura) podem ser usadas para orientar o desenvolvimento. Entretanto, construir os *decoys* apropriados não é uma tarefa trivial (Deng et al., 2016). Desse modo, o programa 3DRobot foi utilizado com o objetivo de construir um conjunto de proteínas com variabilidade conformacional (*decoys*) e esse auxiliar no desenvolvimento de um modelo de avaliação e seleção de conformações considerando as informações de contato.

5.2.1 3D Robot

O 3DRobot é um algoritmo para a geração de *decoys* de estrutura 3D de proteínas. Esse consiste em três etapas: i) identificação de templates, ii) simulação para a remontagem da estrutura por fragmentos (através de *Replica-Exchange*) e iii) seleção e refinamento dos *decoys*. O protocolo do 3DRobot é representado na Figura 5.2.

Inicialmente, os *decoys* seriam gerados pela autora, porém verificou-se que a construção desses levava um tempo considerável de algumas horas para uma proteína de tamanho próximo a 40 resíduos. Desse modo, o conjunto foi selecionado a partir dos dados disponibilizados por Deng et al. (2016)[10].

Sendo assim, o conjunto de proteínas com variabilidade conformacional selecionado inclui *decoys* gerados pelo 3DRobot de proteínas providas dos preditores I-TASSER (Zhang e Zhang, 2010) e ROSETTA (Simons et al., 1997). O primeiro apresenta 400 *decoys* e o segundo 100 *decoys* para cada proteína. Com isso, o conjunto selecionado é composto por 100 proteínas com 400 ou 100 *decoys* (dependendo do preditor pela qual esses foram gerados) contabilizando 22600 conformações.

Além disso, para cada proteína o 3DRobot gera um arquivo (*rst.dat*) que apresenta uma lista dos *decoys* e o valor de RMSD para cada um. O tamanho das proteínas altera de 47 até 146 resíduos, com estruturas do tipo somente hélice-α, somente folha-β ou hélice-α e folha-β. O conjunto completo pode ser conferido no Apêndice C.

[10]https://zhanglab.ccmb.med.umich.edu/3DRobot/

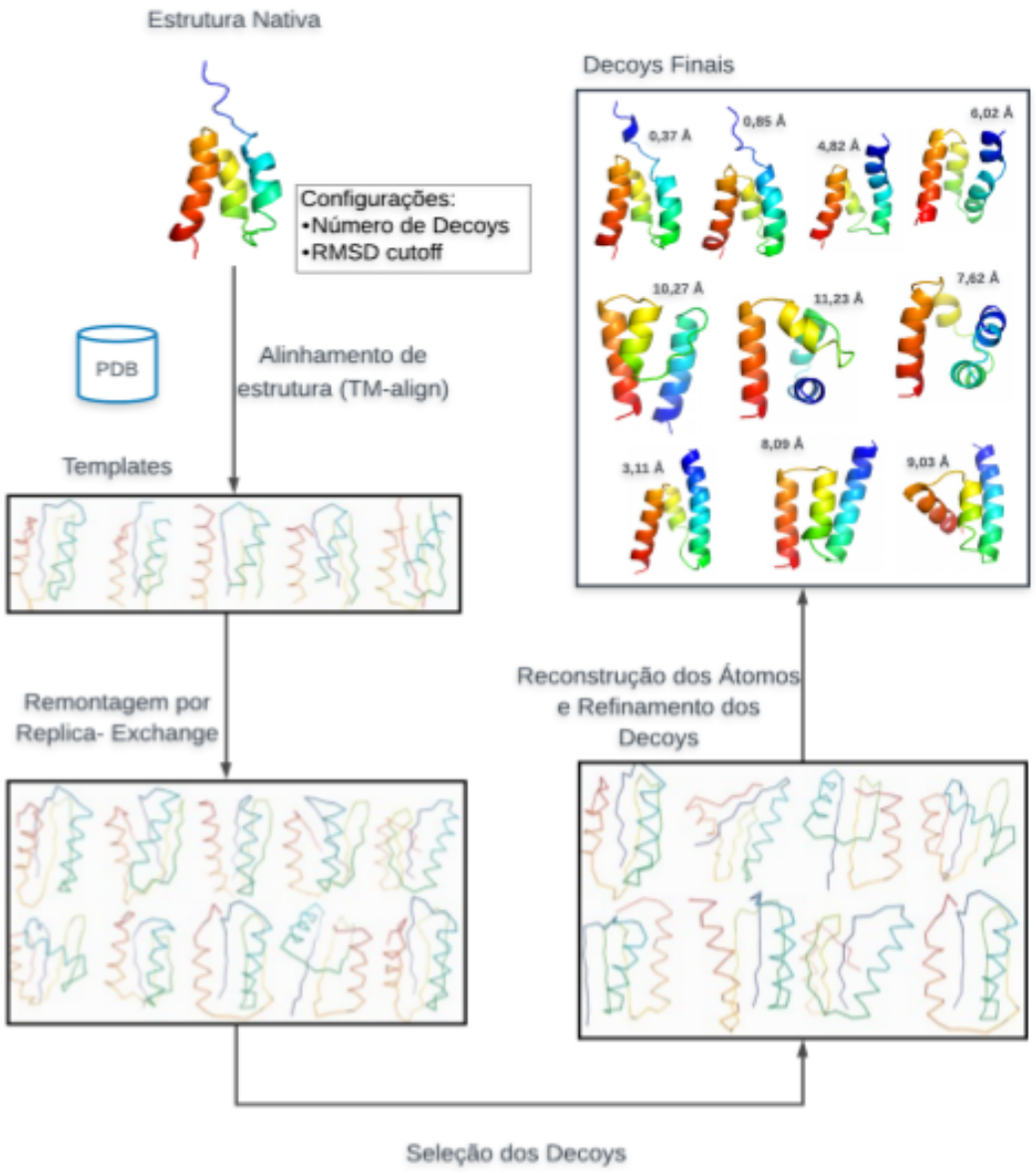


Figura 5.2 – Protocolo para construção de *decoys* de estrutura de proteína através do programa 3DRobot. Proteína usada PDB ID: 1GAB. Fonte: Adaptada de Deng et al. (2016).

Com o conjunto definido, um método de seleção, que busca a melhor conformação dado um conjunto de conformações para uma proteína, foi desenvolvido com o objetivo de avaliar a qualidade dessas.

5.2.2 RMSD Predito

A análise das conformações, levando em consideração a configuração energética, é um passo importante para separar e selecionar àquelas que estão próximas ou longe do estado nativo. Assim, uma função de energia permite que essas ações sejam realizadas. Pensando nisso, uma função que considerasse as informações de contato somados

a um potencial atômico dependente de distância (RW, do inglês *Random-Walk*)[11] (Zhang e Zhang, 2010), o qual apresenta forte correlação com o RMSD de *decoys* de estrutura, foi elaborada.

Desse modo, essa considera os valores de contato (curto, médio e longo alcance) e o potencial RW. Contudo, ao invés de retornar um valor de energia potencial total, pensou-se em retornar um valor que correspondesse ao RMSD. Assim, quanto menor o valor de RMSD predito, melhor a conformação, considerando-o como uma pseudo-energia[12]. Desse modo, a Equação 5.1 descreve o cálculo para o RMSD predito.

$$RMSD_{predito} = RW + c + m + l \tag{5.1}$$

Onde $RMSD_{predito}$ representa o valor obtido a partir da soma dos valores de *RW*, o potencial atômico, *c*, contato de curto alcance, *m* ,contato de médio alcance e *l*, contato de longo alcance.

Todavia, quando não se conhece a estrutura 3D nativa não é possível calcular o RMSD. Para isso, um modelo simples utilizando redes neurais artificiais foi desenvolvido. Assim, dado os valores de contato de curto, médio e longo alcance e um potencial RW, o modelo retorna o valor de RMSD correspondente.

Conjunto de Dados

Para gerar o modelo, o conjunto de dados usado foi os 22600 *decoys* gerados a partir do 3DRobot com os valores de contato e RW correspondentes. Os valores de contato de curto, médio e longo alcance foram calculados a partir do arquivo de contatos nativo gerado no formato indicado na Figura 6.6. Assim, se o par dito em contato em cada *decoy* estivesse presente no arquivo de contatos nativo da proteína correspondente, o par era considerado pontuando aquele *decoy*. Os valores de RW foram calculados conforme Zhang e Zhang (2010)[13], considerando os *decoys* com e sem cadeia lateral.

A partir disso, surgiu a necessidade de normalizar os valores obtidos para os contatos e para o RW, para que assim fosse possível comparar os resultados obtidos por proteínas diferentes. Desse modo, os valores de RW (com e sem cadeia lateral) foram normalizados pelo número de átomos e os valores de contato pelo número de pares em contato correspondente daquela proteína. Ou seja, se uma proteína apresentasse 5 pares em contato de curto alcance, o valor para contato de curto alcance do *decoy* seria dividido por 5.

[11] A partir desse momento, quando a autora escreve RW, ela refere-se ao potencial atômico dependente de distância.

[12] Aqui a autora considera o valor de RMSD predito como pseudo-energia, pois esse apresenta o mesmo comportamento para a seleção das conformações, no qual quanto menor o valor mais próximo da estrutura nativa.

[13] Código disponível em: https://zhanglab.ccmb.med.umich.edu/RW/

A Figura 5.3 representa o esquema para a construção do conjunto o qual é separado em 5 etapas.

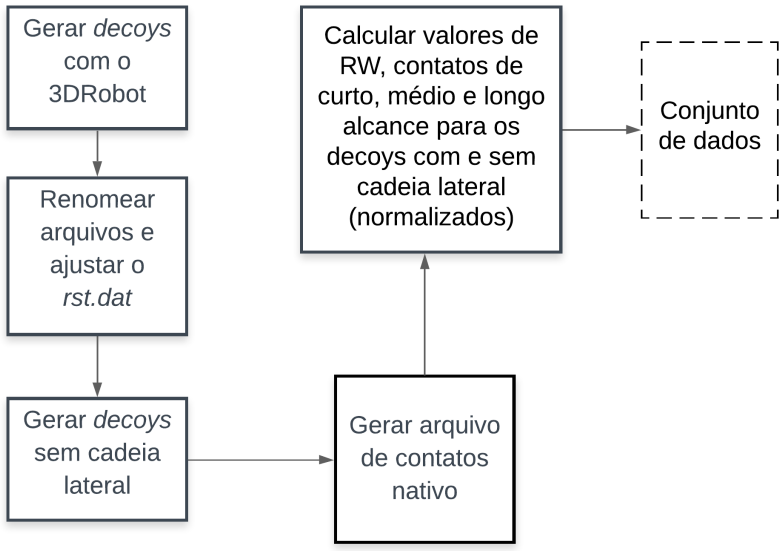


Figura 5.3 – Esquema para a construção do conjunto de dados com os valores de RW e contatos (curto, médio e longo alcance) para cada *decoy* do conjunto gerado pelo 3DRobot. Fonte: Autora.

Modelo

Com o conjunto de dados disponível, modelos que representassem o comportamento desse foram criados através do uso de redes neurais artificias do tipo totalmente conectada (FC, do inglês *Fully Connected*). As redes neurais artificiais foram desenvolvidas inspiradas na capacidade do cérebro humano em processar informações (Norvig e Russell, 2014). É um modelo matemático não-linear baseado no funcionamento dos neurônios (Haykin, 1999). O primeiro modelo artificial de um neurônio biológico foi desenvolvido por McCulloch e Pitts (McCulloch e Pitts, 1943).

O modelo é representado por um vetor de entradas que recebe pesos. A soma ponderada das entradas é submetida a uma função de ativação que determina se o valor é maior que o limiar do neurônio, propagando o resultado para as ligações de saída. A representação do modelo foi desenhada por Norvig e Russel e está representada na Figura 5.4 (Norvig e Russell, 2014). Assim, a arquitetura aqui consiste em uma série de camadas com neurônios totalmente conectados.

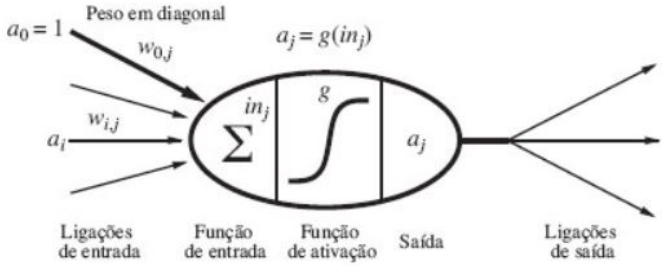


Figura 5.4 – Representação em forma de desenho do primeiro modelo artificial de um neurônio biológico. Os valores de entrada recebem pesos. A soma ponderada é submetida a uma função de ativação que determina se o valor obtido deverá ser propagado as ligações de saída. Fonte: Norvig e Russell (2014).

A escolha do tipo de rede foi dada pela simplicidade, já que a quantidade de *decoys* e de variáveis não era grande. A partir disso, três arquiteturas com diferença no número de neurônios e camadas foram testadas. Essas consistiam em i) camada de entrada; ii) camadas ocultas e iii) camada de saída. Como não havia conhecimento prévio sobre os valores a assumir nas camadas ocultas, esses foram escolhidos arbitrariamente.

- Arquitetura 1: 3 camadas ocultas com 32, 64, e 4 neurônios;
- Arquitetura 2: 6 camadas ocultas com 32, 64, 64, 64, 128 e 128 neurônios;
- Arquitetura 3: 3 camadas ocultas com 128, 128 e 128 neurônios.

Para cada uma das arquiteturas, três modelos foram gerados. Esses consideravam, respectivamente, os valores de RW com cadeia lateral, contatos de curto, médio e longo; valores de RW sem cadeia lateral, contatos de curto, médio e longo alcance; e somente contatos de curto médio e longo alcance. Os principais hiperparâmetros utilizados foram 25 épocas, 32 para tamanho do lote (*batch size*), função de ativação ReLU (unidade linear retificada), otimizador *Adam* com taxa de aprendizado de 0,001.

Desse modo, o conjunto de dados foi dividido em 70% para treino e 30% para teste. Para validação, 20% do conjunto de treino foi utilizado. Assim, para avaliar a qualidade e comparar os modelos, as métricas em relação ao Erro Médio Absoluto (MAE, do inglês *Mean Absolute Error*), Erro Médio Quadrático (MSE, do inglês *Mean-Squared Error*), R^2 ajustado foram utilizadas.

Métricas de Avaliação da Qualidade

Para a seleção do modelo foram utilizadas as métricas citadas anteriormente. Assim, o R^2 ajustado (Equação 5.2) é uma métrica que calcula qual a porcentagem da variância dos dados que pode ser prevista pelo modelo. Desse modo, quanto maior o seu

valor, maior é a capacidade do modelo em explicar os dados. Na Equação, *N* representa o número de instâncias e *p* o número de variáveis de entrada.

$$R_a^2 = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1} \tag{5.2}$$

As outras duas métricas selecionadas foram o MSE e o MAE. O primeiro é a média das distâncias entre o valor predito e real ao quadrado (Equação 5.3). Assim, quanto maior o valor, pior o modelo. Do mesmo modo, o segundo utiliza a média das distâncias entre os valores preditos e reais, porém sua interpretação é mais intuitiva (Equação 5.4). Com isso, ambas foram utilizadas. Nas Equações, $\hat{Y}i$ é o valor predito e Yi o valor real.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\widehat{Y}i - Yi)^2 \tag{5.3}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\widehat{Y}i - Yi| \tag{5.4}$$

5.3 Simulação Molecular

Para resolver o problema de amostragem do CReF, um módulo de simulação molecular foi desenvolvido paralelamente.

5.3.1 PepDice3

PepDice3 é um módulo desenvolvido que permite o programa CreF executar simulações e fazer manipulações na estrutura da proteína. As principais funcionalidades são:

- Alterações dos ângulos da cadeia principal (ϕ, ψ e ω);
- Alteração dos ângulos das cadeias laterais (não usados nesse trabalho);
- Inserção de fragmentos;
- Cálculo de energia (podendo incluir diferentes termos);
- Imposição de restrições de estrutura secundária baseada em potenciais harmônicos (comprimento e ângulos de ligação);

- Simulações por Método de Monte Carlo com decaimento de temperatura por *simulated annealing*.

A função de energia utilizada no programa PepDice3 inclui os termos RW, contato de curto, médio e longo alcance conforme a Equação 5.1. Os pesos para cada tipo de contato são ajustados de acordo com as necessidades do usuário.

No desenvolvimento desse, um modelo para a representação da proteína inspirado nos campos de força dos programas QUARK (Xu e Zhang, 2012), UNRES (Czaplewski et al., 2018) e CABSDOCK (Blaszczyk et al., 2016) foi utilizado. A estrutura da proteína inclui os átomos da cadeia principal e também os carbonos β dos resíduos de aminoácidos (exceto glicina). Os outros átomos de cadeia lateral são desconsiderados no modelo.

Desse modo, o espaço de busca por conformações e as penalidades por sobreposições de átomos (*clashes*) são reduzidos. Apesar da cadeia lateral apresentar papel importante para o enovelamento, as principais interações que mantêm a topologia são realizadas pelos átomos pertencentes a cadeia principal por meio de ligações de hidrogênio. Assim, as interações da cadeia lateral são mimetizadas no termo de energia que considera as informações de contato. As simulações são realizadas utilizando uma adaptação do método de monte carlo - metropolis (Metropolis et al., 1953), *simulated annealing* (Van Laarhoven e Aarts, 1987).

5.3.2 Simulação por Monte Carlo

É dito como método de Monte Carlo (MMC) qualquer método estatístico que baseia-se em amostragens aleatórias numerosas. O método permite explorar aleatoriamente o espaço de configurações do sistema, aqui sendo, o espaço de conformações possíveis para uma proteína.

A primeira simulação computacional foi realizada há mais de meio século por Metropolis et al. (1953). Eles desenvolveram um método que hoje é conhecido por algoritmo Metropolis. Esse, provavelmente, seja o MMC mais utilizado na física. Infelizmente, o método não permite o cálculo de quantidades dependentes do tempo como ocorre na dinâmica molecular. Apesar disso, simulações por MMC permitem que o espaço de busca a ser explorado seja maior, apresentando menores chances de cair em mínimos locais.

Levando em consideração as etapas do algoritmo metropolis, uma adaptação conhecida é o *simulated annealing*. Essa é uma técnica probabilística para otimizar o ótimo global que mimetiza um processo térmico. Assim sendo, esse consiste em duas etapas:

1. Aumentar a temperatura T;

2. Resfriar o sistema gradativamente até que esse estabilize ou chegue a um critério estabelecido.

Na etapa 2, é preciso acompanhar e controlar o resfriamento, pois é nela que se espera que os átomos se organizem em uma estrutura uniforme com energia mínima. Desse modo, quanto maior o valor atribuído a T, maior será o componente aleatório na próxima conformação escolhida.

Aqui, T inicia em 100 e sofre um decaimento de 10 até que T_{final} seja 10 (critério de parada). Para cada T ocorre 100 passos de monte carlo (um ciclo). Assim, totalizando 9 ciclos de 100 passos cada. Com o decréscimo da temperatura, os critérios de metropolis para que aceite uma nova conformação com energia mais alta ficam mais rigorosos. Para o algoritmo, geralmente, a conformação inicial é gerada aleatoriamente, porém aqui, a conformação gerada pelo CReF é utilizada para iniciar a simulação. Desse modo, o algoritmo metropolis adaptado para as simulações segue as seguintes etapas:

1. Selecionar a conformação inicial gerada pelo CReF denominada C_1 e calcular a energia correspondente E_1;
2. Gerar uma nova conformação C_{n+1} através da amostragem de fragmentos e calcular a energia E_{n+1};
3. Verificar se E_{n+1} é menor que E_n; Se afirmativo, C_{n+1} é aceita, caso contrário segue as seguintes etapas:
   3.1. Gerar número aleatório entre 0 e 1;
   3.2. Verificar se o valor obtido na etapa 3.1 é menor que o valor P. Esse é calculado conforme a Equação 5.5[14] , onde

$$P = e^{\frac{\Delta E}{kT}} \tag{5.5}$$

   Se afirmativo, C_{n+1} é aceita, caso contrário é rejeitada.
4. Repetir as etapas 2 e 3 até o critério de parada.

5.3.3 Conjunto de Proteínas

Com o objetivo de avaliar a qualidade das simulações e da efetividade das informações de contato para a aceitar as conformações geradas, um conjunto de proteínas com

[14] k é a constante de Boltzmann e T a temperatura

diferentes estruturas secundárias foi selecionado e testado. O comportamento dos valores de RMSD e energia (calculado através dos valores de RW e contatos de curto, médio e longo alcance) ao longo das simulações foi analisado. Para todos os estudos de caso apresentados nesta etapa, as bibliotecas de fragmentos utilizadas foram construídas a partir da própria estrutura nativa e um conjunto de *decoys* gerados pelo programa 3DRobot. Isso garantiu que houvesse um controle na variabilidade dos fragmentos gerados.

Ao todo, cada conjunto continha um total de sete (7) conformações, onde os valores de RMSD variavam entre 0 e 11Å. A presença da própria estrutura nativa é importante para avaliar a robustez do método em encontrar a estrutura correta tendo a informação necessária para isso. Os *decoys* de baixo RMSD são, em geral, topologicamente corretos, apresentando pequenas variações na estrutura (semelhantes às variantes estruturais encontradas em modelos da RMN) e atuam conferindo variabilidade aos fragmentos gerados.

Por fim, os *decoys* com altos valores de RMSD são topologicamente deturpados e atuam gerando fragmentos indesejáveis com objetivo de introduzir erros nas bibliotecas de fragmentos. Desse modo, dadas as conformações aceitas, aquelas com menor RMSD, menor energia e a última conformação da simulação são mostradas no capítulo 8. Para cada uma das proteínas, foi identificado o quão perto as conformações amostradas estão do modelo de referência e se aquela com menor energia apresenta também o menor RMSD.

6. MELHORIAS NA IMPLEMENTAÇÃO

Neste capítulo, são mostrados os resultados relacionados ao funcionamento e a execução do método, a identificação de problemas pré-existentes que pudessem explicar os problemas nas predições e os novos arquivos de saída gerados pelo método.

6.1 Funcionamento e Execução

Inicialmente, para utilizar o método CReF, era necessário que cada programa ou dependência fosse instalada manualmente via terminal. Com isso, diversos problemas ocorriam como dependência de bibliotecas com versões específicas e links desatualizados. Assim, para facilitar a instalação e o uso, um instalador foi desenvolvido.

Após isso, mas antes de qualquer alteração no método em si, um conjunto de proteínas foi selecionado e submetido ao preditor, para que assim, fosse conhecido o estado atual desse. As proteínas foram escolhidas através dos estudos anteriores realizados com o método (Apêndice A) e aquelas disponibilizadas pelos CASP10, CASP11, CASP12 e CASP13 (Apêndice B). Como o preditor apresentava problemas no dobramento das proteínas, para o conjunto selecionado, optou-se por proteínas com tamanho até 105 resíduos de aminoácidos. Sendo assim, proteínas com tamanho superior a 105 resíduos de aminoácidos não foram selecionadas. Para exemplificar as execuções, as proteínas de código PDB ID: 1ZDD, 1YWJ e 1CSP são apresentadas a seguir.

6.1.1 Proteína PDB ID: 1ZDD

A proteína de código PDB ID: 1ZDD foi utilizada como teste no CReF desde sua primeira versão. A proteína é composta por 34 resíduos de aminoácidos e apresenta estruturas secundárias regulares do tipo hélice-α. A estrutura experimental, a estrutura predita pelo CReF e a sobreposição das duas conformações podem ser visualizadas na Figura 6.1.

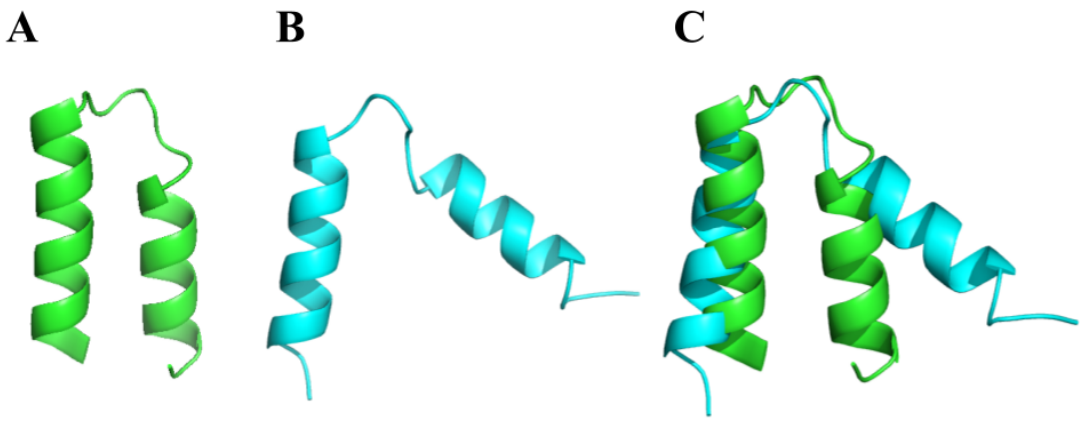


Figura 6.1 – Comparação de conformações predita e experimental da proteína 1ZDD: (A) estrutura experimental da 1ZDD, (B) conformação predita com 6 grupos e 200 templates e (C) sobreposição da estrutura experimental (verde) e predita (ciano) (RMSD pelo CReF: 4,722 Å alinhando 158 átomos, considerando todos os átomos RMSD: 8,26 Å). Fonte: Autora.

É possível identificar que a estrutura predita está razoavelmente próxima da estrutura experimental, porém problemas nas regiões de estrutura secundária irregular não permitem que a proteína predita fique mais próxima. Alterações no número de templates e de identidade para a exclusão dos templates foram realizadas, mas nenhuma apresentou melhora. Além disso, ao deixar todas as informações, ou seja, usando as informações da própria proteína, mesmo assim, a proteína predita não chegou na estrutura experimental.

6.1.2 Proteína PDB ID: 1YWJ

A proteína de código PDB ID: 1YWJ é composta por 41 resíduos de aminoácidos e apresenta estruturas secundárias regulares do tipo folhas-β. A estrutura experimental, a estrutura predita pelo CReF e a sobreposição das duas conformações podem ser visualizadas na Figura 6.2.

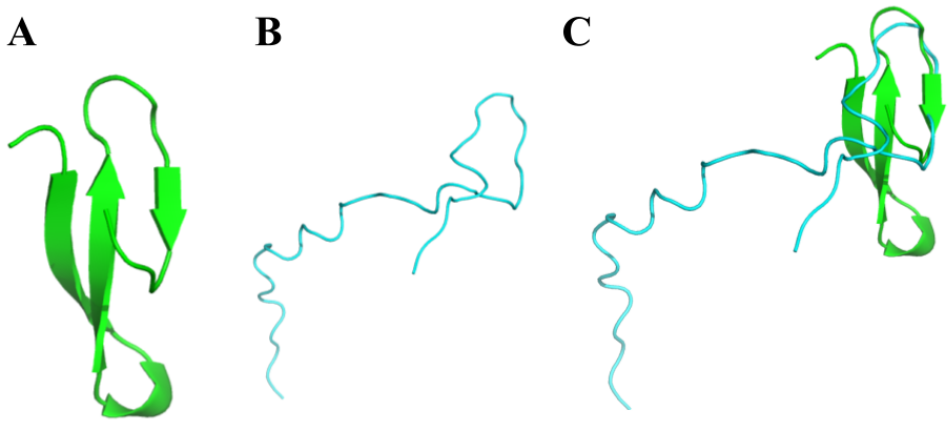


Figura 6.2 – Comparação de conformações predita e experimental da proteína 1YWJ: (A) estrutura experimental da 1YWJ, (B) conformação predita com 6 grupos e 200 templates e (C) sobreposição da estrutura experimental (verde) e predita (ciano) (RMSD pelo CReF: 1,387 Å alinhando 45 átomos, considerando todos os átomos RMSD: 15,03 Å). Fonte: Autora.

Aqui, a estrutura predita não conseguiu enovelar de forma correta mesmo com alterações dos parâmetros e utilizando as informações da própria estrutura experimental. Além disso, foi possível observar que se apenas o valor obtido pela métrica de avaliação fosse considera, ou seja, sem visualizar a estrutura, poderia ser inferido que a predição estaria muito próxima da experimental. Essa ocorria, por depender, exclusivamente, do valor obtido pela sobreposição das estruturas pelo PyMOL.

Assim, o alinhamento *super* do programa é baseado em estrutura e segue uma série de refinamentos com o objetivo de melhorar o ajuste nas sobreposições, além de ser mais robusto para proteínas com baixa similaridade de sequência. Porém, como o RMSD é calculado entre duas estruturas que apresentam a mesma sequência de resíduos de aminoácidos e a mesma estrutura 3D, outro modo de avaliar a qualidade das estruturas 3D preditas poderia ser aplicado.

6.1.3 Proteína PDB ID: 1CSP

A proteína de código PDB ID: 1CSP é composta por 67 resíduos de aminoácidos e apresenta estruturas secundárias regulares do tipo folhas-β e uma quase hélice-α. A estrutura experimental, a estrutura predita pelo CReF e a sobreposição das duas conformações podem ser visualizadas na Figura 6.3.

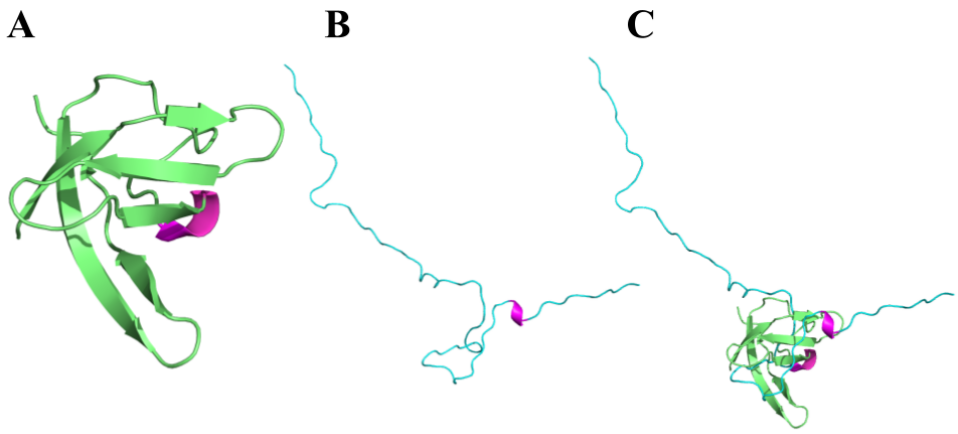


Figura 6.3 – Comparação de conformações predita e experimental da proteína 1CSP: (A) estrutura experimental da 1CSP, (B) conformação predita com 6 grupos na clusterização e 200 templates e (C) sobreposição da estrutura experimental (verde) e predita (ciano) (RMSD pelo CReF: 13,146 Å alinhando 282 átomos, considerando todos os átomos RMSD: 33,13 Å). Fonte: Autora.

Como a proteína 1CSP apresenta um número de resíduos de aminoácidos maior que as outras duas, a complexidade envolvida no seu enovelamento também aumenta. Além disso, do mesmo modo que a predição da proteína 1YWJ, as folhas-β não conseguiram ser formadas, apesar da região de hélice-α ter sido identificada.

6.1.4 Ângulos

Com as análises iniciais, questões relacionadas a problemas envolvendo a etapa dos ângulos ϕ e ψ surgiram, pois mesmo usando as informações da estrutura experimental a proteína predita não a reproduzia. Assim, modificações no algoritmo de clusterização dos ângulos foi proposto, sendo modificado de *k-means* para *Expectation Maximization* (EM), mas sem sucesso ou melhoria significativa. Desse modo, outros métodos de agrupamento não foram testados.

Apoiado nisso, especulou-se que o problema poderia estar no próprio modo como os ângulos eram captados. O par de ângulos era calculado pelo software *Torsions* e não dava controle ao CReF no modo como o cálculo era feito. Assim, alterações dos ângulos foram realizadas utilizando o módulo *PepDice3* (ver Capítulo 8).

Ademais, guardar as informações apenas dos ângulos dos resíduos centrais dos fragmentos, quando se tem as informações dos ângulos de todo o fragmento, abre um número grande de possibilidades e maiores chances de ruídos. Por exemplo, dado um fragmento utilizando 100 templates, ocorrerá 100 valores para cada um dos ângulos que não necessariamente podem ser considerados, mas que não há como saber por não utilizar

a informação do resíduo antecessor e sucessor. Assim, os valores dos ângulos de todo o fragmento começaram a ser armazenados.

6.2 Arquivos de saída

Com o intuito de acessar as informações geradas pelo CReF posteriormente, arquivos de saída foram elaborados. Assim, um arquivo de fragmentos com a informação da posição do resíduo inicial do fragmento no template, acompanhado da predição da estrutura secundária referente ao fragmento; um arquivo no formato json dos ângulos calculados pelo *PepDice3*; arquivos em formato PDB da estrutura estendida e da estrutura predita com e sem cadeia lateral (apenas Cβ); arquivo de controle de templates excluídos dependendo da identidade com a sequência alvo; e o arquivo com os contatos a serem considerados foram desenvolvidos. Desse modo, os formatos dos arquivos de fragmentos e estrutura secundária, dos ângulos e de contatos são explicados a seguir.

O arquivo de fragmentos surgiu com o objetivo de identificar a posição inicial do fragmento na sequência alvo, o template associado e a posição inicial das coordenadas. Além disso, a informação da predição da estrutura secundária para aquele fragmento foi adicionada a ser utilizada em trabalhos futuros. O formato do arquivo é representado na Figura 6.4.

Pos. Seq.	Frag.	PDB ID	Pos. Temp.	Id.	Escore	Est. Sec.
0	FNMEC	4uu8	208	80	39.0	EE-S-
0	FNMQE	3ri7	235	80	36.0	--GGG
0	FNMQE	2bf2	949	80	36.0	--THH
1	NMVCQ	2ce7	345	80	30.0	HHHHT
1	MMQCQ	3k4q	382	80	37.0	EEEET
1	VMQCQ	2ia5	237	80	37.0	EEEEE
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.

Figura 6.4 – Formato do arquivo de fragmentos. Esse apresenta sete colunas: i) pos. do frag. na sequência, ii) resíduos que fazem parte do frag., iii) PDB ID, iv) pos. inicial das coordenadas do frag. no temp., v) id. do alinhamento, vi) escore do alinhamento, e vii) predição da est. sec. do frag. Fonte: Autora.

O arquivo dos ângulos foi desenvolvido para que as informações desses fossem armazenadas em um formato compatível ao módulo *PepDice3*. Assim, o arquivo dos ângu-

los segue um formato json, no qual cada valor calculado para os ângulos ϕ, ψ e ω é atribuído a um resíduo de aminoácidos e atribuído a posição específica desse na estrutura primária. Assim, o formato é representado na Figura 6.5.

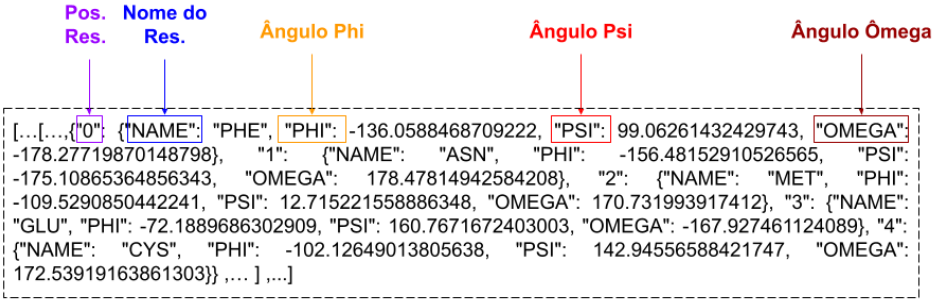


Figura 6.5 – Formato do arquivo dos ângulos. Para cada resíduo é possível identificar a posição do mesmo na sequência, o nome do aminoácido e os valores dos ângulos ϕ, ψ e ω. Fonte: Autora.

Pensando no melhor modo que as informações de contato poderiam ser extraídas, um formato de arquivo de contatos foi elaborado. Com isso, esse formato apresenta 5 colunas que representam o tipo de contato que o par de resíduos faz, assim curto (C), médio (M) ou longo (L), a posição do primeiro e segundo resíduo e a letra que representa o primeiro e o segundo resíduo do par. A necessidade de separação dos contatos surgiu com a ideia de, futuramente, realizar a ponderação da importância dos contatos dependendo do alcance desses. Além disso, armazenar a informação de qual resíduo está em determinada posição foi pensada para utilizar propriedades conhecidas de cada aminoácido e as interações que esses fazem. Assim, o formato do arquivo de contato é ilustrado na Figura 6.6.

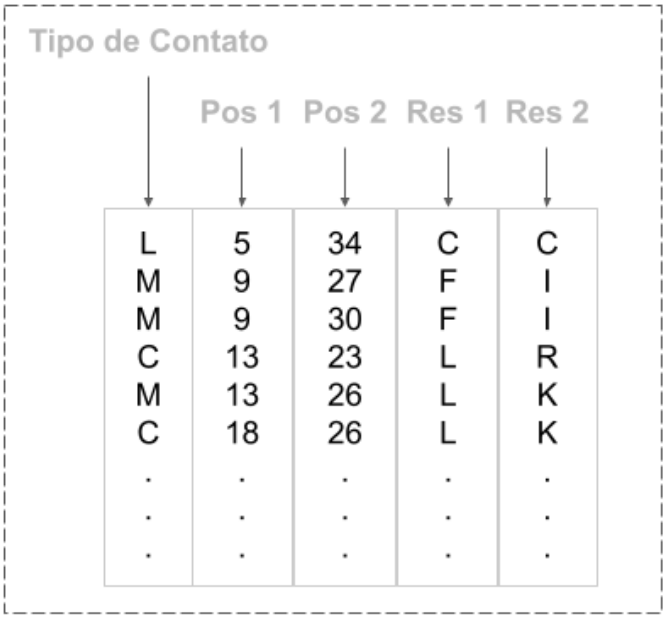


Figura 6.6 – Formato do arquivo de contato. Esse apresenta cinco colunas: i) tipo de contato, ii) posição do primeiro resíduo, iii) posição do segundo resíduo, iv) nome do primeiro resíduo e v) nome do segundo resíduo. Fonte: Autora.

Originalmente, o método CReF não possui uma rotina de amostragem, sendo seu produto final um único modelo para a estrutura aproximada da proteína de interesse (Dorn, 2008). Com os formatos de arquivos definidos e o acesso às informações, percebeu-se que as informações de contato poderiam ser não apenas utilizadas com parte da métrica que avalia a qualidade do modelo gerado, mas também como parte da função de energia para uma rotina de amostragem, que ampliaria a diversidade de estruturas e, portanto, aumentaria a robustez na busca pela conformação nativa.

7. MELHORIAS NA AVALIAÇÃO DA QUALIDADE DA ESTRUTURA

Neste capítulo, são mostrados os resultados relacionados a seleção e a análise dos modelos criados para avaliar a qualidade da estrutura através de um valor de $RMSD_{predito}$, assim como, o comportamento do modelo selecionado para um conjunto de *decoys*.

7.1 Seleção e Análise dos Modelos

Analisando os valores obtidos para cada métrica utilizada, observou-se que esses eram próximos e que mais estudos e alterações poderiam ser realizados com o intuito de identificar possíveis melhorias para o modelo. Entre os modelos gerados, o modelo 5 foi selecionado. Esse apresentou o maior valor para R^2 ajustado, menor valor para MSE e segundo menor valor para MAE (com diferença apenas na segunda casa decimal para o primeiro). Os valores obtidos para cada um dos modelos são encontrados na Tabela 7.1.

Tabela 7.1 – Tabela comparativa entre os 9 modelos gerados. Os valores de R^2 ajustado, MSE e MAE são mostrados para os modelos usando valores de RW (com cadeia lateral), RW_SC (sem cadeia lateral), contatos (curto, médio e longo alcance) e diferentes arquiteturas. Fonte: Autora.

Modelo	Informação	Arquitetura	R^2 Ajustado (%)	MSE	MAE
1	RW+ Contatos	1	70,30	3,32	1,37
2	RW_SC + Contatos	1	71,36	3,20	1,35
3	Apenas Contatos	1	69,18	3,44	1,40
4	RW+ Contatos	2	70,07	3,34	1,37
5	RW_SC + Contatos	2	**72,20**	**3,10**	1,35
6	Apenas Contatos	2	70,09	3,34	1,41
7	RW+ Contatos	3	70,16	3,33	1,39
8	RW_SC + Contatos	3	71,99	3,13	**1,34**
9	Apenas Contatos	3	69,76	3,38	1,40

7.1.1 Novo conjunto de *Decoys*

Para determinar se o modelo selecionado apresentaria um comportamento semelhante ao apresentado com o conjunto de teste, (mais informações disponíveis no Apêndice

D) 7 proteínas, que não estavam presentes no conjunto inicial, passaram pelo protocolo do 3D robot descrito na Figura 5.2. Foram gerados 10 *decoys* a partir do PDB para cada uma delas com valor máximo para RMSD de 12 Å. O tempo de execução das principais etapas são descritos na Tabela 7.2.

Tabela 7.2 – Tempo de execução das principais etapas do 3DRobot para gerar 10 *decoys* para cada uma das 7 proteínas selecionadas. No total, contabilizando 70 *decoys*. Fonte: Autora.

ID PDB	Tamanho Proteína	Tipo	Identificar Templates	Gerar Decoys	Selecionar Decoys	Refinar Decoys	Tempo Total
2M7T	33	Alfa e Beta	39m 57s	14m 53s	11s	42m 37s	1h 37m 42s
2ERL	40	Alfa	41m 48s	12m 18s	11s	47m 23s	1h 41m 44s
1GAB	53	Alfa	42m 49s	19m 42s	10s	1h 1m 52s	2h 4m 39s
1GB1	56	Alfa e Beta	43m 11s	26m 19s	12s	1h 18m 44s	2h 28m 31s
4F98	69	Beta	44m 26s	25m 24s	12s	1h 9m 37s	2h 19m 43s
1C5A	73	Alfa	46m 5s	23m 52s	12s	1h 14m 34s	2h 24m 47s
1ERV	105	Alfa e Beta	54m 34s	1h 14m 25s	12s	1h 46m 54s	3h 56m 11s

Após isso, esse conjunto foi submetido ao processo explicado na Figura 5.3. Com os valores de RW (sem cadeia lateral), contatos de curto, médio e longo alcance foram passados ao modelo selecionado. Assim, os valores de RMSD real e RMSD predito para todos os decoys das 7 proteínas são mostrados na Figura 7.1.

Desse modo, identificou-se que quando o valor de RMSD real era em torno de 10 Å, o modelo tinha maior dificuldade para predizer valores próximos ao real. Porém, como os *decoys* de alto RMSD, a priori, não deveriam ser selecionados, isso não foi constatado como problema. Assim, para os *decoys* com RMSD menor foi possível observar que o modelo apresentou valores próximos do real, sendo, algumas vezes, um pouco conservador. Adicionalmente, o comportamento da distribuição do RMSD real parece ter sido captado pelo modelo, assim como observado no conjunto de teste. Com isso, dado valor de RW, contatos de curto, médio e longo alcance, o modelo retorna um valor para RMSD, não dependendo da estrutura nativa para essa avaliação.

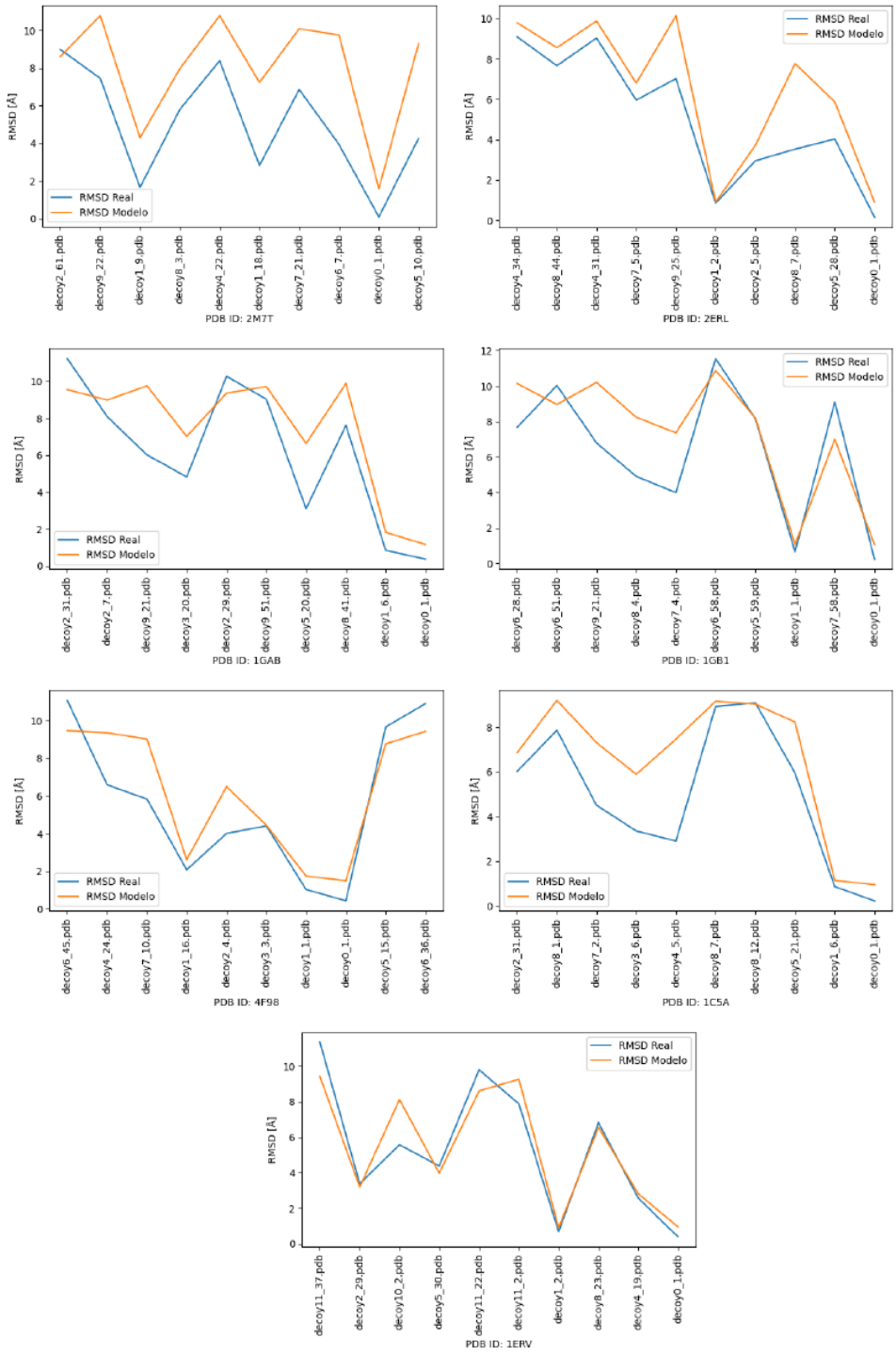


Figura 7.1 – Comparação entre o RMSD predito pelo modelo selecionado e o RMSD real. Para cada proteína há 10 *decoys*. Fonte: Autora.

8. SIMULAÇÃO MOLECULAR

Neste capítulo, são mostrados os resultados relacionados as simulações realizadas com o conjunto de proteínas selecionado utilizando a função de energia com os termos RW e contatos de curto, médio e longo alcance, assim como, as conformações que apresentaram os menores valores de RMSD, energia e a última conformação aceita na simulação.

8.1 Proteína PDB ID: 2WXC

A proteína de PDB ID: 2WXC é composta por 47 resíduos de aminoácidos e apresenta estruturas secundárias regulares do tipo hélice-α. O número de pares em contato é 46, dos quais 17 são curtos, 15 médios e 14 longos. A conformação inicial do CReF apresentou RMSD de 14,02 Å e energia de -2.131,1 kcal/mol. Ao longo da simulação quando a temperatura sofria o decaimento a conformação com menor RMSD (calculado a partir da estrutura nativa) foi constatada no frame 153 com 1,85 Å e energia -3.028,3 kcal/mol. A conformação com menor energia atribuída pela função envolvendo os contatos foi verificada no frame 315 com -3.555,9 kcal/mol e RMSD igual a 3,44 Å. A última conformação aceita pela simulação foi a de frame 332 com RMSD igual a 2,78 Å e energia de -3.519,6 kcal/mol. A simulação para esta proteína foi interessante, pois apresentou conformações na qual a região irregular que conecta as duas hélices foi muito próxima da apresentada na estrutura nativa. Assim, a Figura 8.1 apresenta uma visão geral do comportamento da proteína ao longo da simulação.

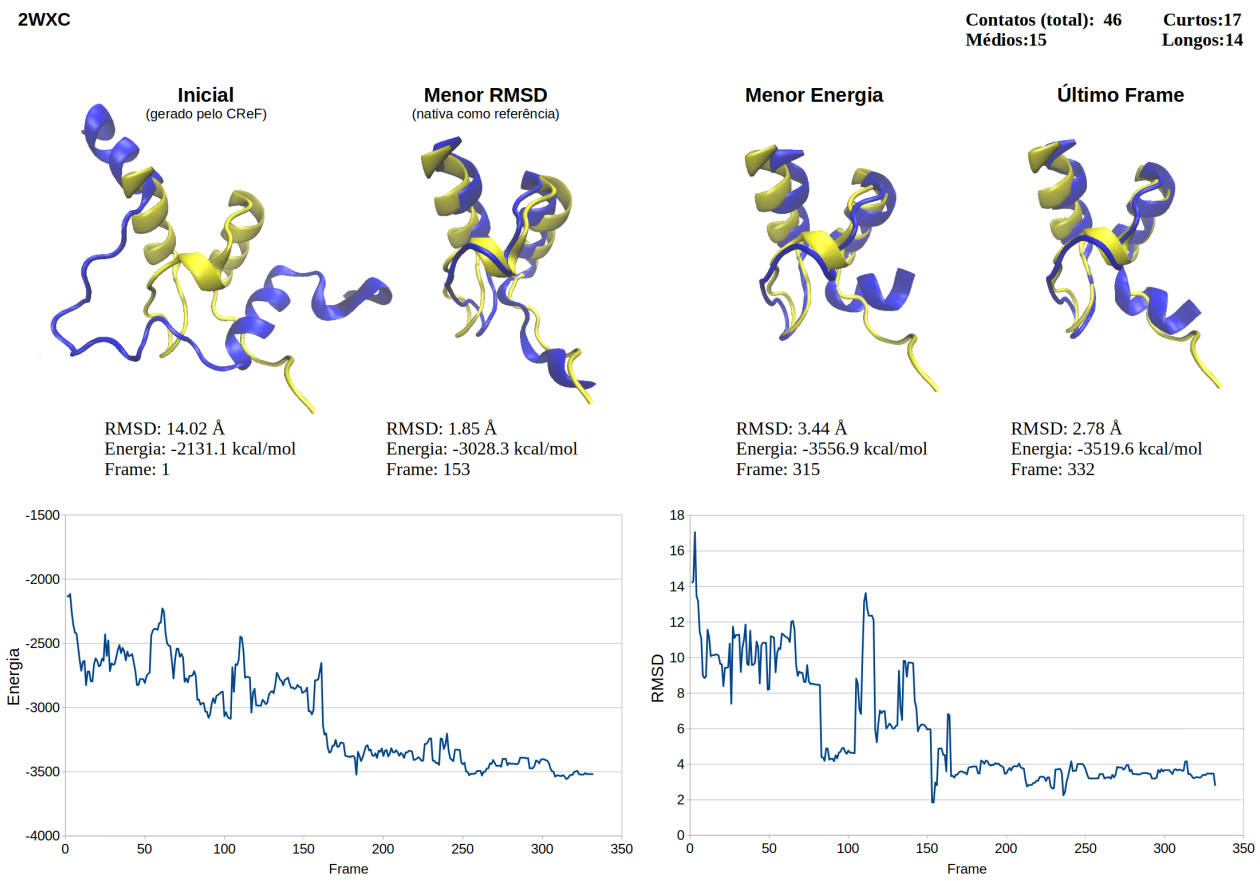


Figura 8.1 – Visão geral do comportamento da proteína de PDB ID: 2WXC ao longo da simulação. Fonte: Autora.

8.2 Proteína PDB ID: 1E0N

A proteína de PDB ID: 1E0N é composta por 27 resíduos de aminoácidos e apresenta estruturas secundárias regulares do tipo folha-β antiparalela. O número de pares em contato é 36, dos quais 22 são curtos e 14 médios, não apresentando contatos de longo alcance. A conformação inicial do CReF apresentou RMSD de 15,68 Å e energia de -950,7 kcal/mol. Ao longo da simulação, a conformação com menor RMSD (calculado a partir da estrutura nativa) foi constatada no frame 288 com 0,778 Å e energia -1.454,9 kcal/mol. A conformação com menor energia atribuída pela função envolvendo os contatos foi verificada no frame 254 com -1.505,6 kcal/mol e RMSD igual a 1,22 Å. A última conformação aceita pela simulação foi o frame 298 com RMSD igual a 0,98 Å e energia de -1.493,1 kcal/mol. A simulação mostrou que as fitas conseguiram enovelar a ponto de formar a folha-β. Assim, a Figura 8.2 apresenta uma visão geral do comportamento da proteína ao longo da simulação.

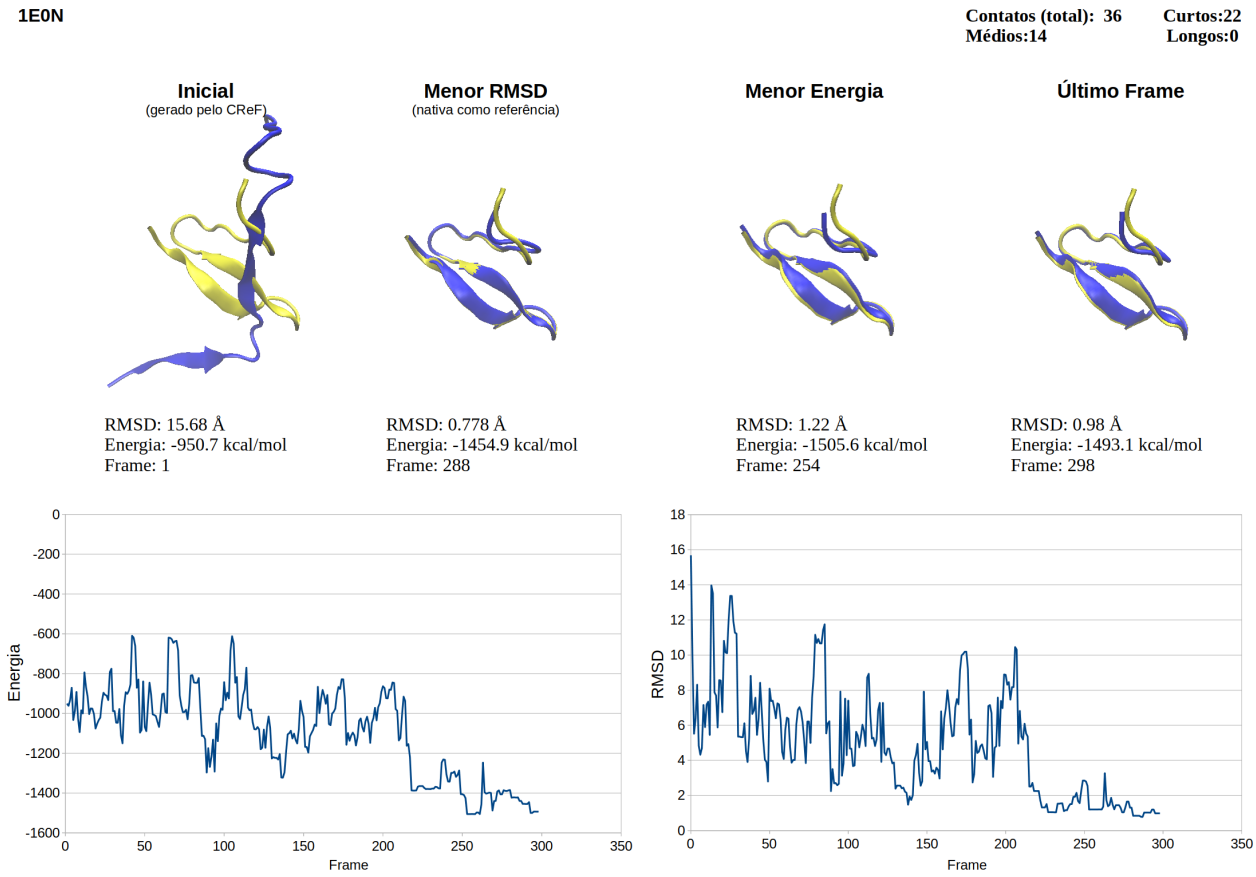


Figura 8.2 – Visão geral do comportamento da proteína de PDB ID: 1E0N ao longo da simulação. Fonte: Autora.

8.3 Proteína PDB ID: 1FME

A proteína de PDB ID: 1FME é composta por 28 resíduos de aminoácidos e apresenta estruturas secundárias regulares do tipo hélice-α e folha-β antiparalela. O número de pares em contato é 21, dos quais 14 são curtos e 7 médios, não apresentando contatos de longo alcance. A conformação inicial do CReF apresentou RMSD de 9,90 Å e energia de -1.017,9 kcal/mol. Ao longo da simulação, a conformação com menor RMSD (calculado a partir da estrutura nativa) foi constatada no frame 218 com 1,23 Å e energia -1.483,0 kcal/mol. A conformação com menor energia atribuída pela função envolvendo os contatos foi verificada no frame 398 com -1.622,2 kcal/mol e RMSD igual a 2,01 Å. A última conformação aceita pela simulação foi a de frame 486 com RMSD igual a 2,49 Å e energia de -1.547,6 kcal/mol. A Figura 8.3 apresenta uma visão geral do comportamento da proteína ao longo da simulação.

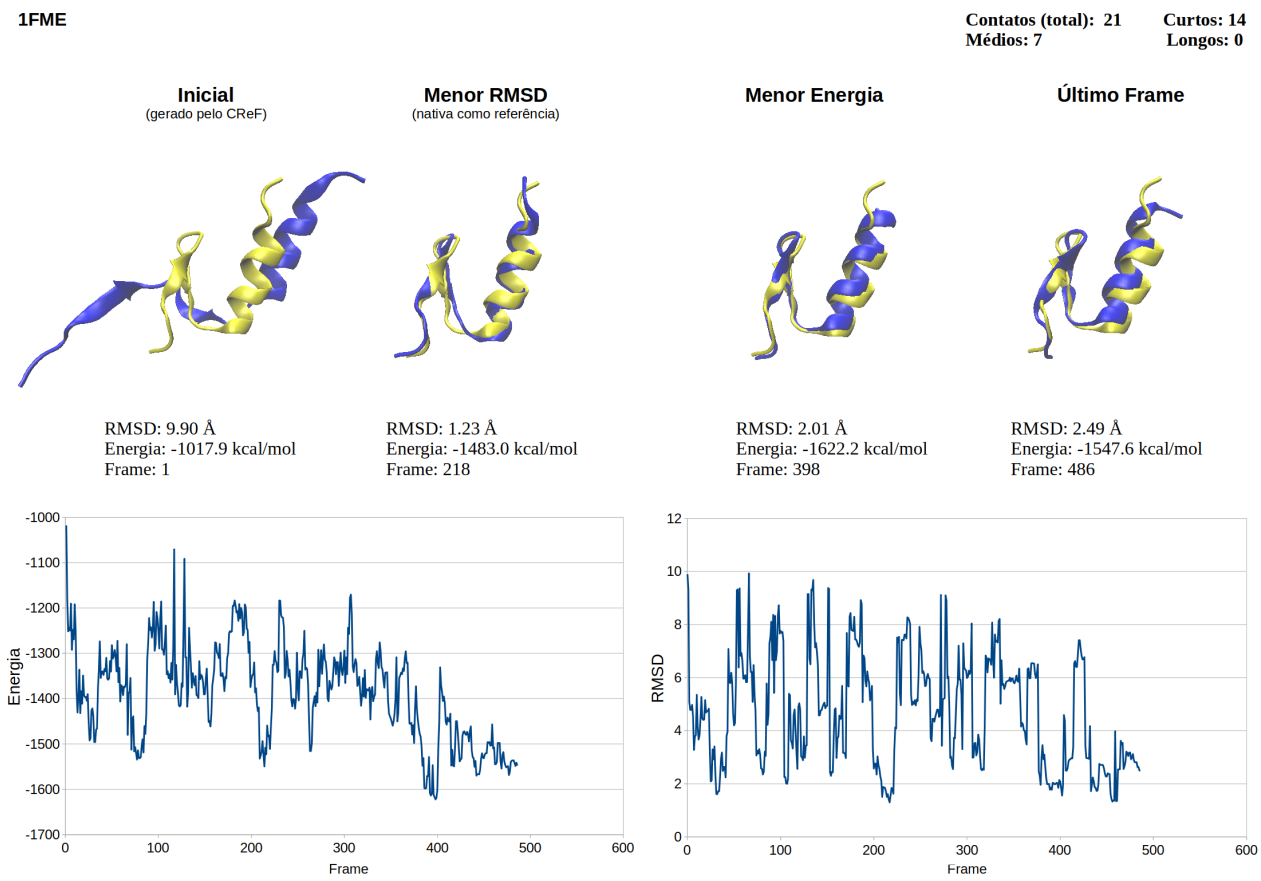


Figura 8.3 – Visão geral do comportamento da proteína de PDB ID: 1FME ao longo da simulação. Fonte: Autora.

8.4 Proteína PDB ID: 2HBA

A proteína de PDB ID: 2HBA é composta por 52 resíduos de aminoácidos e apresenta estrutura em alfa beta sanduíche. O número de pares em contato é 73, dos quais 21 são curtos, 22 médios e 30 longos. A conformação inicial do CReF apresentou RMSD de 15,41 Å e energia de -2.274,9 kcal/mol. Ao longo da simulação, a conformação com menor RMSD (calculado a partir da estrutura nativa) foi constatada no frame 267 com 1,80 Å e energia -3.937,6 kcal/mol. A conformação com menor energia atribuída pela função envolvendo os contatos foi verificada no frame 271 com -4.049,3 kcal/mol e RMSD igual a 3,69 Å. A última conformação aceita pela simulação foi a de frame 281 com RMSD igual a 2,76 Å e energia de -4.006,9 kcal/mol. A Figura 8.4 apresenta uma visão geral do comportamento da proteína ao longo da simulação.

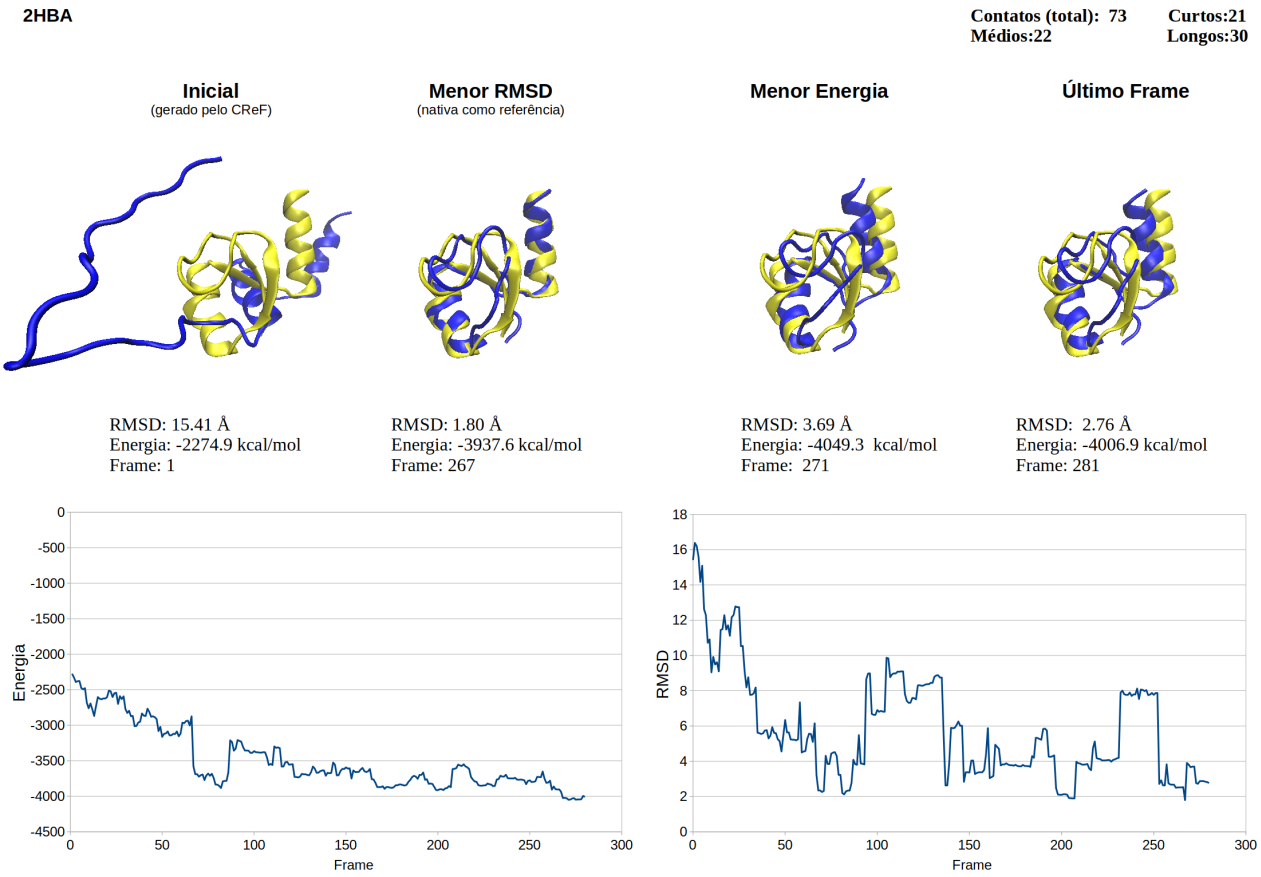


Figura 8.4 – Visão geral do comportamento da proteína de PDB ID: 2HBA ao longo da simulação. Fonte: Autora.

8.5 Proteína PDB ID: 1RES

A proteína de PDB ID: 1RES é composta por 43 resíduos de aminoácidos e apresenta estruturas secundárias regulares do tipo hélice-α. O número de pares em contato é 29, dos quais 16 são curtos, 10 médios e 3 longos. A conformação inicial do CReF apresentou RMSD de 13,85 Å e energia de -2.018,3 kcal/mol. Ao longo da simulação o frame 153 apresentou a conformação com menor valor de RMSD de 1,53 Å e energia -2.675,0 kcal/mol. A conformação com menor energia atribuída pela função envolvendo os contatos foi verificada no frame 374 com -3.282,0 kcal/mol. A última conformação aceita pela simulação foi a de frame 375 com RMSD igual a 3,82 Å e energia de -3.281,6 kcal/mol. A Figura 8.5 apresenta uma visão geral do comportamento da proteína ao longo da simulação.

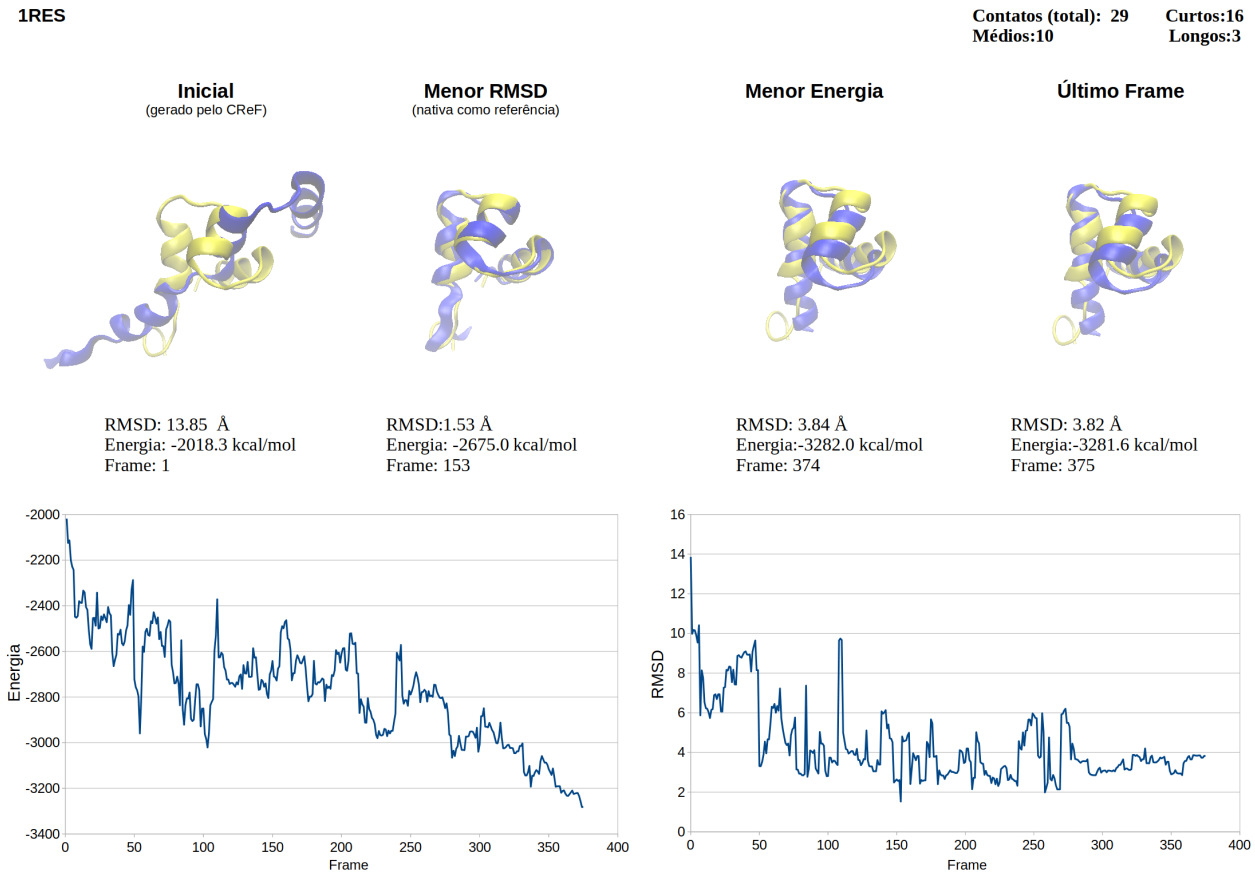


Figura 8.5 – Visão geral do comportamento da proteína de PDB ID: 1RES ao longo da simulação. Fonte: Autora.

8.6 Proteína PDB ID: 1YWJ

A proteína de código PDB ID: 1YWJ é composta por 41 resíduos de aminoácidos e apresenta estruturas secundárias regulares do tipo folhas-β antiparalela e paralela. O número de pares em contato é 40, dos quais 22 são curtos, 16 médios e 2 longos. A conformação inicial do CReF apresentou RMSD de 15,03 Å e energia de -1.329,3 kcal/mol. Ao longo da simulação, a conformação com menor RMSD, menor energia e a última aceita foram a mesma. Assim o frame 178 apresentou valores de 2,48 Å e -2.418,3 kcal/mol para RMSD e energia, respectivamente. A Figura 8.6 apresenta uma visão geral do comportamento da proteína ao longo da simulação.

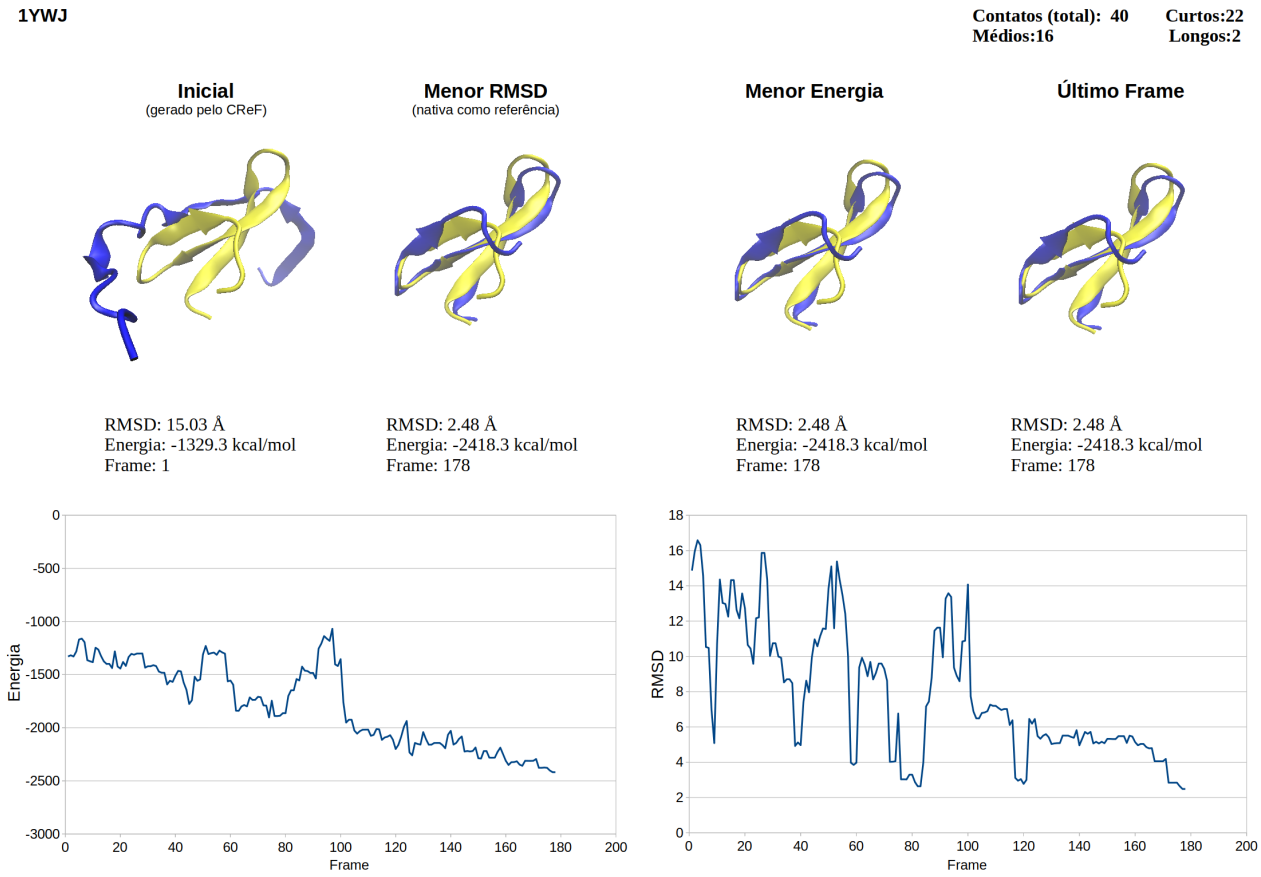


Figura 8.6 – Visão geral do comportamento da proteína de PDB ID: 1YWJ ao longo da simulação. Fonte: Autora.

9. PREDIÇÃO DA ESTRUTURA 3D DE PROTEÍNAS COM O MÉTODO CReF - VERSÃO FINAL

A visão geral do que foi realizado nesta dissertação é apresentado na Figura 9.1. A partir de um preditor de contatos, o arquivo de contatos (formato RR) para uma proteína é gerado. Com base na sequência de resíduos, esses são separados em pares de contatos (curto, médio e longo alcance) e armazenados no novo arquivo. A sequência é dividida em fragmentos de tamanho 5, os quais são alinhados através do BLASTp. Assim, as informações relacionadas aos templates selecionados pelo alinhamento são armazenadas em um novo arquivo junto à predição da estrutura secundária correspondente. A partir do arquivo de fragmentos, o cálculo dos ângulos ϕ, ψ e ω é realizado pelo PepDice3. Esse realiza simulações moleculares com o método monte carlo com *simulated annealing*. Aqui, a estrutura predita pela versão inicial do CReF foi utilizada como conformação inicial para as simulações. Para cada fragmento sorteado, uma nova conformação é atribuída, sua energia é calculada através da função que considera os contato e é comparada à conformação anterior. Isso ocorre até um critério de parada. Como a ideia é utilizar o método para proteínas com estrutura não conhecida, uma métrica de qualidade utilizando um valor de $RMSD_{predito}$ é usada para diferenciar as possíveis conformações.

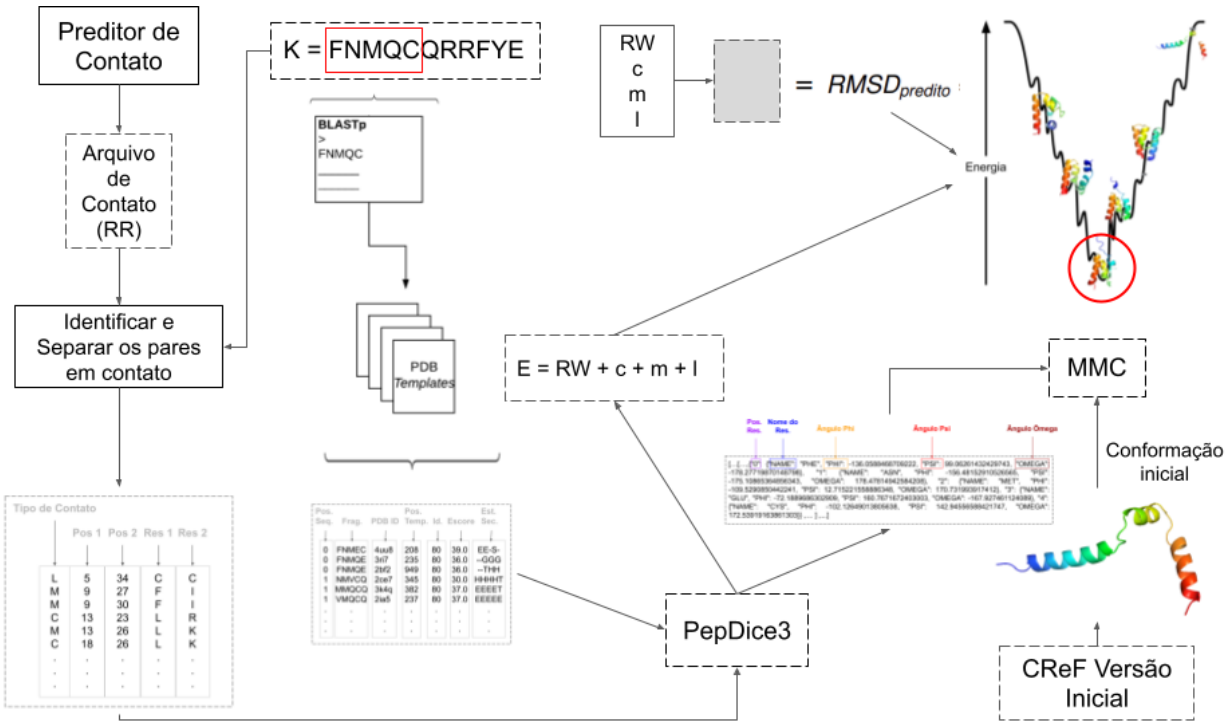


Figura 9.1 – Visão Geral do método proposto. Fonte: Autora.

10. CONCLUSÕES

Nesta dissertação foi proposto o uso de informações providas de contato entre resíduos de aminoácidos com o intuito de identificar se esses contribuiriam para o enovelamento das proteínas utilizando o método CReF (*Central Residue Fragment-based method*). Desse modo, este trabalho aborda o problema da predição de estrutura 3D de proteínas. Assim sendo, para que as informações de contato fossem incorporadas ao método, diversas alterações, desde o funcionamento básico até metodológicas, precisaram ser realizadas.

As principais modificações metodológicas surgiram, inicialmente, pelas execuções iniciais do método, o qual mesmo com todas as informações necessárias disponíveis para encontrar a estrutura nativa não a encontrava. A hipótese de que o problema estaria no algoritmo de clusterização foi refutada ao implementar outro algoritmo de clusterização e esse não apresentar resultados significativos.

Assim, acredita-se que, ao utilizar a clusterização dos ângulos centrais do fragmento, rotulando os valores pela estrutura secundária não parece ser o mais adequado no momento, isso se justifica já que os problemas de enovelamento que o método CReF enfrentou ao longo dos seus anos de existência foram, principalmente, nas regiões de estrutura secundária irregular. Desse modo, quando uma região da proteína tinha a sua estrutura secundária predita como região irregular, a média de um grupo em uma região no mapa de Ramachandran era selecionada para representar os ângulos ϕ e ψ. No entanto, é de conhecimento que os ângulos para as estruturas secundárias irregulares podem ocupar qualquer região do mapa e que impactam diretamente no enovelamento.

Além disso, os ângulos eram calculados pelo programa *Torsions*, o que fazia com que uma parte crucial para o método dependesse totalmente do modo como esse foi implementado. Adicionalmente, esse não calcula os ângulos pertencentes a cadeia lateral, assim, se em trabalhos futuros as informações da cadeia lateral fossem consideradas, outras modificações metodológicas precisariam ser realizadas.

Adicionalmente, guardar as informações apenas dos ângulos do resíduo central do fragmento (quando se conhece todo os resíduos desse) era um tanto adverso ao proposto neste trabalho, pois esse utilizaria informações entre os fragmentos. Assim, os valores dos ângulos para todos os resíduos dos fragmentos começaram a ser calculados, assim como a informação da posição no template que os fragmentos começam, para que possam ser acessados novamente.

Considerando os contatos e o modo como eles são preditos, há dois grandes grupos. O primeiro engloba métodos baseados em coevolução e o segundo métodos baseados em aprendizado de máquina. Contudo, diversos preditores utilizam informações de coevolução e aprendizado de máquina, assim não há uma grande divisão entre as abordagens. Entre os preditores existentes, 11 são listados e avaliados neste trabalho. Entre esses, ape-

nas 3 estão disponíveis atualmente online e entregam os arquivos de contato preditos, os outros dependem de grandes bancos de dados para serem executados.

Com essa análise e para que as informações de contato fossem efetivamente incorporadas ao método CReF, optou-se por utilizar o arquivo de contato no formato adotado pelo CASP como dado de entrada do CReF. Com isso, o método seleciona a probabilidade da predição do par em contato a ser considerada e os separa em curto, médio e longo alcance, dado a sequência de resíduos de aminoácidos. Essas informações são armazenas em um arquivo de saída de contatos no formato criado neste trabalho.

Com as informações incorporadas identificou-se a necessidade de um conjunto de conformações a serem amostradas, para que assim as informações de contato fossem utilizadas. Porém o modo como o CReF estava estruturado, entregando apenas uma conformação final, os contatos não poderiam ser utilizados. Adicionalmente o impacto e quantidade de contatos preditos corretamente necessários para o enovelamento não foram avaliados neste trabalho, pois para realizar essas análises outros aspectos do método precisaram ser realizados antes. Assim, um modelo de avaliação e de amostragem foram desenvolvidos.

O primeiro surgiu com o intuito de selecionar uma conformação dado um conjuntos de conformações a partir das informações de contato de curto, médio e longo alcance e um potencial atômico dependente de distância (RW). Para isso, um conjunto de dados robusto e com variabilidade conformacional (*decoys*) foi utilizado. Assim, apesar da simplicidade do modelo, esse evidencia que, ao retornar um valor de $RMSD_{predito}$, consegue distinguir entre *decoys* com valores de $RMSD_{real}$ maiores daqueles menores. Ademais, é de conhecimento que análises mais profundas e diferentes validações precisam ser realizados com o modelo.

O segundo emergiu da necessidade de um conjunto de conformações que comprovasse que as informações de contato auxiliariam o enovelamento. Assim, os termos de contato foram incorporados a uma função de energia usada para as simulações computacionais. Desse modo, as simulações foram realizadas pelo PepDice3 através de *simulated annealing*, um método de otimização que adapta o método de monte carlo - metropolis. O qual cada nova conformação é determinada pelo sorteio aleatório de um fragmento.

Os resultados e análises realizadas com o conjunto de proteínas submetido às simulações indica que as informações de contato foram suficientes para entregar conformações boas com valores baixos de energia e RMSD, mesmo explorando espaços conformacionais um tanto longe da conformação nativa, sugerindo uma boa distribuição amostral. No entanto, com isso também foi identificado que para choques incoerentes um termo de penalização poderia ser englobado a função de energia.

Além disso, observou-se que, para que os fragmentos obtidos através do CReF sejam de fato utilizados nas simulações, é preciso que algumas etapas sejam realizadas como a clusterização de todo o fragmento por estrutura secundária; identificar como os fragmentos preditos como estrutura irregular podem ser agrupados e mesmo assim representar aquela região; definir uma métrica que avalie se os fragmentos obtidos apresentam

a informação mínima necessária para que a partir da simulação esses sejam suficientes para encontrar uma conformação próxima ao estado nativo.

10.1 Principais Contribuições

Desse modo, as principais contribuições atribuídas a esta dissertação foram:

- Incorporação das informações de contato no método CReF;
- Elaboração de formatos de arquivos de saída;
- Obtenção das informações dos ângulos de todo o fragmento;
- Alteração no cálculo dos ângulos da cadeia principal pelo CReF;
- Modelo de avaliação de qualidade através de um valor de $RMSD_{predito}$ dado valores para RW, contatos de curto, médio e longo alcance para quando não tiver a estrutura nativa como parâmetro;
- Função de energia que utiliza as informações de RW e contatos para as simulações;
- Módulo de simulação molecular.

10.2 Limitações

As principais limitações encontradas neste trabalho foram:

- A falta de conformações totalmente estendidas ou com valores de RMSD acima de 12 Å no conjunto de treinamento;
- O número de proteínas representadas no conjunto de treinamento;
- A não penalização, quando ocorre choques, através de um termo na função de energia;
- A falta de garantia e confiabilidade dos fragmentos gerados diretamente pelo CReF e se esses contêm a informação necessária para o enovelamento, assim como um método de avaliação de qualidade e seleção desses.
- A não avaliação da qualidade da predição dos contatos, assim, o quanto é possível piorar a predição desses e ainda obter pares de contatos que são de fato importante para o enovelamento.

10.3 Perspectivas

O desenvolvimento desta dissertação apresentou diversas ideias e questões de pequisa, as quais podem ser consideradas em trabalhos futuros:

- A fim de garantir e avaliar a qualidade dos fragmentos gerados pelo CReF, desenvolver algoritmos de clusterização para todo o fragmento;
- Construir bibliotecas de fragmentos menores, por exemplo, com 3 resíduos de aminoácidos para usar em regiões de estrutura secundária irregular;
- Utilizar mapas de distribuição de probabilidade dos ângulos ϕ e ψ no mapa de Ramachandran para cada resíduo de aminoácido específico;
- Desenvolver métodos mais robustos para a predição do valor de RMSD com o intuito de utilizá-lo como métrica de avaliação de qualidade das conformações geradas pela simulação molecular quando não se conhece a estrutura nativa correspondente;
- Estudar os contatos preditos (não nativos) com cortes de probabilidade para aceitar o par de resíduos em contato e avaliar até que ponto as informações ainda conseguem enovelar a proteína e assim identificar os pares de resíduos cruciais para isso;
- Estudar a influência dos contatos de curto, médio e longo alcance separadamente para o enovelamento;
- Avaliar e acoplar novos termos à função de energia e meios de penalização.

REFERÊNCIAS BIBLIOGRÁFICAS

Adhikari, B. e Cheng, J. (2016). Protein residue contacts and prediction methods. In: *Data Mining Techniques for the Life Sciences*, vol. 1415, pp. 463–476. Springer, 2 ed..

Adhikari, B., Hou, J. e Cheng, J. (Mai, 2018). Dncon2: improved protein contact prediction using two-level deep convolutional neural networks. *Bioinformatics*, vol. 34, pp. 1466–1472.

Alberts, B., Johnson, A., Lewis, J., Morgan, D., Raff, M., Roberts, K., Walter, P., Wilson, J. e Hunt, T. (2017). *Biologia Molecular da Célula*. Artmed, Porto Alegre, BRA.

AlphaFold (2020). Alphafold: a solution to a 50-year-old grand challenge in biology. Recuperado de https://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology. Fev 2021.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. e Lipman, D. J. (Out, 1990). Basic local alignment search tool. *Journal of Molecular Biology*, vol. 215, pp. 403–410.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. e Lipman, D. J. (Set, 1997). Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Research*, vol. 25, pp. 3389–3402.

Amala, A. e Emerson, I. A. (Jun, 2019). Understanding contact patterns of protein structures from protein contact map and investigation of unique patterns in the globin-like folded domains. *Journal of Cellular Biochemistry*, vol. 120, pp. 9877–9886.

Anfinsen, C. B. (Jul, 1973). Principles that govern the folding of protein chains. *Science*, vol. 181, pp. 223–230.

Arnold, K., Bordoli, L., Kopp, J. e Schwede, T. (Jan, 2006). The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics*, vol. 22, pp. 195–201.

Baker, D. e Sali, A. (Out, 2001). Protein structure prediction and structural genomics. *Science*, vol. 294, pp. 93–96.

Baumeister, W. e Steven, A. C. (Dez, 2000). Macromolecular electron microscopy in the era of structural genomics. *Trends in Biochemical Sciences*, vol. 25, pp. 624–631.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. e Wheeler, D. L. (Jan, 2005). GenBank. *Nucleic Acids Research*, vol. 33, pp. D34–38.

Berger, B. e Leighton, T. (1998). Protein Folding in the Hydrophobic-hydrophilic (HP) is NP-complete. In: *Proceedings of the Second Annual International Conference on Computational Molecular Biology*, pp. 30–39, New York, USA. ACM.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. e Bourne, P. E. (Jan, 2000). The Protein Data Bank. *Nucleic Acids Research*, vol. 28, pp. 235–242.

Bhowmick, A. e Head-Gordon, T. (Jan, 2015). A monte carlo method for generating side chain structural ensembles. *Structure*, vol. 23, pp. 44–55.

Björkholm, P., Daniluk, P., Kryshtafovych, A., Fidelis, K., Andersson, R. e Hvidsten, T. R. (Mai, 2009). Using multi-data hidden markov models trained on local neighborhoods of protein structure to predict residue–residue contacts. *Bioinformatics*, vol. 25, pp. 1264–1270.

Blaszczyk, M., Kurcinski, M., Kouza, M., Wieteska, L., Debinski, A., Kolinski, A. e Kmiecik, S. (Jan, 2016). Modeling of protein–peptide interactions using the cabs-dock web server for binding site search and flexible docking. *Methods*, vol. 93, pp. 72–83.

Bowers, P. M., Strauss, C. E. e Baker, D. (Dez, 2000). De novo protein structure determination using sparse nmr data. *Journal of Biomolecular NMR*, vol. 18, pp. 311–318.

Bowie, J. U., Luthy, R. e Eisenberg, D. (Jul, 1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, vol. 253, pp. 164–170.

Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. e Karplus, M. (Jan, 1983). CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry*, vol. 4, pp. 187–217.

Chen, P. e Li, J. (Mai, 2010). Prediction of protein long-range contacts using an ensemble of genetic algorithm classifiers with sequence profile centers. *BMC Structural Biology*, vol. 10, pp. 1–13.

Cheng, J. e Baldi, P. (Abr, 2007). Improved residue contact prediction using support vector machines and a large feature set. *BMC Bioinformatics*, vol. 8, pp. 1–9.

Chothia, C. e Lesk, A. M. (Abr, 1986). The relation between the divergence of sequence and structure in proteins. *The EMBO Journal*, vol. 5, pp. 823–826.

Christen, M., Hünenberger, P. H., Bakowies, D., Baron, R., Bürgi, R., Geerke, D. P., Heinz, T. N., Kastenholz, M. A., Kräutler, V., Oostenbrink, C., Peter, C., Trzesniak, D. e Gunsteren, W. F. v. (Dez, 2005). The GROMOS software for biomolecular simulation: GROMOS05. *Journal of Computational Chemistry*, vol. 26, pp. 1719–1751.

Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W. e Kollman, P. A. (Mai, 1995). A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *Journal of the American Chemical Society*, vol. 117, pp. 5179–5197.

Crescenzi, P., Goldman, D., Papadimitriou, C., Piccolboni, A. e Yannakakis, M. (Fev, 1998). On the complexity of protein folding. *Journal of Computational Biology*, vol. 5, pp. 423–465.

Crick, F. (Ago, 1970). Central dogma of molecular biology. *Nature*, vol. 227, pp. 561–563.

Czaplewski, C., Karczyńska, A., Sieradzan, A. K. e Liwo, A. (Jul, 2018). Unres server for physics-based coarse-grained simulations and prediction of protein structure, dynamics and thermodynamics. *Nucleic Acids Research*, vol. 46, pp. W304–W309.

da Motta Dall'Agno, K. C. (2012). Um estudo sobre a predição da estrutura 3D aproximada de proteínas utilizando o método cref com refinamento. Dissertação de Mestrado, Faculdade de Informática – PUCRS, Porto Alegre, BRA.

Dall'Agno, K. C. e Norberto de Souza, O. (Jun, 2013). An expert protein loop refinement protocol by molecular dynamics simulations with restraints. *Expert Systems with Applications*, vol. 40, pp. 2568–2574.

Dayhoff, M. O., Schwartz, R. M. e Orcutt, B. (1978). Chapter 22: A model of evolutionary change in proteins. In: *Proceedings of the Atlas of Protein Sequence and Structure*, pp. 345–352, Washington, USA. NBRF.

de Oliveira, S. H. P., Shi, J. e Deane, C. M. (Fev, 2017). Comparing co-evolution methods and their application to template-free protein structure prediction. *Bioinformatics*, vol. 33, pp. 373–381.

Deng, H., Jia, Y. e Zhang, Y. (Fev, 2016). 3drobot: automated generation of diverse and well-packed protein structure decoys. *Bioinformatics*, vol. 32, pp. 378–387.

Di Lena, P., Nagata, K. e Baldi, P. (Out, 2012). Deep architectures for protein contact map prediction. *Bioinformatics*, vol. 28, pp. 2449–2457.

Dorn, M. (2008). Uma proposta para a predição computacional da estrutura 3D aproximada de polipeptídeos com redução do espaço conformacional utilizando análise de intervalos. Dissertação de Mestrado, Faculdade de Informática – PUCRS, Porto Alegre, BRA.

Dorn, M. e de Souza, O. N. (2008). CReF: A central-residue-fragment-based method for predicting approximate 3-d polypeptides structures. In: *Proceedings of the ACM Symposium on Applied Computing*, pp. 1261–1267, Fortaleza, BRA. ACM.

Efimov, A. (Jan, 1993). Standard structures in proteins. *Progress in Biophysics and Molecular Biology*, vol. 60, pp. 201–239.

Feng, S.-H., Xu, J.-Y. e Shen, H.-B. (2020). Artificial intelligence in bioinformatics: Automated methodology development for protein residue contact map prediction. In: *Biomedical Information Technology*, vol. 1, pp. 217–237. Elsevier, 2 ed..

Fiser, A., Do, R. K. G. e Šali, A. (Set, 2000). Modeling of loops in protein structures. *Protein Science*, vol. 9, pp. 1753–1773.

Fiser, A., Feig, M., Brooks, C. L. e Sali, A. (Jun, 2002). Evolution and physics in comparative protein structure modeling. *Accounts of Chemical Research*, vol. 35, pp. 413–421.

Floudas, C., Fung, H., McAllister, S., Mönnigmann, M. e Rajgaria, R. (Fev, 2006). Advances in protein structure prediction and de novo protein design: A review. *Chemical Engineering Science*, vol. 61, pp. 966–988.

Frenkel, D. e Smit, B. (2001). *Understanding molecular simulation: from algorithms to applications*. Elsevier, San Diego, USA.

Gu, J. e Bourne, P. E. (2009). *Structural Bioinformatics*. Wiley-Blackwell, Hoboken, USA.

Guex, N. e Peitsch, M. C. (2006). Principles of protein structure, comparative protein modelling and visualisation. Recuperado de https://swissmodel.expasy.org/course/text/chapter3.htm. Out 2020.

Hamelryck, T. e Manderick, B. (Nov, 2003). Pdb file parser and structure class implemented in python. *Bioinformatics*, vol. 19, pp. 2308–2310.

Hanson, J., Paliwal, K., Litfin, T., Yang, Y. e Zhou, Y. (Jun, 2018). Accurate prediction of protein contact maps by coupling residual two-dimensional bidirectional long short-term memory with convolutional neural networks. *Bioinformatics*, vol. 34, pp. 4039–4045.

Haykin, S. S. (1999). *Neural Networks: A Comprehensive Foundation*. Prentice Hall, Singapore, MYS.

Heilmann, N., Wolf, M., Kozlowska, M., Sedghamiz, E., Setzler, J., Brieg, M. e Wenzel, W. (Out, 2020). Sampling of the conformational landscape of small proteins with monte carlo methods. *Scientific Reports*, vol. 10, pp. 1–13.

Höltje, H.-D., Folkers, G., Mannhold, R., Kubinyi, H. e Timmerman, H. (2008). *Molecular Modeling: Basic Principles and Applications*. Wiley-VCH Verlag GmbH, Weinheim, DEU.

Jana, N. D., Das, S. e Sil, J. (2018). *A Metaheuristic Approach to Protein Structure Prediction: Algorithms and Insights from Fitness Landscape Analysis*. Springer, Cham, CHE.

Johnson, L. S., Eddy, S. R. e Portugaly, E. (Ago, 2010). Hidden markov model speed heuristic and iterative hmm search procedure. *BMC Bioinformatics*, vol. 11, pp. 1–8.

Jones, D. T., Buchan, D. W., Cozzetto, D. e Pontil, M. (Jan, 2012). Psicov: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics*, vol. 28, pp. 184–190.

Jones, D. T. e Kandathil, S. M. (Out, 2018). High precision in protein contact prediction using fully convolutional neural networks and minimal sequence features. *Bioinformatics*, vol. 34, pp. 3308–3315.

Jones, D. T., Singh, T., Kosciolek, T. e Tetchner, S. (Abr, 2015). Metapsicov: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. *Bioinformatics*, vol. 31, pp. 999–1006.

Jones, D. T., Taylort, W. e Thornton, J. M. (Jul, 1992). A new approach to protein fold recognition. *Nature*, vol. 358, pp. 86–89.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Tunyasuvunakool, K., Ronneberger, O., Bates, R., Zidek, A., Bridgland, A. et al. (Dez, 2020). High accuracy protein structure prediction using deep learning. *Fourteenth Critical Assessment of Techniques for Protein Structure Prediction*, vol. 22, pp. 24.

Kabsch, W. e Sander, C. (Dez, 1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers: Original Research on Biomolecules*, vol. 22, pp. 2577–2637.

Kaján, L., Hopf, T. A., Kalaš, M., Marks, D. S. e Rost, B. (Mar, 2014). Freecontact: fast and free software for protein contact prediction from residue co-evolution. *BMC Bioinformatics*, vol. 15, pp. 1–6.

Källberg, M., Margaryan, G., Wang, S., Ma, J. e Xu, J. (2014). Raptorx server: a resource for template-based protein structure modeling. In: *Protein Structure Prediction*, vol. 1137, pp. 17–27. Springer, 3 ed..

Kamisetty, H., Ovchinnikov, S. e Baker, D. (Set, 2013). Assessing the utility of coevolution-based residue–residue contact predictions in a sequence-and structure-rich era. *Proceedings of the National Academy of Sciences*, vol. 110, pp. 15674–15679.

Kandathil, S. M., Greener, J. G. e Jones, D. T. (Jul, 2019). Prediction of interresidue contacts with deepmetapsicov in casp13. *Proteins: Structure, Function, and Bioinformatics*, vol. 87, pp. 1092–1099.

Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. e Sternberg, M. J. (Mai, 2015). The phyre2 web portal for protein modeling, prediction and analysis. *Nature Protocols*, vol. 10, pp. 845–858.

Kendrew, J. C., Bodo, G., Dintzis, H. M., Parrish, R., Wyckoff, H. e Phillips, D. C. (Mar, 1958). A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*, vol. 181, pp. 662–666.

Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K. e Moult, J. (Out, 2019). Critical assessment of methods of protein structure prediction CASP—round XIII. *Proteins: Structure, Function, and Bioinformatics*, vol. 87, pp. 1011–1020.

Laskowski, R. A., Jabłońska, J., Pravda, L., Vařeková, R. S. e Thornton, J. M. (Jan, 2018). Pdbsum: Structural summaries of pdb entries. *Protein Science*, vol. 27, pp. 129–134.

Laskowski, R. A., MacArthur, M. W., Moss, D. S. e Thornton, J. M. (Abr, 1993). Procheck: a program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography*, vol. 26, pp. 283–291.

Lesk, A. M. (2001). *Introduction to Protein Architecture: The Structural Biology of Proteins*. Oxford University Press, Oxford, GBR.

Lesk, A. M. (2008). *Introdução à Bioinformática*. Artmed, Porto Alegre, BRA.

Levinthal, C. (Jan, 1968). Are there pathways for protein folding? *Journal de Chimie Physique*, vol. 65, pp. 44–45.

Li, Y., Hu, J., Zhang, C., Yu, D.-J. e Zhang, Y. (Nov, 2019). Respre: high-accuracy protein contact prediction by coupling precision matrix with deep residual neural networks. *Bioinformatics*, vol. 35, pp. 4647–4655.

Liu, Y., Palmedo, P., Ye, Q., Berger, B. e Peng, J. (Jan, 2018). Enhancing evolutionary couplings with deep convolutional neural networks. *Cell Systems*, vol. 6, pp. 65–74.

Lodish, H., Berk, A., Zipursky, S. L., Matsudaira, P., Baltimore, D. e Darnell, J. (2008). *Molecular Cell Biology*. W. H. Freeman, New York, USA.

Lorenzen, S. e Zhang, Y. (Dez, 2007). Monte carlo refinement of rigid-body protein docking structures with backbone displacement and side-chain optimization. *Protein Science*, vol. 16, pp. 2716–2725.

Luscombe, N. M., Greenbaum, D. e Gerstein, M. (Fev, 2001). What is bioinformatics? a proposed definition and overview of the field. *Methods of Information in Medicine*, vol. 40, pp. 346–358.

Ma, J. (2015). *Protein Structure Prediction by Protein Alignments*. Tese de Doutorado, Toyota Technological Institute at Chicago, Chicago, IL, EUA.

Machado, V. S. (2016). wCReF: uma interface web para o método CReF de predição da estrutura 3D aproximada de proteínas. Dissertação de Mestrado, Escola Politécnica - PUCRS, Porto Alegre, BRA.

Maggio, E. T. e Ramnarayan, K. (Jul, 2001). Recent developments in computational proteomics. *Trends in Biotechnology*, vol. 19, pp. 266–272.

Magnan, C. N. e Baldi, P. (Set, 2014). Sspro/accpro 5: almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity. *Bioinformatics*, vol. 30, pp. 2592–2597.

Martí-Renom, M. A., Stuart, A. C., Fiser, A., Sánchez, R., Melo, F. e Šali, A. (Jun, 2000). Comparative protein structure modeling of genes and genomes. *Annual Review of Biophysics and Biomolecular Structure*, vol. 29, pp. 291–325.

Marzzoco, A. e Torres Baptista, B. (2015). *Bioquímica Básica*. Guanabara Koogan, Rio de Janeiro, BRA.

McCulloch, W. S. e Pitts, W. (Dez, 1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, vol. 5, pp. 115–133.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. e Teller, E. (Mar, 1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, vol. 21, pp. 1087–1092.

Miller, C. S. e Eisenberg, D. (Jul, 2008). Using inferred residue contacts to distinguish between correct and incorrect protein models. *Bioinformatics*, vol. 24, pp. 1575–1582.

Mirny, L. e Domany, E. (Dez, 1996). Protein fold recognition and dynamics in the space of contact maps. *Proteins: Structure, Function, and Bioinformatics*, vol. 26, pp. 391–410.

Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T. e Tramontano, A. (Mar, 2018). Critical assessment of methods of protein structure prediction (casp)—round xii. *Proteins: Structure, Function, and Bioinformatics*, vol. 86, pp. 7–15.

Nagaraj, K., Sharvani, G. e Sridhar, A. (Jan, 2018). Emerging trend of big data analytics in bioinformatics: a literature review. *International Journal of Bioinformatics Research and Applications*, vol. 14, pp. 144–205.

Nelson, D. e Cox, M. (2014). *Princípios de Bioquímica de Lehninger*. Artmed, Porto Alegre, BRA.

Norvig, P. e Russell, S. (2014). *Inteligência artificial*. Elsevier Brasil, Rio de Janeiro, BRA.

Okamoto, Y. (Nov, 2019). Protein structure predictions by enhanced conformational sampling methods. *Biophysics and Physicobiology*, vol. 16, pp. 344–366.

Osguthorpe, D. J. (Abr, 2000). Ab initio protein folding. *Current Opinion in Structural Biology*, vol. 10, pp. 146–152.

Ovchinnikov, S., Kamisetty, H. e Baker, D. (Mai, 2014). Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information. *Elife*, vol. 3, pp. e02030.

Pauling, L. e Corey, R. B. (Mai, 1951). Atomic coordinates and structure factors for two helical configurations of polypeptide chains. *Proceedings of the National Academy of Sciences*, vol. 37, pp. 235.

Perutz, M. F., Muirhead, H., Cox, J. M. e Goaman, L. C. (Jul, 1968). Three-dimensional fourier synthesis of horse oxyhaemoglobin at 2.8 å resolution: the atomic model. *Nature*, vol. 219, pp. 131–139.

Perutz, M. F., Rossmann, M. G., Cullis, A. F., Muirhead, H., Will, G. e North, A. C. (Fev, 1960). Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5-A. resolution, obtained by X-ray analysis. *Nature*, vol. 185, pp. 416–422.

Pevsner, J. (2015). *Bioinformatics and functional genomics*. Wiley-Blackwell, Oxford, GBR.

Ramachandran, G. T. e Sasisekharan, V. (Jan, 1968). Conformation of polypeptides and proteins. *Advances in Protein Chemistry*, vol. 23, pp. 283–437.

Remmert, M., Biegert, A., Hauser, A. e Söding, J. (Dez, 2012). Hhblits: lightning-fast iterative protein sequence searching by hmm-hmm alignment. *Nature Methods*, vol. 9, pp. 173–175.

Rino, J. P. e Costa, B. V. d. (2013). *ABC da Simulação Computacional*. Editora Livraria da Física, São Paulo, BRA.

Rohl, C. A., Strauss, C. E., Misura, K. M. e Baker, D. (2004). Protein structure prediction using rosetta. In: *Numerical Computer Methods, Part D*, vol. 383, pp. 66–93. Academic Press, 1 ed..

Roy, A., Kucukural, A. e Zhang, Y. (Mar, 2010). I-tasser: a unified platform for automated protein structure and function prediction. *Nature Protocols*, vol. 5, pp. 725–738.

Rupp, B. (Oct, 2009). *Biomolecular Crystallography: Principles, Practice, and Application to Structural Biology*. Garland Science, New York, USA.

Sali, A. e Blundell, T. L. (Dez, 1993). Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology*, vol. 234, pp. 779–815.

Sanger, F. (Jun, 1949). The terminal peptides of insulin. *Biochemical Journal*, vol. 45, pp. 563.

Satoh, A. (2010). *Introduction to practice of molecular simulation: molecular dynamics, Monte Carlo, Brownian dynamics, Lattice Boltzmann and dissipative particle dynamics*. Elsevier, Burlington, USA.

Schneider, M. e Brock, O. (Out, 2014). Combining physicochemical and evolutionary information for protein contact prediction. *PloS One*, vol. 9, pp. e108438.

Schrödinger, LLC (2015). The PyMOL molecular graphics system, version~1.8. Recuperado de https://pymol.org/2/. Fev 2020.

Seemayer, S., Gruber, M. e Söding, J. (Nov, 2014). Ccmpred—fast and precise prediction of protein residue–residue contacts from correlated mutations. *Bioinformatics*, vol. 30, pp. 3128–3130.

Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A. W., Bridgland, A. et al. (Dez, 2019). Protein structure prediction using multiple deep neural networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13). *Proteins: Structure, Function, and Bioinformatics*, vol. 87, pp. 1141–1148.

Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A. W., Bridgland, A. et al. (Jan, 2020). Improved protein structure prediction using potentials from deep learning. *Nature*, vol. 577, pp. 706–710.

Shindyalov, I., Kolchanov, N. e Sander, C. (Mar, 1994). Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? *Protein Engineering, Design and Selection*, vol. 7, pp. 349–358.

Simons, K. T., Kooperberg, C., Huang, E. e Baker, D. (Abr, 1997). Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *Journal of Molecular Biology*, vol. 268, pp. 209–225.

Siow, L. (2018). Welcome into the fold. Recuperado de https://medium.com/proteinqure/welcome-into-the-fold-bbd3f3b19fdd. Jun 2018.

Skwark, M. J., Raimondi, D., Michel, M. e Elofsson, A. (Nov, 2014). Improved contact predictions using the recognition of protein like contact patterns. *PLoS Computational Biology*, vol. 10, pp. e1003889.

Söding, J. (Abr, 2005). Protein homology detection by hmm–hmm comparison. *Bioinformatics*, vol. 21, pp. 951–960.

Soletti, L. V. (2015). Um modelo de workflow científico para o refinamento da estrutura 3D aproximada de proteínas. Dissertação de Mestrado, Faculdade de Informática – PUCRS, Porto Alegre, BRA.

Sousa, S. F., Fernandes, P. A. e Ramos, M. J. (Jul, 2006). Protein–ligand docking: Current status and future challenges. *Proteins: Structure, Function, and Bioinformatics*, vol. 65, pp. 15–26.

Srinivasan, R. e Rose, G. D. (Jun, 1995). Linus: a hierarchic procedure to predict the fold of a protein. *Proteins: Structure, Function, and Bioinformatics*, vol. 22, pp. 81–99.

Svedberg, T. e Fåhraeus, R. (Fev, 1926). A new method for the determination of the molecular weight of the proteins. *Journal of the American Chemical Society*, vol. 48, pp. 430–438.

Tien, M. Z., Sydykova, D. K., Meyer, A. G. e Wilke, C. O. (Mai, 2013). Peptidebuilder: A simple python library to generate model peptides. *PeerJ*, vol. 1, pp. 80.

Timberlake, K. C. (2015). *General, Organic, and Biological Chemistry: Structures of Life*. Prentice Hall, New York, USA.

Touw, W. G., Baakman, C., Black, J., Te Beek, T. A., Krieger, E., Joosten, R. P. e Vriend, G. (Jan, 2015). A series of pdb-related databanks for everyday needs. *Nucleic Acids Research*, vol. 43, pp. D364–D368.

Tramontano, A. (Mar, 1998). Homology modeling with low sequence identity. *Methods*, vol. 14, pp. 293–300.

Van Laarhoven, P. J. e Aarts, E. H. (1987). *Simulated annealing: Theory and applications*. Springer, Dordrecht, NLD.

Vendruscolo, M. e Domany, E. (2000). Protein folding using contact maps. In: *Hormones and Stem Cells*, vol. 58, pp. 171–212. Academic Press, 1 ed..

Venkatesan, A., Gopal, J., Candavelou, M., Gollapalli, S. e Karthikeyan, K. (Jun, 2013). Computational approach for protein structure prediction. *Healthcare Informatics Research*, vol. 19, pp. 137.

Verli, H. (2014). *Bioinformática: da biologia à flexibilidade molecular*. Sociedade Brasileira de Bioquímica e Biologia Molecular, Porto Alegre, BRA.

Voet, D. e Voet, J. G. (2013). *Bioquímica*. Artmed, Porto Alegre, BRA.

Wang, S., Sun, S. e Xu, J. (Mar, 2018). Analysis of deep learning methods for blind protein contact prediction in casp12. *Proteins: Structure, Function, and Bioinformatics*, vol. 86, pp. 67–77.

Wang, Z., Eickholt, J. e Cheng, J. (Jun, 2011). Apollo: a quality assessment service for single and multiple protein models. *Bioinformatics*, vol. 27, pp. 1715–1716.

Watson, J. D., Baker, T. A., Bell, S. P., Gann, A., Levine, M. e Losicke, R. (2015). *Biologia molecular do gene*. Artmed, Porto Alegre, BRA.

Webb, B. e Sali, A. (Jun, 2016). Comparative Protein Structure Modeling Using MODELLER. *Current Protocols in Bioinformatics*, vol. 54, pp. 5.6.1–5.6.37.

Wüthrich, K. (Jul, 1986). NMR with Proteins and Nucleic Acids. *Europhysics News*, vol. 17, pp. 11–13.

Xiang, Z. (Jun, 2006). Advances in Homology Protein Structure Modeling. *Current Protein and Peptide Science*, vol. 7, pp. 217–227.

Xiong, J. (2006). *Essential Bioinformatics*. Cambridge University Press, New York, USA.

Xu, D. e Zhang, Y. (Jul, 2012). Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins: Structure, Function, and Bioinformatics*, vol. 80, pp. 1715–1735.

Yang, Y., Faraggi, E., Zhao, H. e Zhou, Y. (Ago, 2011). Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates. *Bioinformatics*, vol. 27, pp. 2076–2082.

Zhang, J., Lin, M., Chen, R., Liang, J. e Liu, J. S. (Jan, 2007). Monte carlo sampling of near-native structures of proteins with applications. *Proteins: Structure, Function, and Bioinformatics*, vol. 66, pp. 61–68.

Zhang, J. e Zhang, Y. (Out, 2010). A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction. *PloS One*, vol. 5, pp. e15386.

Zhang, Y. (Jun, 2008). Progress and challenges in protein structure prediction. *Current Opinion in Structural Biology*, vol. 18, pp. 342–348.

APÊNDICE A – CONJUNTO INICIAL DE PROTEÍNAS

Tabela A.1 – Proteínas selecionadas como conjunto inicial para identificar o estado atual do método CReF antes de qualquer alteração. Essas foram definidas através dos estudos realizados por da Motta Dall'Agno (2012). Fonte: Autora.

Fonte	PDB ID	Tamanho da Proteína
Karina, 2012	1K43	14
	1ZDD	34
	2ERL	40
	1YWJ	41
	1GPT	47
	1GAB	53
	1GB1	56
	1ROP	63
	1CSP	67
	1UTG	70
	1C5A	73
	1CTF	74
	1OPD	85
	2EZK	99
	1KSR	100
	1ERV	105

APÊNDICE B – CONJUNTO INICIAL DE PROTEÍNAS - CASP

Tabela B.1 – Proteínas selecionadas como conjunto inicial para identificar o estado atual do método CReF antes de qualquer alteração. Essas foram definidas através dos CASP10, CASP11, CASP12 e CASP13 com tamanho de até 105 resíduos de aminoácidos. Fonte: Autora.

Fonte	PDB ID	Tamanho da Proteína
CASP10	2M7T	33
	6MM4	33
	4HFX	86
	5SYQ	86
	4F98	88
CASP11	4OJK	44
	5A1Q	68
	2N2U	77
CASP12	5G3Q	89
	5JMB	91
CASP13	5W9F	41
	6GNX	52
	6F45	72
	6QEK	79
	6MSP	80
	6BTC	96
	6D7Y	96
	6G57	97

APÊNDICE C – CONJUNTO DE PROTEÍNAS COM VARIABILIDADE CONFORMACIONAL

Tabela C.1 – Conjunto de proteínas com variabilidade conformacional (*decoys*) selecionado a partir do programa 3DRobot. Fonte: Autora.

ID PDB	Tamanho Proteína	Tipo	Número de Decoys	Fonte
1A19	89	Alfa e Beta	100	ROSETTA
1A32	65	Alfa	100	ROSETTA
1A68	87	Alfa e Beta	100	ROSETTA
1ABV	103	Alfa	400	I-TASSER
1ACF	125	Alfa e Beta	100	ROSETTA
1AF7	72	Alfa	400	I-TASSER
1Ah9	63	Beta	400	I-TASSER
1AIL	70	Alfa	100	ROSETTA
1AIU	105	Alfa e Beta	100	ROSETTA
1AOY	65	Alfa e Beta	400	I-TASSER
1B3A	55	Alfa e Beta	100	ROSETTA
1B4B	71	Alfa e Beta	400	I-TASSER
1B72	49	Alfa	400	I-TASSER
1BGF	118	Alfa	100	ROSETTA
1BK2	57	Beta	100	ROSETTA
1BKR	108	Alfa	100	ROSETTA
1BM8	99	Alfa e Beta	400	I-TASSER
1BM8	99	Alfa e Beta	100	ROSETTA
1BQ9	51	Alfa e Beta	100	ROSETTA
1C8C	62	Alfa e Beta	100	ROSETTA
1C9O	66	Beta	100	ROSETTA
1CC8	72	Alfa e Beta	100	ROSETTA
1CEI	85	Alfa	100	ROSETTA
1CEW	108	Alfa e Beta	400	I-TASSER
1CG5	141	Alfa	100	ROSETTA
1CQK	101	Alfa e Beta	400	I-TASSER
1CSP	67	Beta	400	I-TASSER
1CTF	68	Alfa e Beta	100	ROSETTA
1DCJ	73	Alfa e Beta	400	I-TASSER

Continua na próxima página

Tabela C.1 – Continuação da página anterior

ID PDB	Tamanho Proteína	Tipo	Número de Decoys	Fonte
1DHN	121	Alfa e Beta	100	ROSETTA
1DTJ	74	Alfa e Beta	400	I-TASSER
1E6I	110	Alfa	100	ROSETTA
1EGX	115	Alfa e Beta	400	I-TASSER
1ELW	117	Alfa	100	ROSETTA
1ENH	54	Alfa	100	ROSETTA
1EW4	106	Alfa e Beta	100	ROSETTA
1EYV	131	Alfa	100	ROSETTA
1FAD	92	Alfa	400	I-TASSER
1FKB	107	Alfa e Beta	100	ROSETTA
1FNA	91	Beta	100	ROSETTA
1FO5	85	Alfa e Beta	400	I-TASSER
1G1C	98	Beta	400	I-TASSER
1GJX	77	Beta	400	I-TASSER
1GPT	47	Alfa e Beta	400	I-TASSER
1GVP	87	Beta	100	ROSETTA
1GYV	117	Beta	400	I-TASSER
1HZ6	61	Alfa e Beta	100	ROSETTA
1IG5	75	Alfa	100	ROSETTA
1IIB	103	Alfa e Beta	100	ROSETTA
1ITP	68	Alfa e Beta	400	I-TASSER
1JNU	104	Alfa e Beta	400	I-TASSER
1KJS	74	Alfa	400	I-TASSER
1KPE	108	Alfa e Beta	100	ROSETTA
1KVI	68	Alfa e Beta	400	I-TASSER
1LIS	125	Alfa	100	ROSETTA
1LOU	92	Alfa e Beta	100	ROSETTA
1MKY	81	Alfa e Beta	400	I-TASSER
1MLA	70	Alfa e Beta	400	I-TASSER
1N0U	69	Alfa e Beta	400	I-TASSER
1NE3	56	Beta	400	I-TASSER
1NPS	88	Beta	400	I-TASSER
1NPS	88	Beta	100	ROSETTA
1O2F	77	Alfa e Beta	400	I-TASSER
1OF9	77	Alfa	400	I-TASSER
1OPD	85	Alfa e Beta	100	ROSETTA
1PGX	55	Alfa e Beta	100	ROSETTA

Continua na próxima página

Tabela C.1 – Continuação da página anterior

ID PDB	Tamanho Proteína	Tipo	Número de Decoys	Fonte
1PTQ	50	Beta	100	ROSETTA
1R69	61	Alfa	400	I-TASSER
1R69	61	Alfa	100	ROSETTA
1RNB	109	Alfa e Beta	100	ROSETTA
1SCJ	66	Alfa e Beta	100	ROSETTA
1SHF	59	Beta	400	I-TASSER
1SHF	59	Beta	100	ROSETTA
1SRO	71	Beta	400	I-TASSER
1TEN	89	Beta	100	ROSETTA
1TFI	47	Beta	400	I-TASSER
1TIF	59	Alfa e Beta	400	I-TASSER
1TIG	88	Alfa e Beta	400	I-TASSER
1TIG	88	Alfa e Beta	100	ROSETTA
1TUL	102	Beta	100	ROSETTA
1UBI	71	Alfa e Beta	100	ROSETTA
1UGH	82	Alfa e Beta	100	ROSETTA
1URN	90	Alfa e Beta	100	ROSETTA
1UTG	70	Alfa	100	ROSETTA
1VCC	76	Alfa e Beta	400	I-TASSER
1VCC*	77	Alfa e Beta	100	ROSETTA
1VIE	56	Beta	100	ROSETTA
1VLS	146	Alfa	100	ROSETTA
1WHO	94	Beta	100	ROSETTA
2ACY	98	Alfa e Beta	100	ROSETTA
2CHF	128	Alfa e Beta	100	ROSETTA
2CI2	62	Alfa e Beta	100	ROSETTA
2CR7	60	Alfa	400	I-TASSER
2F3N	65	Alfa	400	I-TASSER
2PCY	99	Beta	400	I-TASSER
2REB	60	Alfa e Beta	400	I-TASSER
4UBP	100	Alfa e Beta	100	ROSETTA
5CRO	55	Alfa e Beta	100	ROSETTA
256B	106	Alfa	400	I-TASSER
256B	106	Alfa	100	ROSETTA
Tamanho:	47 - 146	Total:	22600	

* adição de V como último resíduo

APÊNDICE D – COMPARAÇÃO ENTRE RMSD REAL E RMSD PREDITO NO CONJUNTO DE TESTE PELO MODELO SELECIONADO. FONTE: AUTORA

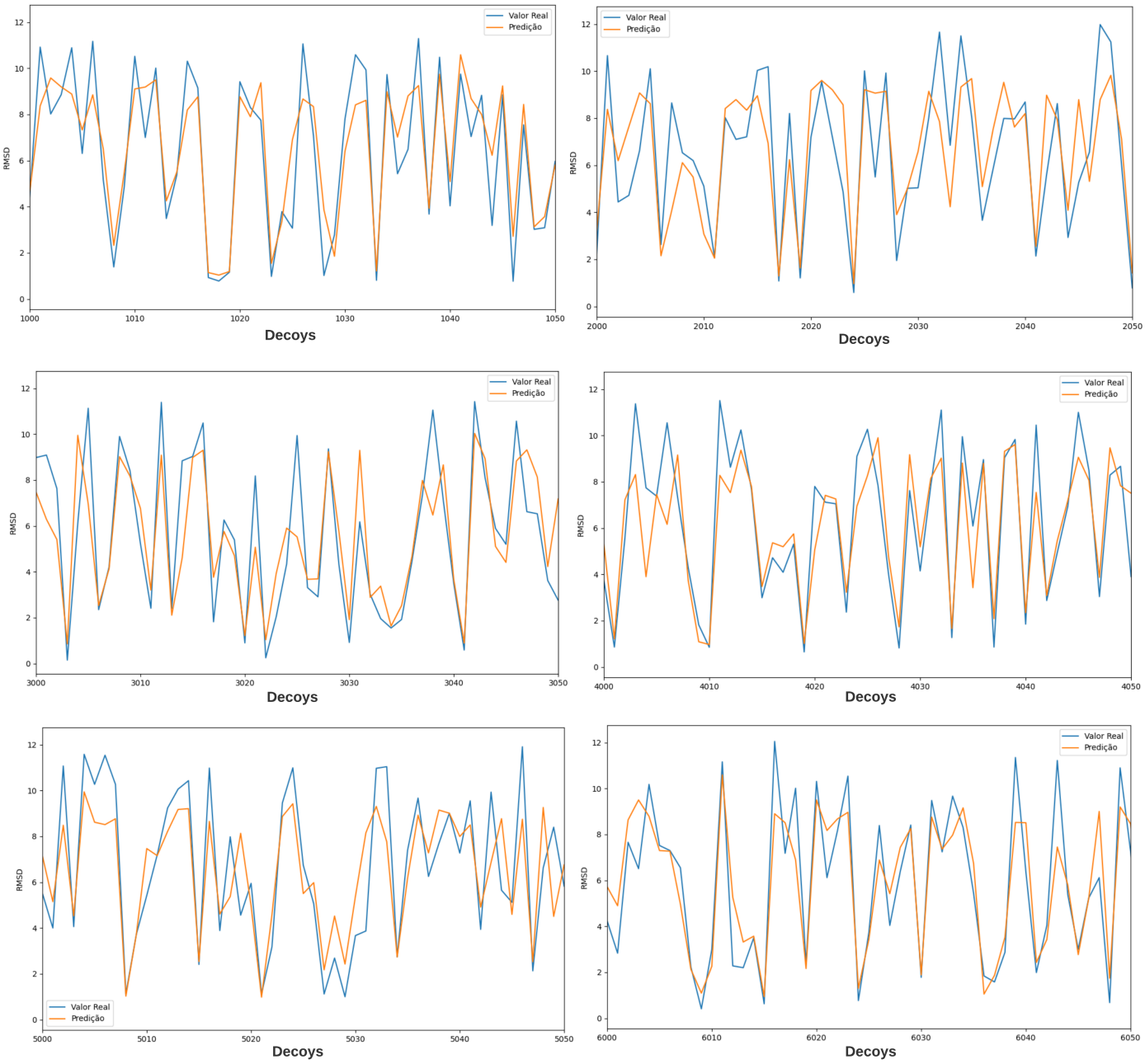


Figura D.1 – Comparação entre o RMSD predito pelo modelo selecionado e o RMSD real em intervalos de 1000 *decoys* mostrando 50 em cada gráfico. Em geral, o comportamento do modelo parece reproduzir o comportamento real. Fonte: Autora.