UNIVERSIDADE DO ESTADO DE SANTA CATARINA – UDESC
CENTRO DE CIÊNCIAS TECNOLÓGICAS – CCT
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO APLICADA - PPGCA

NILCIMAR NEITZEL WILL

UM ALGORITMO EVOLUTIVO DE ESPECIAÇÃO DINÂMICA COM INSERÇÃO DE FRAGMENTOS E ESTRATÉGIA DE SELEÇÃO ADAPTATIVA BASEADA EM INFORMAÇÃO PARA PREDIÇÃO DE ESTRUTURA DE PROTEÍNAS

JOINVILLE
2021

NILCIMAR NEITZEL WILL

UM ALGORITMO EVOLUTIVO DE ESPECIAÇÃO DINÂMICA COM INSERÇÃO DE FRAGMENTOS E ESTRATÉGIA DE SELEÇÃO ADAPTATIVA BASEADA EM INFORMAÇÃO PARA PREDIÇÃO DE ESTRUTURA DE PROTEÍNAS

Dissertação apresentada ao Programa de Pós–Graduação em Computação Aplicada do Centro de Ciências Tecnológicas da Universidade do Estado de Santa Catarina, como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

Orientador: Rafael Stubs Parpinelli

JOINVILLE
2021

Ficha catalográfica elaborada pelo programa de geração automática da

Biblioteca Setorial do CCT/UDESC,

com os dados fornecidos pelo(a) autor(a)

Will, Nilcimar Neitzel
Um algoritmo evolutivo de especiação dinâmica com inserção de fragmentos e estratégia de seleção adaptativa baseada em informação para predição de estrutura de proteínas / Nilcimar Neitzel Will. -- 2021.
84 p.

Orientador: Rafael Stubs Parpinelli
Dissertação (mestrado) -- Universidade do Estado de Santa Catarina, Centro de Ciências Tecnológicas, Programa de Pós-Graduação em Computação Aplicada, Joinville, 2021.

1. Predição de estrutura de proteína. 2. Algoritmos evolutivos. 3. Inserção de fragmentos. 4. Mapas de contato. 5. Especiação. I. Parpinelli, Rafael Stubs. II. Universidade do Estado de Santa Catarina, Centro de Ciências Tecnológicas, Programa de Pós-Graduação em Computação Aplicada. III. Título.

NILCIMAR NEITZEL WILL

UM ALGORITMO EVOLUTIVO DE ESPECIAÇÃO DINÂMICA COM INSERÇÃO DE FRAGMENTOS E ESTRATÉGIA DE SELEÇÃO ADAPTATIVA BASEADA EM INFORMAÇÃO PARA PREDIÇÃO DE ESTRUTURA DE PROTEÍNAS

Dissertação apresentada ao Programa de Pós–Graduação em Computação Aplicada do Centro de Ciências Tecnológicas da Universidade do Estado de Santa Catarina, como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

Orientador: Rafael Stubs Parpinelli

BANCA EXAMINADORA:

Dr. Rafael Stubs Parpinelli
Orientador

Dr. Chidambaram Chidambaram
CCT-UDESC

Dr. César Manuel Vargas Benitez
UTFPR

Joinville, 28 de julho de 2021

Às minhas filhas, Catarina e Lauren, vocês são a minha melhor parte!

AGRADECIMENTOS

Primeiramente, agradeço a Deus e aos meus pais, pois são o motivo da minha existência.

Agradeço a mim, por nunca desistir, apesar das dificuldades.

Agradeço ao meu esposo Maikon Will pela paciência em me ouvir nas intermináveis conversas, pelo apoio e suporte em todos os momentos. Às minhas filhas, Catarina e Lauren, pois minha força e inspiração vem delas.

A todos os professores que tive a oportunidade de assistir as aulas durante o curso, em especial durante a pandemia do Covid-19, muito obrigada pela dedicação. Também em especial a professora Avanilde Kemczinski, pelo incentivo e pelas palavras de motivação que me fizeram persistir.

Aos colegas e amigos que fizeram parte deste percurso, em especial ao Mateus Boiani e a Ellen Flávia Weis Leite, obrigada por toda ajuda.

Por último, deixo um agradecimento especial ao meu orientador por todo apoio, paciência e compreensão. Muito obrigada!

RESUMO

A previsão de estrutura de proteína é um dos problemas fundamentais da Bioinformática Estrutural, que ainda está em aberto. Nos últimos anos diversos métodos tem sido propostos, porém a distância entre o número de sequências de aminoácidos identificadas e a determinação de estruturas tridimensionais de proteína ainda é grande. O uso de informação do problema através da inserção de fragmento, estrutura secundária e mapas de contato podem ajudar a melhorar a exploração do espaço de busca conformacional. Neste trabalho, é proposto um algoritmo evolutivo que usa informação do problema para a predição de estruturas de proteína. O método proposto é dividido em três etapas principais: Inicialização, Otimização e Pós-processamento. A primeira consiste em gerar uma biblioteca de fragmentos com maior diversidade. Para isso, é utilizado o protocolo Quota do Rosetta em conjunto com 3 preditores de estrutura secundária: PSIPRED, SPIDER2 e RaptorX. O objetivo é tentar evitar más conformações devido a previsões incorretas de estruturas secundárias. Na segunda etapa, é executado um algoritmo evolutivo que utiliza uma técnica de especiação dinâmica, chamada DST, para agrupar a população em espécies. O objetivo da técnica DST é identificar diversos bons indivíduos em diferentes locais do espaço de busca. Para orientar a busca pelo espaço conformacional, é utilizada uma estratégia de cruzamento baseada na informação de estrutura secundária do preditor RaptorX. A técnica de inserção de fragmentos é utilizada na mutação, com o objetivo de trazer diversidade. Duas estratégias de seleção, uma baseada em estrutura e outra baseada em contato, são utilizadas para selecionar indivíduos mais aptos para a próxima geração. Um mecanismo auto-adaptativo, baseado na estrutura secundária da população atual, fornece informações para a escolha entre as duas estratégias. Isso favorece um equilíbrio entre contato e estrutura, e evita uma possível dependência dos preditores. Na etapa de Pós-processamento, o melhor indivíduo passa por um procedimento de reembalagem para ter uma representação completa da conformação. O método proposto foi testado em um conjunto com 9 proteínas. Os resultados obtidos foram comparados, em termos de RMSD e GDT, com outros trabalhos da literatura. Os resultados do protocolo Quota foram comparados com o protocolo Best, para isto, foram geradas bibliotecas de fragmentos para ambos os protocolos do Rosetta. Os protocolos foram comparados em termos de RMSD, GDT, tempo de processamento e convergência, além da análise da representação visual das conformações. Os resultados mostram que o método proposto com o protocolo Quota é competitivo com outros métodos da literatura.

Palavras-chave: Predição de Estrutura de Proteína. Algoritmos Evolutivos. Algoritmo baseado em conhecimento. Rosetta. Inserção de fragmentos. Estrutura secundária. Mapas de contato. Especiação.

ABSTRACT

The protein structure prediction problem is one of the fundamental problems of Structural Bioinformatics, that still remains open. In recent years, several methods have been proposed, but the distance between the number of amino acid sequences identified and the determination of three-dimensional protein structures is still great. The use of problem information through fragment insertion, secondary structure, and contact maps can help explore the search space better. An evolutionary algorithm is proposed in this work, which uses this problem information for protein structure prediction. The proposed method is divided into three main stages: Initialization, Optimization and Post-processing. The first stage to generate a fragment library with greater diversity. For this, the Rosetta Quota protocol is used with 3 secondary structure predictors: PSIPRED, SPIDER2 and RaptorX. The objective is to try to avoid bad conformations due to incorrect predictions of secondary structures. In the second step, an evolutionary algorithm is executed that uses a dynamic speciation technique, called DST, to group the population into species. The aim of the DST technique is to identify several good individuals at different locations in the search space. To guide the search for the conformational space, a crossover strategy based on secondary structure information from the RaptorX predictor is used. The fragment insertion is used on mutation, with the aim of diversity. Two selection strategies, one based on structure and one based on contact, are used to select the fittest individuals for the next generation. Based on the SS of the current population, a self-adaptive mechanism provides information for choosing between the two strategies. This favors a balance between contact and structure and avoids possible dependence on predictors. In the Post-Processing step, the best individual goes through a repackaging procedure to have a complete representation of the conformation. The proposed method was tested in a set with 9 proteins. The results obtained were compared, in terms of RMSD and GDT, with other works in the literature. The results of the Quota protocol were compared with the Best protocol, for this, fragment libraries were generated for both Rosetta protocols. The protocols were compared in terms of RMSD, GDT, processing time and convergence, and analysis of the visual representation of the conformations. The results show that the method proposed with the Quota protocol is competitive with other methods in the literature.

Keywords: Protein Structure Prediction. Evolutionary Algorithms. Knowledge-based algorithm. Rosetta. Fragment insertion. Secondary structure. Contact maps. Speciation.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

3D	Tridimensional
ABC	Algoritmo *Artificial Bee Colony*
AE	Algoritmo Evolutivo
AG	Algoritmo Genético
AMBER	*Assisted Model Building with Energy Refinement*
CASP	*Critical Assessment of Protein Structure Prediction*
CHARMM	*Chemistry at HARvard Macromolecular Mechanics*
DE	*Differential Evolution*
DSM-DE	*Dynamic Speciation-based Mutation Differential Evolution*
DST	*Dynamic Speciation Technique*
ES	Estrutura Secundária
FM	*Free-Modelling*
GDT	*Global Distance Test*
GROMOS	*GROningen MOlecular Simulation*
MC	Monte Carlo
PDB	*Protein Data Bank*
PSO	*Particle Swarm Optimization*
PSP	*Protein Structure Prediction*
RMN	Ressonância Magnética Nuclear
RMSD	*Root-Mean-Square Deviation*
TBM	*Template-Based Modelling*

LISTA DE SÍMBOLOS

ϕ	Ângulo Phi
ψ	Ângulo Psi
ω	Ângulo Ômega

SUMÁRIO

1 INTRODUÇÃO

A previsão da estrutura tridimensional de proteína (PSP, *Protein Structure Prediction* em inglês) é um dos problemas fundamentais da Bioinformática, cuja solução ainda está em aberto. A Bioinformática surgiu quando se iniciou a utilização de ferramentas computacionais para a análise de dados genéticos, bioquímicos e de biologia molecular. É uma ciência que envolve áreas como engenharia de software, matemática, estatística, Ciência da Computação e biologia (CARDOSO, 2007). A Bioinformática Estrutural é uma das áreas da Bioinformática, e se preocupa em estudar a estrutura tridimensional de moléculas e macromoléculas, dentre elas pode-se citar a predição de estruturas de proteínas.

As proteínas são macromoléculas formadas por sequências de aminoácidos ligados por ligações peptídicas. Existem 20 tipos de aminoácidos conhecidos, que podem formar milhares de combinações diferentes de proteínas, e cada proteína possui uma sequência única (LOPES; VENSKE, 2015). A maioria é formada por mais de 100 aminoácidos, podendo chegar a milhares (CARDOSO, 2007). As proteínas estão presentes em todos os seres vivos e são responsáveis por regular diversas atividades complexas, o que inclui quase todos os processos biológicos fundamentais necessários a vida. Realizam funções como replicação de DNA, conversão de uma molécula em outra, transporte de oxigênio, catálise, além de funções de defesa, hormonal, estrutural, entre outras (CARDOSO, 2007). Sabendo a função exercida por uma proteína num organismo é possível intervir ativando ou inibindo tal função. Também vale destacar sua importância dentro da Bioinformática Estrutural para o desenvolvimento de novos fármacos.

Estas funções exercidas pelas proteínas estão diretamente relacionadas a sua forma tridimensional. A estrutura da proteína pode ser classificada em primária, secundária, terciária e quaternária, mas é ao assumir a sua estrutura terciária que ela desenvolve a sua funcionalidade, sendo possível prever ou analisar a função que a mesma exerce no organismo (LOPES; VENSKE, 2015). Desta forma, conhecer sua estrutura 3D implica em também conhecer a sua função, o que é fundamental no desenvolvimento de soluções capazes de estimular, restringir ou suspender sua ação biológica (CARDOSO, 2007). Pois, conhecer a estrutura tridimensional de uma proteína fornece informações para o estudo e desenvolvimento de medicamentos para cura e tratamento de certas doenças. Com isso, é possível dizer que o PSP se trata dos estudos para determinar a estrutura tridimensional de proteínas.

Apesar de toda tecnologia e recursos atualmente, das milhares de proteínas existentes, apenas uma pequena parcela das estruturas é conhecida. Com o projeto Genoma dos anos 90 (DORN; BURIOL; LAMB, 2011) houve um aumento na descoberta de novas sequências de aminoácidos, porém o número de estruturas 3D de proteína não seguiu a mesma tendência. Atualmente existe uma grande lacuna entre o número de sequências e o número de estruturas 3D conhecidas, a qual não pode ser preenchida através dos métodos experimentais como Cristalografia por Raio X e Ressonância Magnética Nuclear. Tais métodos tem diversas limitações, mas o principal fato é que são métodos caros, muitas vezes demorados e não tem garantia de

sucesso. Por isso, pesquisadores tem se dedicado a encontrar métodos computacionais de forma a reduzir tempo e os custos necessários para determinar a estrutura terciária da proteína. Visto a importância do PSP na descoberta de novas medicações ou até mesmo cura de certas doenças, se faz necessário a aceleração deste processo de predição de estrutura.

Nos últimos anos diversos métodos tem sido propostos para o problema do PSP, como alternativa aos métodos experimentais. Estes métodos são classificados em 4 grupos de acordo com a utilização da informação estrutural do problema (CORREA; DORN, 2018). (a) Os métodos *ab initio* sem informação de banco de dados, são métodos que utilizam apenas a informação da sequência de aminoácidos, o que torna o espaço de busca para esse problema gigantesco; (b) Os métodos *de novo*, utilizam como entrada a sequência de aminoácidos combinada com informações de banco de dados, o que reduz consideravelmente o espaço de busca; (c) Os métodos de reconhecimento de dobras, e (d) Os métodos de modelagem comparativa. Os métodos (c) e (d) tem a desvantagem de não descobrir novas estruturas, apenas similares as já conhecidas pois são métodos comparativos.

Sob outra perspectiva, ainda é possível dividir estes métodos em 2 grandes grupos (DHINGRA et al., 2020): *Template-Based Modelling* (TBM) e *Template-Free Modelling/Free-Modelling* (FM). No primeiro grupo se encaixam os métodos (c) e (d), os quais utilizam modelos de estruturas já conhecidas em banco de dados para prever novas estruturas. Neles é comum a utilização de técnicas como o alinhamento de sequências para fazer as predições. Este grupo pode alcançar ótimos resultados quando há certa similaridade entre as sequências comparadas, caso contrário deve ser utilizada outra abordagem. No segundo grupo se encaixam os métodos (a) e (b), estes métodos não utilizam modelos de estruturas já conhecidas para fazer previsões, apenas a sequência primária da proteína e suas propriedades físico-químicas. Os métodos puramente *ab initio*, ou métodos do tipo (a), não permitem a utilização de qualquer tipo de informação de banco de dados, por isso conseguem prever novas dobras, mas o custo para isso é muito alto. Neste caso, o espaço de busca é tão grande que torna este tipo de abordagem inviável computacionalmente.

Uma maneira de minimizar o espaço de busca conformacional para métodos *Free-Modelling* (FM) é usar as informações do problema durante o processo de otimização. Entre as fontes mais conhecidas de informação do problema no PSP estão a montagem de fragmentos, previsões de estrutura secundária (ES), mapas de contato e dinâmica molecular. Os métodos do tipo (b) ou *de novo* tem o intuito de fazer isto. Ao mesmo tempo que não comparam sequências inteiras de proteínas, tentam melhorar a exploração do espaço de busca utilizando informações que podem vir de pequenos fragmentos de estruturas já conhecidas, ou de preditores de estrutura secundária (ES) e preditores de contato. Atualmente existem diversos preditores que podem fornecer informações diversas, como a ES prevista para cada aminoácido, e alguns com maior precisão informam a probabilidade de ocorrência para cada tipo de estrutura, além de outras informações disponíveis. Basicamente é isto que diferencia os preditores um do outro, o tipo de informação fornecida.

O uso de informação do problema aplicado ao PSP tem se tornado cada vez mais

frequente, pois além de reduzir a complexidade dos métodos *ab initio*, ainda consegue prever novas dobras. De acordo com (DHINGRA et al., 2020), a técnica de montagem de fragmentos é a técnica de maior sucesso entre os métodos FM. Mais recentemente tem sido aplicado em conjunto com o uso de informações vindas de preditores de ES ou mapas de contato. Em (ZHANG et al., 2020), os autores utilizaram fragmentos, ES e contato para melhorar a predição de estrutura de proteína num modelo de objetivo único. Já em (MARCHI; PARPINELLI, 2021) os 3 tipos de informação são utilizados para compor um modelo multi-objetivo para o PSP. Em (CORREA; DORN, 2020; ZHANG et al., 2019) foram utilizados fragmentos e ES. Já em (PENG; ZHOU; ZHANG, 2020) os fragmentos foram utilizados em conjunto com os mapas de contato para melhorar as previsões. Em (SILVA, 2019) os fragmentos foram utilizados em um algoritmo evolutivo com auto adaptação de parâmetros. O método proposto neste trabalho se encaixa na abordagem *de novo*, pois não há utilização de modelos nas previsões, mas é utilizado informação do problema na forma de fragmentos, estrutura secundária e mapas de contato.

Em termos computacionais, para se trabalhar com o PSP, é necessário que se tenha uma representação computacional da proteína, uma função de energia eficaz e um algoritmo de otimização para encontrar possíveis soluções no espaço de busca conformacional (DORN et al., 2014). Representar uma proteína computacionalmente pode ser uma tarefa bastante desafiadora, pois está diretamente relacionada a sua complexidade. Obviamente quanto mais características da estrutura real de uma proteína se consegue representar, mais próxima ela estará de representar sua forma como encontrada na natureza. Porém, representar uma proteína em toda a sua complexidade tem custo computacional muito alto. Por isso, existem algumas representações mais simplificadas, que conseguem reduzir a complexidade do problema através da abstração de alguns conceitos (CORREA; DORN, 2020). Dentre elas, duas ganham maior destaque na literatura. Uma representa os aminoácidos por sua posição Cartesiana (x, y, z), e a outra através dos ângulos diédricos (ϕ, ψ, ω). Neste trabalho os polipeptídeos serão representados pelos ângulos diédricos, o que reduz a complexidade do modelo completo mas mantém as características necessárias para a predição tridimensional.

Além da representação computacional da proteína, trabalhar com uma abordagem *ab initio* ou *de novo* requer que se tenha uma função de energia. Esta é necessária segundo o postulado de (ANFINSEN, 1973), que diz que a estrutura tridimensional ou nativa de uma proteína se encontra no seu estado de energia potencial mínima. Logo, uma função de energia deve estimar o quanto um modelo está próximo da sua conformação nativa. Uma função de energia pode incluir diversos termos como ângulos de ligação, interações eletrostáticas, ligações de hidrogênio, entre outras. As funções de energia podem ser baseadas em física e mecânica quântica, ou podem ser baseadas em conhecimento, e neste caso elas são deduzidas empiricamente a partir de estruturas já resolvidas em banco de dados. Neste trabalho, são utilizadas as funções de energia do Rosetta[1], que são funções baseadas em conhecimento, ou seja, foram desenvolvidas a partir de uma análise

[1] Rosetta é um pacote de software que inclui vários algoritmos para a análise de estruturas de proteínas. Pode ser acessado em <https://www.rosettacommons.org/>

empírica das estruturas contidas no PDB[2] (*Protein Data Bank*) (COMBS et al., 2013).

Com relação ao método de busca, partindo-se do postulado de Anfinsen, o método deve ser capaz de procurar uma conformação de menor energia potencial, em meio a inúmeras conformações em um amplo espaço de busca. Ou seja, trata-se de um algoritmo de otimização, cujo o objetivo deve ser minimizar a energia potencial da proteína. De acordo com (HAO et al., 2016), a alta dimensionalidade do espaço conformacional da proteína e sua complexidade tornam o PSP um problema NP-difícil. Esta característica traz a necessidade de trabalhar com métodos meta-heurísticos, nos quais Algoritmos Evolutivos (EA) como Algoritmos Genéticos (AG) e Evolução Diferencial (ED) são representantes (PARPINELLI et al., 2019). Esses são algoritmos inspirados no processo darwiniano de seleção natural. Com base na população, eles têm indivíduos (ou soluções candidatas) que evoluem ao longo das gerações por meio de processos de cruzamento, mutação e seleção.

Neste trabalho, é aplicado um Algoritmo Evolutivo baseado em especiação dinâmica, utilizando informação de estrutura secundária, mapas de contato e inserção de fragmentos. A combinação da técnica de especiação dinâmica com uso de informações do problema, visa melhorar a exploração do espaço de busca sem perder a diversidade. Além disso, são usados diversos preditores de estrutura secundária para a geração dos fragmentos. O objetivo é aumentar a diversidade dos fragmentos e evitar que estruturas incorretas limitem o modelo. Em vista disso, pode-se dizer que as principais contribuições deste trabalho estão no uso e avaliação do protocolo Quota aliado aos diversos tipos de informação do problema.

1.1 MOTIVAÇÃO

Recentemente houve um grande avanço no que diz respeito a predição de estruturas de proteínas com a utilização de técnicas de aprendizado profundo (JUMPER et al., 2020). Porém, acredita-se que o problema ainda está em aberto, pois os métodos computacionais atuais ainda não conseguem prever uma solução ótima para o problema. Apesar de todo o avanço dos últimos anos, ainda há muito o que se investigar a respeito de como as estruturas tridimensionais se formam. Na verdade, o PSP tem se mostrado um problema bastante atual, o que desperta cada vez mais o interesse de novos pesquisadores e ressalta a importância de novas soluções.

Os preditores de estrutura secundária também tem evoluído, se tornando cada vez mais precisos e trazendo diversos tipos de informação sobre o problema. Da mesma forma, os mapas de contato têm se mostrado um grande aliado na predição de estruturas de proteínas. Sendo cada vez mais utilizados de maneira complementar a outras técnicas e como forma de contornar possíveis informações incorretas. Nos últimos anos, métodos baseados em conhecimento tem provado obter os melhores resultados para a abordagem FM. Combinando meta-heurísticas com diversos tipos de informação do problema é possível melhorar a exploração do espaço de busca e acelerar o processo de predição sem perder a diversidade.

[2] Protein Data Bank disponível em <https://www.rcsb.org/>

1.2 OBJETIVOS

O principal objetivo deste trabalho é desenvolver um modelo meta-heurístico que incorpore diversas formas de informação do problema, visando a diversidade e a melhora na exploração do espaço de busca, aplicado ao problema de predição de estrutura de proteínas. Para isto, as seguintes tarefas devem ser cumpridas:

1. Estudar as principais características do problema do PSP, principalmente relacionadas a abordagem *de novo*;
2. Pesquisar os algoritmos de otimização para o PSP;
3. Realizar um estudo a cerca das informações do problema que podem ser aplicadas ao PSP, com foco em fragmentos, estrutura secundária e mapas de contato;
4. Implementar um modelo para previsão de estruturas 3D de proteínas, utilizando uma meta-heurística e informações do problema;
5. Avaliar o desempenho do método proposto em várias proteínas;
6. Comparar o desempenho alcançado com abordagens recentes.

1.3 ESTRUTURA DO DOCUMENTO

O restante deste documento está organizado como a seguir: Capítulo 2 se trata do *Background*, onde é apresentado todo o embasamento teórico necessário a compreensão deste trabalho, bem como as questões biológicas das proteínas, as questões que envolvem o problema do PSP, a montagem de fragmentos, a predição de estrutura secundária e os mapas de contato. O Capítulo 3 apresenta os trabalhos relacionados, tanto com relação as meta-heurísticas utilizadas para o PSP, quanto da utilização de informação do problema. O Capítulo 4 apresenta o método proposto, bem como um fluxograma e todas as etapas envolvidas no processo. O Capítulo 5 apresenta os experimentos, resultados e suas respectivas análises. O Capítulo 6 apresenta as considerações finais e trabalhos futuros.

2 BACKGROUND

Este capítulo apresenta diversos conceitos necessários a compreensão deste trabalho. Na Seção 2.1 são apresentados os conceitos de fundamentação biológica das proteínas, que vai desde sua composição por aminoácidos até os níveis estruturais. Este embasamento teórico é necessário para facilitar o entendimento do problema. Na Seção 2.2 é feita a descrição do problema de predição de estruturas de proteína, onde é discutido sua complexidade, função de energia e representação computacional, além das abordagens computacionais para o problema do PSP. Na Seção 2.3 é apresentada a técnica de montagem de fragmentos e o pacote Rosetta que contém diversas funcionalidades baseadas em fragmentos. Por último, na Seção 2.4 é apresentada a predição de estrutura secundária e sua importância para a predição de estrutura terciária. O capítulo termina com a Seção 2.5 onde é apresentada a técnica de mapas de contato e como são utilizados no PSP.

2.1 PROTEÍNAS

As proteínas são macromoléculas formadas por cadeias lineares de aminoácidos. Existem 20 aminoácidos naturais conhecidos (ver Tabela 1), e todos possuem uma estrutura comum conforme a Figura 1. Cada aminoácido é uma pequena molécula que possui um carbono α central ou C_α, ligado a um átomo de Hidrogênio (H), um grupo Carboxila ($COOH$), um grupo Amina (NH_2), e uma cadeia lateral, também chamada de grupo R. A cadeia lateral R é o que diferencia um aminoácido do outro e gera as propriedades físico-químicas de cada um (LODISH et al., 2008).

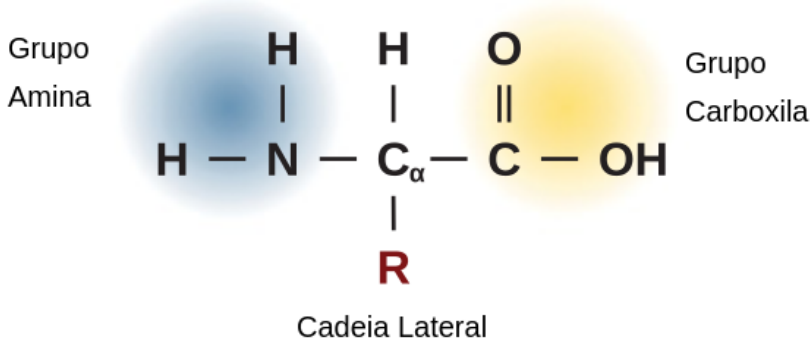


Figura 1 – Estrutura de um aminoácido. Fonte: Adaptado de (GARZA-FABRE et al., 2016)

Os aminoácidos se ligam entre si através de ligações peptídicas conforme a Figura 2, onde um átomo de oxigênio do grupo carboxila de um aminoácido se liga a um átomo de hidrogênio do grupo amina de outro aminoácido, liberando assim uma molécula de água (H_2O), e por isso também são conhecidos como resíduos de aminoácido. Uma cadeia com 2 ou mais aminoácidos ligados é também conhecida como peptídeo, e cadeias maiores são conhecidas como polipeptídeos ou proteínas (SILVA, 2019).

Tabela 1 – Os 20 aminoácidos naturais conhecidos.

Nome	Sigla 3 letras	Sigla 1 letra
Alanina	Ala	A
Arginina	Arg	R
Asparagina	Asn	N
Ácido aspártico	Asp	D
Ácido glutâmico	Glu	E
Cisteína	Cys ou Cis	C
Glicina	Gli ou Gly	G
Glutamina	Gln	Q
Histidina	His	H
Isoleucina	Ile	I
Leucina	Leu	L
Lisina	Lis ou Lys	K
Metionina	Met	M
Fenilalanina	Phe ou Fen	F
Prolina	Pro	P
Serina	Ser	S
Tirosina	Tir ou Tyr	Y
Treonina	Tre ou Thr	T
Triptofano	Trp	W
Valina	Val	V

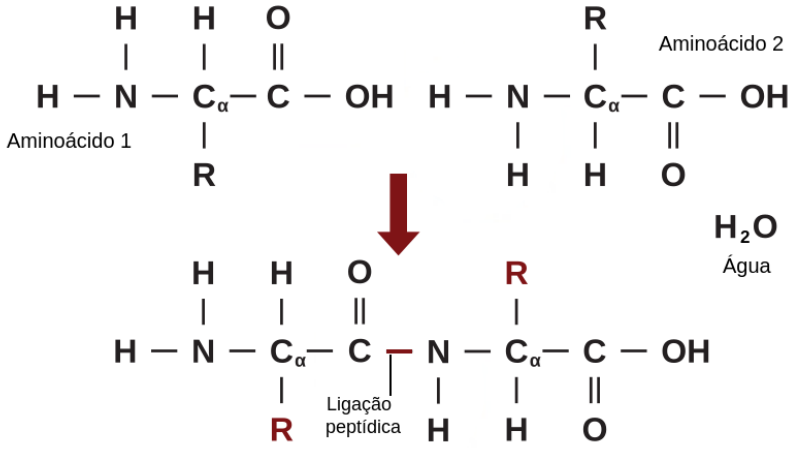


Figura 2 – Ligação peptídica. Fonte: Adaptado de (GARZA-FABRE et al., 2016)

A ligação peptídica entre dois aminoácidos, representada pela ligação $C-N$, forma o ângulo diedro ω (Omega). Sendo uma ligação dupla tem menos liberdade para girar, podendo assumir valores próximos a 0° (cis) ou 180° (trans) (DORN et al., 2014). As rotações que são livres para girar estão ligadas ao carbono α central e formam os ângulos diedros ϕ (Phi) e ψ (Psi). Sendo ϕ o ângulo de $N-C_\alpha$ e ψ o ângulo de $C_\alpha-C$, ambos podem assumir valores de -180° a 180°, como mostra a Figura 3. A tupla ϕ, ψ e ω, é conhecida como ângulos diédricos, e sua

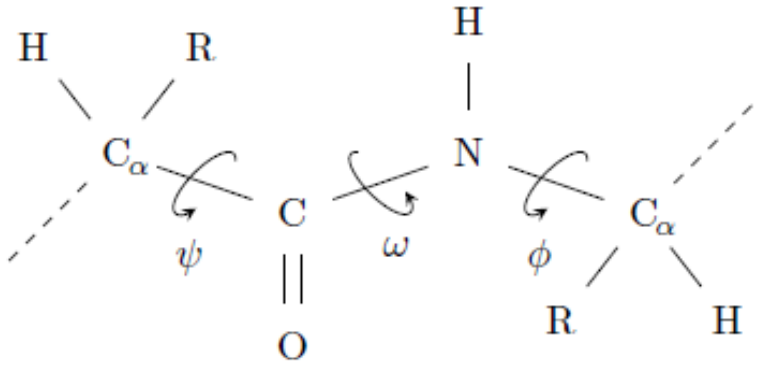


Figura 3 – Ângulos diédricos de um aminoácido. Fonte: (SILVA, 2019)

sequência de ligações é conhecida como backbone (SILVA, 2019). O conjunto destes 3 ângulos é responsável por determinar a conformação do esqueleto peptídico da proteína (CORREA; DORN, 2020). Tem ainda os ângulos da cadeia lateral, estes também são livres para girar e são chamados de ângulos Chi (χ). Como a estrutura da cadeia lateral depende de cada aminoácido, a quantidade de ângulos χ também varia, podendo ser de 0 a 4 ângulos por aminoácido. Sendo que todos podem assumir valores de -180° a 180°.

2.1.1 Estruturas de Proteínas

Como visto até o momento, as proteínas são sequências de aminoácidos, que sob condições fisiológicas se dobram até atingir sua forma tridimensional. Durante este processo de dobramento, é possível destacar 4 níveis estruturais da proteína, com o intuito de facilitar o estudo das conformações adotadas, sendo elas: estrutura primária; estrutura secundária; estrutura terciária e estrutura quaternária (BRANDEN; TOOZE, 1999).

A estrutura primária é a sequência de aminoácidos ligados através de ligações peptídicas. Esta sequência é única para cada proteína e é o que determina a sua conformação (DORN et al., 2014). Devido a estrutura dos aminoácidos, visto na Figura 1, sua sequência tem duas pontas distintas. A sequência inicia na ponta esquerda, com um grupo Amina livre, por isso é chamado de terminal amina ou N-terminal, e termina na ponta direita, onde tem um grupo Carboxila livre, chamado terminal carboxila ou C-terminal (Khan Academy, 2021) (ver Figura 4). Uma simples mudança na sequência de aminoácidos da proteína, como a alteração de apenas um aminoácido na sequência pode afetar a estrutura da proteína e sua função, o que pode estar associado a doenças. A Figura 4 representa uma proteína em sua forma de estrutura primária. Na imagem, cada esfera representa um aminoácido.

A estrutura secundária consiste em padrões estruturais formados a partir de ligações de hidrogênio entre os átomos do backbone, ou seja, não envolve os átomos da cadeia lateral *R*. São interações que ocorrem entre os átomos de *H* dos grupos Amina e os átomos de *O* dos grupos Carboxila de diferentes aminoácidos, durante o processo de dobramento. Em geral, estas ligações garantem a estabilidade das estruturas secundárias (DORN et al., 2014).

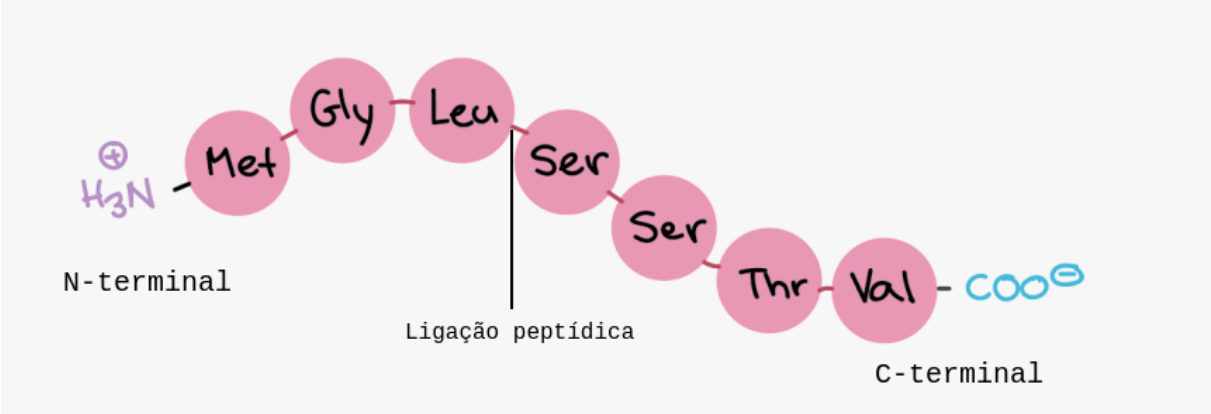


Figura 4 – Estrutura primária. Fonte: Adaptado de (Khan Academy, 2021)

Atualmente existem 8 tipos de estruturas secundárias aceitas, sendo 3 do tipo hélice, 2 do tipo folha, voltas, alças e bobinas (GEOURJON; DELéAGE, 1995). As estruturas do tipo hélice e folha são mais estáveis e por isso são conhecidas como estruturas regulares. Já as estruturas voltas, alças e bobinas são menos estáveis, por isso são chamadas de irregulares.

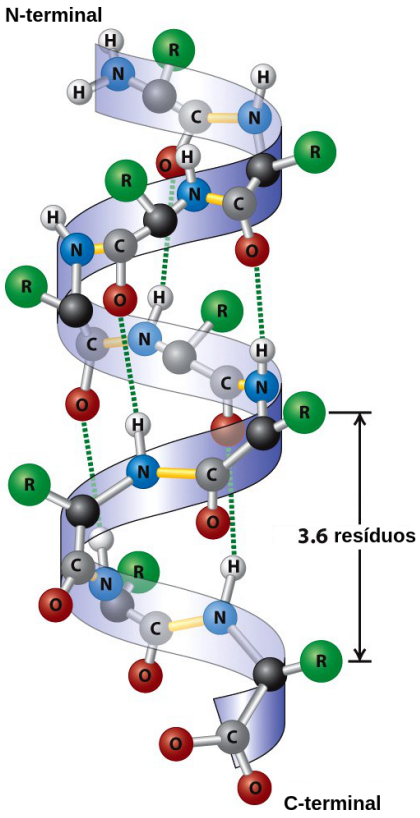


Figura 5 – α-*helix*. Fonte: Adaptado de (LODISH et al., 2008)

As hélices se caracterizam pelas ligações de hidrogênio que ocorrem entre 2 aminoácidos distantes entre si por k aminoácidos, onde k pode assumir valores de 3, 4 ou 5. Em outras palavras, as ligações ocorrem entre um grupo Amina de um aminoácido e um grupo Carboxila de outro aminoácido após k posições. Isso garante o formato de hélice deste tipo de estrutura. Para

$k = 3$ diz-se que é uma 3^{10}-*helix*, quando $k = 4$ é chamada de α-*helix*, e para $k = 5$ chama-se de π-*helix*. A Figura 5 apresenta o formato de uma α-*helix*, sendo a mais comum entre as hélices, ela corresponde aproximadamente a 34% das estruturas secundárias. Enquanto que a 3^{10}-*helix* e a π-*helix* só correspondem a 1% e 4% das estruturas, respectivamente (BORGUESAN et al., 2015).

Diferente das hélices, as folhas são caracterizadas por sua superfície plana. Chamadas de folhas β, este tipo de estrutura secundária ocorre quando uma cadeia de aminoácidos estendida se liga a outra cadeia vizinha em um mesmo plano (CORREA; DORN, 2020). Esta formação tem uma aparência de folha dobrada. Existem 2 tipos de folhas, β-*strand* e β-*sheet*, e correspondem aproximadamente a 25% e $1,3$% das estruturas secundárias. As folhas β-*strands* são compostas de 3 a 10 aminoácidos, enquanto que as folhas β-*sheets* são formadas por 2 ou mais β-*strands* (SILVA, 2019). As β-*sheets* podem ser paralelas (apontando para a mesma direção, ver Figura 6) ou antiparalelas (em direções opostas, ver Figura 7).

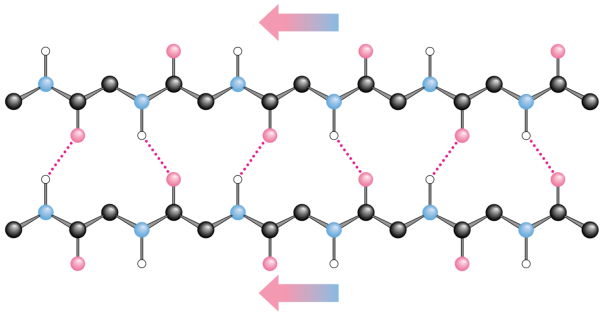


Figura 6 – β-*sheet* paralela. Fonte: (GARRETT; GRISHAM, 1999)

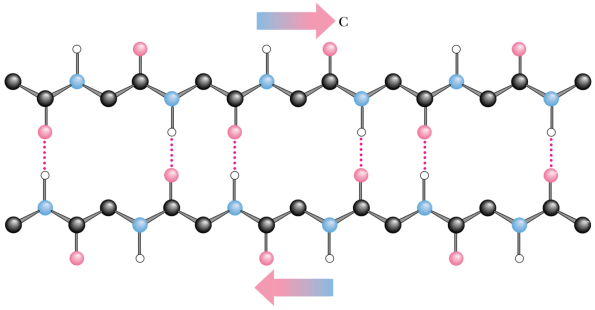


Figura 7 – β-*sheet* antiparalela. Fonte: (GARRETT; GRISHAM, 1999)

Finalmente, as alças (*turns*), voltas (*loops*) e bobinas (*coils*) não representam padrões regulares de estrutura secundária. Este tipo de formação ocorre entre as estruturas regulares, como por exemplo, antes ou depois de uma hélice ou folha. Uma volta liga duas hélices ou folhas, quando a volta é curta e forma um ângulo de 180°, é chamado de alça. As bobinas estão localizadas no início e no final das proteínas. As voltas e alças representam aproximadamente 19% das estruturas e as bobinas 16% (SILVA, 2019). A Figura 8 mostra os diferentes tipos de estruturas secundárias.

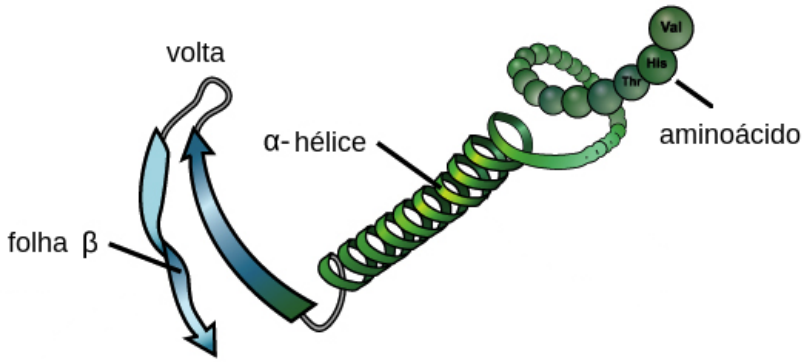


Figura 8 – Estruturas secundárias. Fonte: Adaptado de (OpenStax, 2021)

O próximo nível estrutural é a estrutura terciária da proteína, que é definida pela combinação entre as estruturas secundárias e como estas estão conectadas e posicionadas no espaço 3D. A estrutura terciária é também chamada de estrutura nativa da proteína ou funcional (CORREA; DORN, 2020). O principal responsável pela formação tridimensional da proteína são as interações entre os grupos R, o que incluem ligações de hidrogênio, ligações iônicas, interações hidrofóbicas e hidrofílicas, além das ligações dissulfeto entre outras (DORN et al., 2014). A Figura 9 mostra diversas ligações que podem ocorrer na formação da estrutura terciária.

Acredita-se que ao assumir sua estrutura terciária a proteína se encontra no seu estado mais estável e que sua funcionalidade esteja relacionada ao seu estado tridimensional (ANFINSEN, 1973). Apesar disto, sabe-se que algumas proteínas precisam ser desordenadas para desempenharem sua função, estas representam cerca de 30% das sequências conhecidas (DUNKER et al., 2001). Ainda assim, conhecer a estrutura 3D de uma proteína fornece informações importantes a respeito da função que a mesma irá exercer sobre um organismo (CORREA; DORN, 2020). Este conhecimento é fundamental para o desenvolvimento de medicamentos para o tratamento de diversas doenças.

O último nível hierárquico é a estrutura quaternária, é quando mais de uma cadeia polipeptídica está envolvida na conformação da proteína. Existem proteínas que são formadas por apenas uma cadeia de aminoácidos, neste caso elas tem apenas 3 níveis de estrutura, ou seja, terminam na estrutura terciária. Porém, algumas proteínas necessitam da junção de várias cadeias para desempenharem o seu papel, como por exemplo, a hemoglobina (ver Figura 10). Esta formação é conhecida como estrutura quaternária e é a junção de várias estruturas terciárias, que neste caso são chamadas de subunidades. Uma estrutura quaternária se mantem estável pelos mesmos tipos de interações que contribuem para a estrutura terciária (ligações de hidrogênio, interações hidrofóbicas e hidrofílicas) (LODISH et al., 2008).

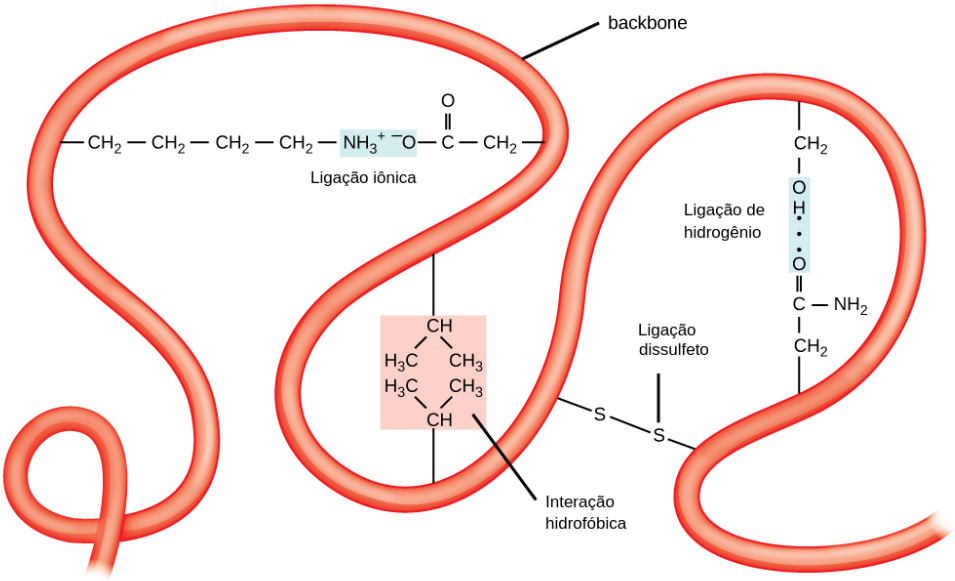


Figura 9 – Interações químicas em uma estrutura terciária. Fonte: Adaptado de (OpenStax, 2021)

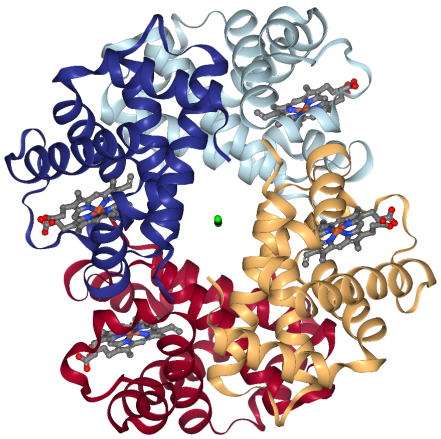


Figura 10 – Estrutura quaternária da Hemoglobina. Fonte: (RCSB PDB, 2021)

2.2 PREDIÇÃO DE ESTRUTURA DE PROTEÍNA

Em resumo, o problema da predição de estrutura de proteína (PSP) consiste em determinar a estrutura tridimensional de uma proteína a partir de sua sequência de aminoácidos. Tal afirmação é baseada no postulado de Anfinsen (1973), o qual afirma que a sequência de aminoácidos de uma proteína contém todas as informações necessárias para determinar sua estrutura terciária. Isso foi provado a partir de experimentos, nos quais uma proteína desdobrada pode ser novamente redobrada em sua forma nativa quando suas condições fisiológicas são restauradas. Isso significa que as informações físico-químicas de uma sequência de aminoácidos é suficiente para prever sua estrutura tridimensional (DORN; SOUZA, 2008).

Existem 2 métodos principais a partir dos quais é possível determinar a estrutura nativa

de uma proteína, são a cristalografia de raios-X e a Ressonância Magnética Nuclear (RMN). São métodos experimentais, também chamados de métodos "in vitro", e necessitam de um ambiente laboratorial para serem executados. A cristalografia de raios X é o método mais antigo e mais utilizado. Também é o método que mais resolveu estruturas 3D até hoje, podendo-se considerar que 90% delas tenha sido por difração de raios-X (CALLAWAY, 2015). Essa técnica consiste em deixar passar um feixe de raios X através de uma substância cristalizada. No entanto, apesar do seu extenso uso existem algumas limitações nesta técnica. A maior dificuldade está na cristalização, pesquisadores podem levar anos para formar cristais de algumas proteínas de forma que seja adequado para análise, além disso, algumas proteínas, como por exemplo as proteínas de membrana, desafiam a cristalização, podendo ocasionar a perda de algumas de suas propriedades (NELSON; LEHNINGER; COX, 2008).

A partir da década de 80, a técnica de RMN começou a se tornar mais popular no estudo de estruturas macromoleculares. A RMN trouxe a vantagem de poder trabalhar com proteínas em meio aquoso, em comparação com as amostras cristalizadas utilizadas pela difração de raios-X. No entando, os 2 métodos apresentam desvantagens, além do custo elevado, do tempo e esforço investidos na resolução de uma única proteína, ambos não garantem a qualidade dos resultados. Mais recentemente, uma nova técnica chamada Cryo-EM (Cryo-Electron Microscopy, em inglês) vem sendo utilizada na determinação de estruturas de proteínas. Esta técnica consiste em fotografar moléculas congeladas para determinar sua estrutura. No entanto, é uma técnica nova e aparenta precisar de alguns ajustes, ainda distante de substituir a difração por raios-X (DHINGRA et al., 2020).

Na década de 90 teve início o Projeto Genoma[1], cujo objetivo era determinar as sequências dos 3 bilhões de bases químicas que compõem o DNA humano. O projeto resultou na descoberta de inúmeras sequências de proteína (GOES; OLIVEIRA, 2014). Porém, com as dificuldades e lentidão dos métodos experimentais a descoberta das estruturas tridimensionais não seguiu a mesma tendência. Hoje, existe uma grande lacuna entre o volume de dados gerados pelo Genoma e o número de estruturas 3D conhecidas. De todas as sequências de proteínas conhecidas e não redundantes, menos de 1% tem sua estrutura tridimensional conhecida (CORREA; DORN, 2020). Por isso, pesquisadores tem se dedicado a encontrar métodos computacionais de forma a reduzir tempo e custos necessários para determinar a estrutura terciária de proteínas.

Nos últimos anos diversas abordagens computacionais tem sido propostas para o problema do PSP, como alternativa aos métodos experimentais. Estes métodos são classificados em 4 grupos de acordo com a utilização da informação estrutural do problema (CORREA; DORN, 2018). (a) Os métodos *ab initio* sem informação de banco de dados; (b) Os métodos *de novo*, utilizam como entrada a sequência de aminoácidos combinada com informações de banco de dados; (c) Os métodos de reconhecimento de dobras, e (d) Os métodos de modelagem comparativa. Sob a perspectiva das proteínas, estes métodos podem ser divididos em 2 grandes grupos, considerando a utilização de modelos na predição ((DHINGRA et al., 2020); (MISHRA;

[1] <https://www.genome.gov/>

HOQUE, 2019)): *Template-Based Modelling* (TBM) e *Template-Free Modelling/Free-Modelling* (FM). O primeiro grupo diz respeito as proteínas que possuem similaridade com sequências já identificadas experimentalmente, neste caso são utilizados os métodos (c) e (d), os quais utilizam modelos nas previsões. O segundo grupo diz respeito as proteínas que não possuem similaridade com sequências já identificadas, neste caso utiliza-se os métodos (a) e (b), que não utilizam modelos em suas previsões. A Figura 11 apresenta um fluxograma dessa classificação de acordo com a avaliação do CASP[2].

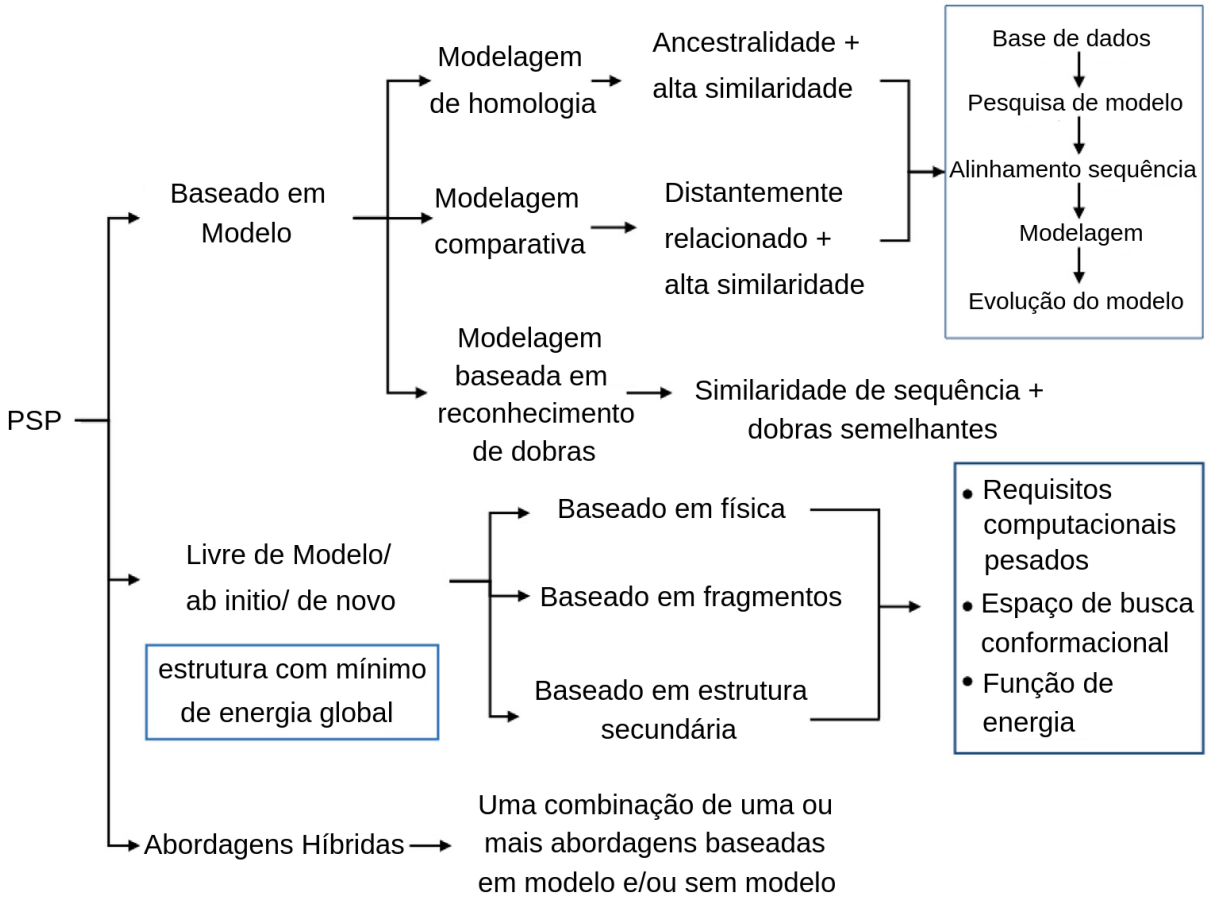


Figura 11 – Abordagens computacionais para o PSP. Fonte: Adaptado de (DHINGRA et al., 2020)

O CASP (*Critical Assessment of protein Structure Prediction*) é uma competição mundial conhecida, que ocorre a cada 2 anos para avaliar a melhoria no campo da previsão computacional de estrutura de proteína. No fluxograma da Figura 11, o primeiro nível apresenta as abordagens principais e ainda uma abordagem híbrida para a previsão de estruturas. O próximo nível apresenta as principais estratégias para cada abordagem. No último nível são citadas as principais etapas envolvidas em cada abordagem. Esta figura apresenta uma classificação ampla das abordagens mais comuns no PSP, mas vale lembrar que existem outras técnicas e estratégias, como por exemplo o uso de mapas de contato na abordagem FM.

[2] CASP website: <https://predictioncenter.org/>

O primeiro grupo utiliza modelos de proteínas semelhantes para fazer previsões. Isso se baseia na ideia de que sequências semelhantes se dobram de forma semelhante. Neste grupo estão incluídas as seguintes abordagens: Modelagem de Homologia (HM), Modelagem Comparativa (CM) e *Threading* (reconhecimento de dobras). A diferença entre HM e CM, é que na HM o modelo compartilha uma ancestralidade com a sequência alvo, já na CM as sequências não tem relação evolutiva, apenas compartilham uma similaridade de sequência. O modelo de *Threading* é utilizado quando não se possui proteínas homólogas conhecidas. Nesta técnica busca-se encontrar um modelo que tenha a maior quantidade de dobras semelhantes através do alinhamento das sequências. Em geral, os 3 métodos são bastante semelhantes, principalmente pelo fato dos 3 precisarem de proteínas homólogas ou quase homólogas para sua aplicação. Estes métodos costumam ser bastante eficazes desde que compartilhem pelo menos 30% de similaridade entre as sequências (DHINGRA et al., 2020). A grande dificuldade destes métodos está em prever novas dobras, uma vez que só podem ser aplicados se já houver informações de estruturas já conhecidas, ou seja, estes métodos são dependentes das bibliotecas de dobras de proteínas.

No segundo grupo estão os métodos que não utilizam modelos homólogos para fazer previsões, apenas a sequência de aminoácidos e suas propriedades físico-químicas. Neste grupo encontram-se os métodos *ab initio* e *de novo*. Os métodos *ab initio* ou métodos de primeiros princípios não fazem uso de informação de banco de dados, mas se baseiam apenas nos princípios baseados em física e termos de energia. Essa ideia é baseada no fato de que ao atingir a estrutura nativa a proteína se encontra no seu estado de menor energia potencial (ANFINSEN, 1973; DORN et al., 2014). O objetivo deste tipo de abordagem é prever a conformação mais estável da proteína com a menor energia livre. Para isto, estes métodos dependem de algoritmos capazes de pesquisar o espaço conformacional em busca da melhor conformação através de uma função de energia eficaz (DHINGRA et al., 2020).

O fato dos métodos *ab initio* não dependerem de modelos de proteínas já existentes torna possível a previsão de novas dobras. Mas, por não utilizarem nenhum outro tipo de informação do problema, o aumento no espaço de busca conformacional torna o PSP um problema quase impossível de ser tratado computacionalmente. De fato, o PSP já é um problema bastante complexo por si só, dado seu extenso espaço de busca e as inúmeras possíveis conformações de uma molécula de proteína, é classificado de acordo com a teoria da complexidade computacional como um problema NP-completo (CORREA; DORN, 2020; DORN et al., 2014). Além disso, o espaço conformacional cresce a medida que aumenta o número de aminoácidos nas sequências de proteínas.

Como forma de melhorar a exploração do espaço de busca e ao mesmo tempo prever novas dobras, surgem os métodos *de novo*. Estes tem o objetivo de aliar o melhor dos 2 mundos, usando os conhecimentos dos métodos baseados em modelo com a vantagem dos métodos FM. Também chamada de classe híbrida, o que estes métodos fazem é utilizar informações de fragmentos de aminoácidos de banco de dados para construir estruturas 3D de proteína. Em outras palavras, a abordagem *de novo* tenta inserir informação do problema de forma a melhorar

a exploração do espaço de busca e acelerar o processo de predição. Porém, diferente dos métodos baseados em modelo, estes não comparam sequências inteiras com estruturas conhecidas, apenas comparam pequenos fragmentos, na tentativa de obter informações que auxiliem na conformação. Assim, estes métodos se baseiam na ideia de que interações locais podem definir estruturas locais. Mas isso não é uma verdade absoluta. Por isso trabalhar com métodos baseados em fragmentos requer alguns cuidados, como por exemplo a definição de funções de pontuação. As funções de pontuação são usadas para avaliar o relacionamento entre os fragmentos, ou seja, ela vai determinar a probabilidade de inserção de um fragmento durante o processo de previsão (DORN et al., 2014). O método proposto nesse trabalho se encaixa na categoria dos métodos *de novo*, pois utiliza informação do problema sem fazer uso de modelos de proteína.

Em resumo, o PSP se trata de um problema de otimização, onde o objetivo é minimizar a energia potencial da proteína de forma a encontrar sua conformação nativa. Em vista disso, para que as aplicações computacionais tenham sucesso, os métodos FM precisam atender 3 requisitos (DORN et al., 2014): uma função de energia eficaz, uma representação computacional da proteína e um método de busca para encontrar conformações no espaço de busca. A seguir, cada um destes pontos são discutidos em maiores detalhes.

2.2.1 Representação Computacional

Para se trabalhar com predição de proteínas de maneira computacional é necessário que estas sejam representadas de alguma forma. Já foi discutido neste trabalho o quão complexo é a estrutura 3D de uma proteína, e que o tamanho da proteína aumenta a complexidade do problema. Em ciência disto, é possível supor que a complexidade computacional também cresça de acordo com o nível de detalhes abordados em sua representação computacional. Em contrapartida, quanto maior o nível de detalhes que se consegue abordar computacionalmente, mais preciso será o modelo e mais próximo da estrutura encontrada na natureza. Logo, quanto menor a quantidade de detalhes, menor a sua representatividade em comparação com uma proteína real. Ou seja, escolher a forma de representar computacionalmente a estrutura tridimensional de uma proteína é um grande desafio, pois está diretamente relacionado a sua complexidade. Claramente, trabalhar com um maior nível de detalhes envolve um custo alto, da mesma forma como reduzir muito os detalhes pode prejudicar a pesquisa pela falta de informação. Então deve haver um equilíbrio entre a quantidade de detalhes abordados, mantendo a quantidade de informação necessária sem sobrecarga computacional.

As representações mais detalhadas são chamadas de *all-atom*. Neste modelo, todos os átomos da sequência são considerados, inclusive as moléculas de solvente necessárias ao processo de enovelamento (NARLOCH; PARPINELLI, 2017b). Mas devido ao custo de se utilizar todos os átomos, é comum encontrar representações mais simplificadas, que buscam abstrair alguns conceitos para reduzir a complexidade do problema. Existem 2 representações reduzidas que são mais comuns na literatura. Uma delas representa os átomos de cada aminoácido por sua posição Cartesiana (x,y,z). A outra, representa a estrutura 3D através dos ângulos de

torção, ou seja, a tupla com os ângulos ϕ, ψ e ω, podendo ou não incluir os ângulos da cadeia lateral R. A vantagem do segundo modelo é ter os graus de liberdade reduzidos, o que reduz a complexidade, porém uma pequena alteração em um dos ângulos pode causar um grande impacto na conformação, ao contrário das representações Cartesianas que neste caso sofreriam pouco impacto (CORREA; INOSTROZA-PONTA; DORN, 2017).

Existem ainda outros modelos mais simplificados, como os modelos *on-lattice* e *off-lattice AB*. Os modelos *on-lattice* são mais simplificados, onde os aminoácidos são representados por pontos em vértices de uma malha. O modelo mais comum é o modelo HP (Hidrofóbico-Hidrofílico) (BOIANI, 2019). Neste modelo cada aminoácido pode ser representado por apenas uma letra, sendo H ou P, onde H representa os aminoácidos hidrofóbicos e P representa os aminoácidos polares ou hidrofílicos. Eles podem ser representados em 2 dimensões (2D, quadrática) ou 3 dimensões (3D, cúbica) (GABRIEL; MELO; DELBEM, 2012). Já os modelos *off-lattice AB*, sua representação está fora da malha, logo os ângulos tem maior liberdade para girar, e neste caso o A representa os aminoácidos hidrofóbicos e B os aminoácidos hidrofílicos (STILLINGER; HEAD-GORDON; HIRSHFELD, 1993). Neste trabalho os polipeptídeos serão representados pelos ângulos diédricos, o que reduz a complexidade do modelo *all-atom* mas mantém as características necessárias para a predição tridimensional.

2.2.2 Função de Energia

De acordo com o postulado de (ANFINSEN, 1973), que diz que a conformação nativa de uma proteína está bem próxima de sua energia potencial mínima, a função de energia tem uma grande importância dentro do PSP. Ela tem a função de estimar o quanto um modelo está próximo da sua conformação nativa. Porém, isso não é tão simples, a começar pela existência de proteínas que, como já comentado anteriormente, precisam ser desordenadas para apresentarem sua funcionalidade. Além disso, analisando o problema do PSP como um problema multimodal, onde se tem diversas soluções estruturais possíveis apontando para diversos mínimos e máximos locais, a possibilidade de encontrar pontos ótimos distantes da estrutura nativa é uma realidade. Toda essa complexidade envolvida no processo de dobramento dificulta a criação de uma função de energia que represente a estrutura tridimensional de fato. Em vista disto, uma função de energia pode envolver diversos termos, como por exemplo, comprimentos de ligações, ângulos de ligações, ângulos de torção proibidos, valores de ângulos de torção, forças de atração e repulsão de van der Waals, interações eletrostáticas, ligações de hidrogênio entre outras (CORREA; DORN, 2020).

As funções de energia podem ser divididas em 2 grupos. As funções baseadas em física e as baseadas em conhecimento, dependendo da abordagem usada para a modelagem 3D. O primeiro grupo visa modelar os campos de força que determinam as conformações das proteínas. São cálculos baseados na física e mecânica quântica. Neste grupo se encaixam as funções AMBER (SALOMON-FERRER; CASE; WALKER, 2013), CHARMM (BROOKS et al., 2009), GROMOS (EICHENBERGER et al., 2011). No segundo grupo, as funções são deduzidas de

forma empírica a partir das estruturas já resolvidas no PDB. O objetivo é determinar o potencial de contato entre pares de resíduos, tanto pela interação entre os átomos quanto pela propensão da estrutura secundária. A função de energia mais conhecida nessa categoria é o Rosetta, que será apresentado em detalhes mais adiante. Como este tipo de função se baseia em estruturas já conhecidas, ela tem a vantagem de poder modelar o que a física ainda não compreendeu, porém não podem prever o que ainda não foi registrado em banco de dados (ALMEIDA, 2016). Neste trabalho, as funções de energia `score0`, `score3` e `scorefxn` do Rosetta são utilizadas. A função `score0` usa apenas as forças de *Van der Waals* e deve ser utilizada com a inserção de fragmentos. A função `score3` é a mais utilizada e usa os termos comuns de pontuação de centroide, assim como a `score0`. A função `scorefxn` utiliza a representação completa da proteína (SILVA, 2019).

2.2.3 Métodos de Busca

Por último, mas não menos importante, é necessário um método de busca para se trabalhar com a abordagem *ab initio* ou *de novo*. Tal método deve ser capaz de procurar uma conformação de menor energia potencial, em meio a milhares de conformações em um espaço de busca gigantesco. Sendo o PSP um problema NP-completo, sabe-se que a utilização de métodos exatos não é adequada. Isso é devido a complexidade e ao espaço de busca muito grande que tornam estes métodos inviáveis pelo tempo de execução não-polinomial. Para problemas da classe NP-completo como o PSP, é mais aconselhável a utilização de métodos heurísticos, ou seja, métodos não exatos, capazes de obter diversas soluções aproximadas em um tempo de execução aceitável. Nos últimos anos, diversos algoritmos de otimização tem sido propostos para o PSP, dentre eles, temos os clássicos como Dinâmica Molecular e Monte Carlo, e também as meta-heurísticas como os Algoritmos Genéticos (AG) e os Algoritmos Evolutivos (AE).

Talvez um dos métodos mais comuns e mais utilizados sejam os métodos de Monte Carlo. Podem ser utilizados para quase todo tipo de problema de otimização, cujo objetivo é a melhoria de trajetória única baseado em amostragem aleatória (SILVA, 2019). Simulação de Dinâmica Molecular (DM) também é uma técnica bastante utilizada no estudo de macromoléculas biológicas. É baseada na Mecânica Molecular (MM) e fornece informações sobre o comportamento dinâmico dos átomos de um sistema através da integração numérica das equações de movimento (NAMBA; SILVA; SILVA, 2008). Os métodos de Monte Carlo e Dinâmica Molecular, comumente são utilizados em conjunto com outros algoritmos de otimização que visam explorar o espaço conformacional (ALMEIDA, 2016). Outra classe que tem sido bastante explorada no PSP, são os Algoritmos Genéticos (ROCHA et al., 2015; DORN; BURIOL; LAMB, 2011), Algoritmos Evolutivos (CORREA; INOSTROZA-PONTA; DORN, 2017), Algoritmos de Otimização por Enxame de Partículas (PSO) (CORREA; DORN, 2018), entre outras meta-heurísticas que tem sido bem sucedidas neste tipo de problema. Em geral, são algoritmos baseados em população, que tentam melhorar as aproximações de Monte Carlo através dos ciclos evolutivos (DORN et al., 2014).

Neste sentido, ganham destaque os algoritmos de Evolução Diferencial (DE), pois tem

sido amplamente utilizados para o problema do PSP. O DE é um Algoritmo Evolutivo bastante utilizado para otimização contínua por ser fácil de entender e implementar, eficiente, com poucos parâmetros e de baixa complexidade, permitindo aplicação em problemas com alta dimensionalidade (DAS; SUGANTHAN, 2011). Existem diversas adaptações na literatura, entre elas: jDE (BREST et al., 2006), SADE (QIN; SUGANTHAN, 2005), SHADE (TANABE; FUKUNAGA, 2013), entre outras. Basicamente, elas se distinguem entre si pela forma como é feito o ajuste de parâmetros. Em (NARLOCH; PARPINELLI, 2017a), foi utilizado um DE com auto ajuste de parâmetros, combinado a uma estratégia de controle de diversidade e informação da estrutura secundária para resolver o PSP. Em (DENG et al., 2019) é proposto um novo algoritmo de evolução diferencial baseado em especiação dinâmica e duas estratégias de mutação, chamado de DSM-DE. No DSM-DE, os autores utilizam uma técnica de agrupamento em espécies para orientar as estratégias de mutação, gerando 2 vetores de mutação baseado em diferenças, um com foco em exploração e outro com foco em aproveitamento. Ele tem grande potencial em acelerar a convergência do algoritmo sem perder a diversidade, o que o torna bastante atraente para aplicação no PSP.

2.2.3.1 Dynamic Speciation-based Mutation Differential Evolution (DSM-DE)

DSM-DE é um algoritmo de evolução diferencial que usa especiação dinâmica e duas estratégias de mutação para resolver problemas de otimização de objetivo único. O principal objetivo deste algoritmo em utilizar duas estratégias de mutação é acelerar o processo de convergência sem perder a diversidade. Para isso, ele utiliza dois vetores de mutação simultaneamente que são baseados nas sementes das espécies, onde cada semente representa o melhor indivíduo de cada espécie. A primeira estratégia é chamada de "DE/seed-to-seed", ou seja, "semente para semente", e foi projetada para acelerar a convergência do algoritmo, o que levará a população ao ótimo global rapidamente. Neste caso, o vetor de mutação é construído a partir de 3 vetores pais, sendo eles a semente da espécie atual, a semente de outra espécie selecionada aleatoriamente e um indivíduo aleatorio da espécie atual (DENG et al., 2019).

A segunda estratégia é chamada de "DE/seed-to-rand", e foi projetada para diversificar a população e ajudar a população a sair do ideal local (no PSP, o ideal é o mínimo local), expandindo os intervalos de pesquisa em comparação com a primeira. Neste caso, ao invés de considerar 2 sementes para a construção do vetor de mutação, é considerado apenas a semente atual e os outros 2 indivíduos são aleatórios de toda a população. A cada geração, as duas mutações geram 2 filhos, que competem entre si e o que tiver melhor aptidão é selecionado para competir com o vetor alvo. Porém, as duas estratégias de mutação dependem muito da estrutura populacional resultante da Técnica de Especiação Dinâmica (DST), que é uma técnica para dividir a população em espécies. A ideia é que essa técnica possa localizar vários grupos de indivíduos potencialmente úteis em diferentes regiões através da distância euclidiana entre esses indivíduos. Segundo os autores, ele foi projetado para classificar os indivíduos a cada iteração, ou seja, a cada nova população. Acredita-se que esta técnica possa ajudar o DE a encontrar a

solução ideal global real entre inúmeras soluções ótimas locais (DENG et al., 2019) (no PSP, uma solução ótima local é um mínimo local). O Algoritmo 1 apresenta o pseudocódigo para o DST, onde *S_num* é o número de espécies, *C_min* e *C_max* são o tamanho mínimo e máximo que uma espécie pode ter.

P, Uma população contendo *NP* indivíduos no espaço de busca;
Di, Uma matriz *NP* * *NP* armazenando a Distância Euclidiana entre cada indivíduo no conjunto *P*;
Set $S\_num = 0$ e $m = 0$;
Set $C\_min = 3$ e $C\_max = NP/5$;
S, Uma célula que contém *S_num* elementos;
while *população P não está vazia* **do**
    $S\_num = S\_num + 1$;
    Encontre o indivíduo mais apto $\vec{X}_{seedi}$ na população *P* como a semente da espécie S_num_{th} e remova ela de *P* para *S*;
    Calcule o tamanho da espécie S_num_{th} como em (1);
    Incorpore $m(i) - 1$ indivíduos restantes na população *P* para *S*, que estão mais próximos de $\vec{X}_{seedi}$ segundo a Distância Euclidiana, com $\vec{X}_{seedi}$ para formar a i_{th} espécie e eliminar esses indivíduos selecionados da população *P*;
end

Algoritmo 1: Pseudocódigo da Técnica de Especiação Dinâmica

O DST é usado para classificar a nova população gerada no início de cada iteração. No DST, as espécies podem ter tamanhos diferentes e isso é calculado a cada etapa. Então primeiro, são calculados os valores de aptidão de todos os indivíduos e a distância euclidiana entre cada par de indivíduos da população. Seleciona-se o indivíduo com o melhor condicionamento físico da população. Esse será a semente da espécie. Em seguida, calcula-se o tamanho da espécie com base na porcentagem do valor de aptidão da semente, usando a equação (1). O próximo passo é selecionar os indivíduos $m(i) - 1$ mais próximos da semente, de acordo com a distância euclidiana. Cada indivíduo selecionado é removido da população e adicionado à espécie. Esse processo se repete enquanto houver indivíduos na população. Depois de dividir a população em espécies, o algoritmo segue o ciclo evolutivo de Evolução Diferencial. A cada nova geração, o processo de especiação se repete. A Figura 12 demonstra o funcionamento do DST.

$$m(i) = round\left\{\frac{fit(i) - maxFit}{minFit - maxFit} * (C\_max - C\_min) + C\_min\right\} \quad (1)$$

A equação 1 mostra como é feito o cálculo para o tamanho de cada espécie, note que o tamanho de cada espécie é calculado levando em consideração o valor de aptidão da semente da espécie atual. Na equação, $fit(i)$ é a aptidão da semente da espécie atual, *maxFit* e *minFit* são a pior e melhor aptidão da população em geral, *C_max* representa o tamanho máximo da espécie e *C_min* é o tamanho mínimo da espécie. Vale destacar que o algoritmo proposto neste trabalho segue o DSM-DE, porém com apenas uma estratégia de mutação e alguns ajustes nos parâmetros de controle, adaptados para o problema em questão.

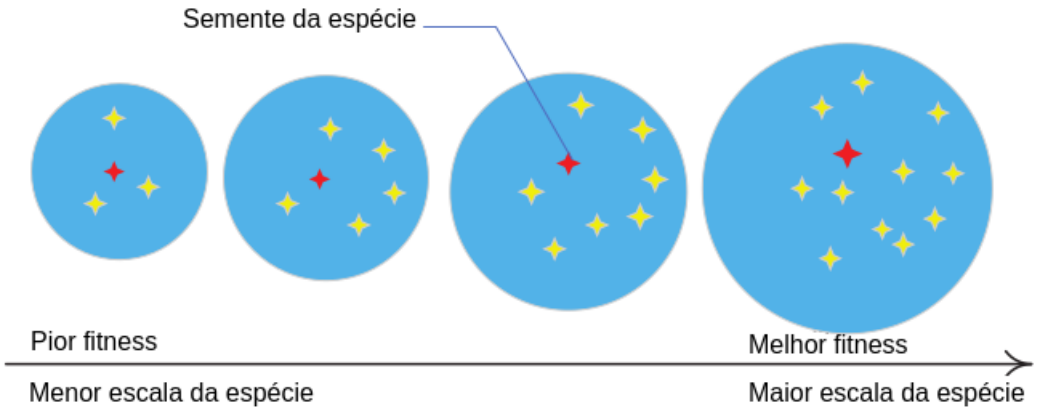


Figura 12 – Mapa de esboço da divisão de especiação dinâmica. Fonte: Adaptado de (DENG et al., 2019)

2.3 MONTAGEM DE FRAGMENTOS

Resumidamente, pode-se dizer que fragmentos são pedaços de proteínas, ou seja, sequências curtas de aminoácidos. O tamanho do fragmento define quantos aminoácidos cabem na sequência. Na literatura é possível encontrar tamanhos de 3 (HAO et al., 2016) resíduos, 5 (DORN; SOUZA, 2008), 7 (DORN; SOUZA, 2010), 9 (GARZA-FABRE et al., 2016), 11 (DORN; SOUZA, 2010) e 15 (LEE; KIM; LEE, 2005) resíduos de aminoácidos. Fragmentos de tamanho maior podem causar grande impacto na conformação da proteína e por isso representam uma etapa de exploração, enquanto que fragmentos de tamanho menor tem menos impacto na conformação e representam uma etapa de refinamento (ZHANG et al., 2019).

A técnica de montagem de fragmentos consiste em dividir a sequência da proteína alvo em fragmentos contínuos, e encontrar estes fragmentos em proteínas já conhecidas que possam substituir os fragmentos da proteína alvo, de forma a trazer informação que auxilie em sua conformação. De maneira geral, os fragmentos trazem informação dos ângulos de torção ϕ (phi) e ψ (psi) e a estrutura secundária para cada aminoácido. A Figura 13 ilustra o processo de fragmentação da proteína alvo. No exemplo é utilizado um tamanho de fragmento de 9 resíduos.

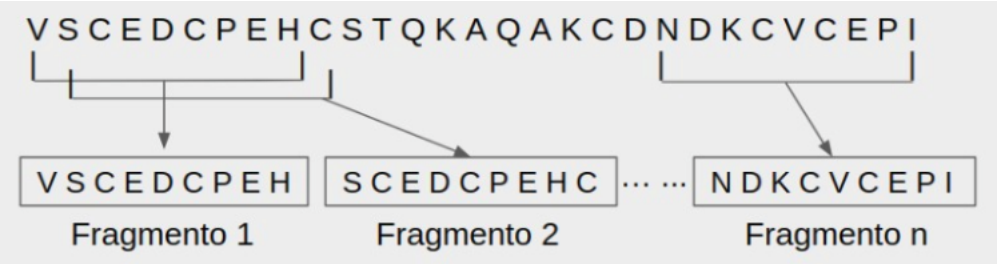


Figura 13 – Geração de fragmentos da proteína alvo. Adaptado de (DORN; SOUZA, 2008)

Geralmente o que se faz é criar uma biblioteca com diversos fragmentos que possam ser utilizados ao longo do processo de otimização. Para isso, utiliza-se banco de dados de

proteínas já conhecidas, como o Protein Data Bank (PDB)[3], ou servidores, como o PISCES (Protein Sequence Culling Server) (WANG; DUNBRACK ROLAND L., 2003), para identificar fragmentos que possam ser substituídos. Não cabe aqui descrever o passo a passo para a geração destas bibliotecas, uma vez que já existem diversas ferramentas que auxiliam neste processo, ficando de certa forma transparente ao usuário, como por exemplo, a ferramenta Rosetta, cujo processo é explicado em detalhes por (GRONT et al., 2011). Porém, em trabalhos anteriores, como em (DORN; SOUZA, 2008) os autores criam suas próprias bibliotecas de fragmentos com todas as suas etapas.

A aplicação de fragmentos de proteína dentro do PSP se resume a causar perturbações na conformação através da inserção de pequenos fragmentos. Segundo (DHINGRA et al., 2020), os métodos que utilizam fragmentos são os mais bem sucedidos da literatura com relação a predição terciária *ab initio*. De maneira geral, a estratégia de fragmentos é utilizada com o objetivo de gerar conformações mais estáveis (HAO et al., 2016) e com isso melhor explorar o espaço conformacional da proteína sem perder as informações da estrutura principal (ZHANG et al., 2016). Neste trabalho os fragmentos são utilizados para gerar a população inicial e na fase de mutação, nesta última tem o objetivo de promover diversidade na população.

2.3.1 Suite Rosetta

Como visto anteriormente, trabalhar com os métodos FM, envolve certa complexidade, como por exemplo, a definição de uma função de energia, a qual envolve diversos cálculos. Neste sentido, existem algumas ferramentas que fornecem esse tipo de cálculo, facilitando o trabalho dos pesquisadores. A Suite Rosetta é um pacote de software gratuito para uso acadêmico e oferece várias funções para trabalhar com proteínas e macromoléculas. Entre elas estão as funções `score0`, `score3` e `scorefxn`, que são utilizadas neste trabalhos. A função `score0` é usada para gerar a população inicial, e a função `score3` é usada durante o ciclo evolutivo. Ambos consideram apenas os átomos do centroide para calcular a energia. Lembrando que a representação por centroide não considera os átomos da cadeia lateral. A função `scorefxn` é usada no final de todo o processo para produzir uma representação completa da proteína (SILVA, 2019).

Um recurso essencial fornecido pelo Rosetta e usado neste trabalho é a inserção de fragmentos. Essa técnica consiste em dividir a sequência da proteína alvo em fragmentos contínuos e encontrar esses fragmentos em proteínas já conhecidas que podem substituir os fragmentos da proteína alvo. Rosetta trabalha com 2 tamanhos de fragmento, 3 e 9 (GRONT et al., 2011), e fornece 2 operadores de inserção de fragmento: clássico e suave. O operador clássico funciona como uma busca global, onde uma pequena mudança pode impactar significativamente a conformação. Por outro lado, o operador suave funciona como uma busca local, causando menos impacto na conformação. Em geral, esses fragmentos devem compor uma biblioteca

[3] Protein Data Bank disponível em <https://www.rcsb.org/>

de fragmentos que pode ser utilizada em todo o processo de otimização. Os fragmentos são coletados de um banco de dados e selecionados de acordo com uma função de pontuação. Além disso, o método Monte Carlo do Rosetta é utilizado durante a inserção de fragmentos, tanto na população inicial quanto durante o ciclo evolutivo.

Rosetta também fornece 2 protocolos para gerar uma biblioteca de fragmentos: Best e Quota (GRONT et al., 2011). O primeiro considera apenas um preditor de ES na composição da função de pontuação. Isso significa que o selecionador de fragmentos priorizará por fragmentos com a mesma ES vencedora no preditor. Em outras palavras, se o preditor em questão previu que um determinado aminoácido tem 51% de chance de ser uma hélice, este protocolo priorizará fragmentos que possuem uma hélice para esse mesmo aminoácido, mesmo que haja uma chance de que seja 49% folha. Se o preditor estiver correto em sua previsão, isso pode ser muito bom. Caso contrário, a previsão da estrutura terciária pode nunca encontrar uma folha nesse trecho da sequência. Já o protocolo Quota tem a função de inserir mais diversidade nos fragmentos e tentar evitar que este efeito "saturado" do protocolo Best ocorra (GRONT et al., 2011). Portanto, considera o uso de 3 preditores de ES para auxiliar na função de pontuação. Nesse caso, o seletor trará os elementos de ES de acordo com as taxas de previsão para os 3 tipos, hélice, folha e voltas. O protocolo Quota também permite definir a participação de cada preditor.

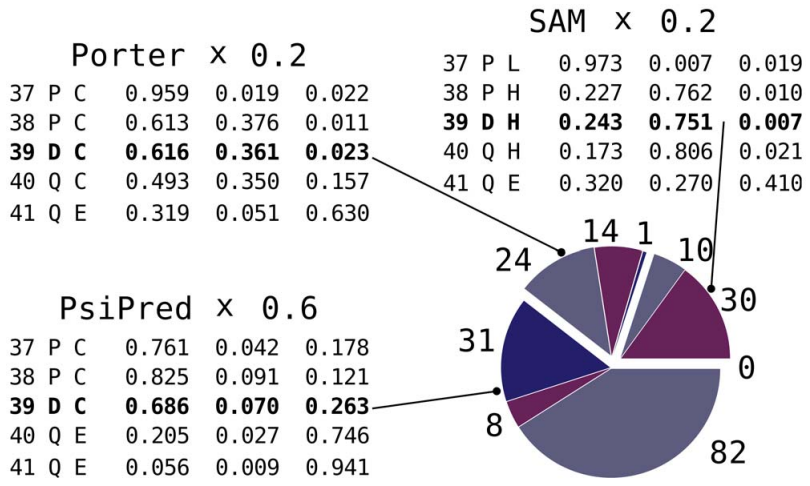


Figura 14 – Exemplo de atuação do protocolo Quota. Fonte: (GRONT et al., 2011)

A Figura 14 mostra o funcionamento do protocolo Quota. Na figura, é mostrado um gráfico de pizza e o arquivo de saída com a previsão de ES de cada um dos 3 preditores. No arquivo com as ES, a primeira coluna representa a posição do aminoácido na sequência; a segunda coluna é o aminoácido em sua representação de uma letra; a terceira coluna é a ES prevista para o aminoácido; a quarta até a sexta coluna é a probabilidade deste aminoácido ser do tipo C, H ou E, respectivamente. O gráfico de pizza representa o aminoácido D da posição 39 e a proporção de elementos de ES de acordo com cada preditor. No gráfico, cada cor representa um tipo de ES, sendo C em cinza, H em roxo e E em azul. No exemplo, foi definido 60% de participação para PSIPRED, e 20% para cada um dos demais. Assim, 60% dos fragmentos selecionados serão de acordo com a previsão do PSIPRED. Estes 60% ainda são divididos entre

os 3 tipos de ES, de acordo com as respectivas previsões. Claramente a imagem reflete o objetivo do protocolo, que é trazer diversidade aos fragmentos.

Outra função essencial para este trabalho e que também é fornecida pelo Rosetta, é o DSSP (*Define Secondary Structure of Proteins*). Esta função fornece as informações de ES de uma conformação. Ela é utilizada durante todo o ciclo evolutivo, toda vez que uma conformação é alterada. Pois é com base nesta informação que ocorre o procedimento de cruzamento entre os indivíduos. A próxima seção explica com maiores detalhes a respeito da predição de ES.

2.4 PREDIÇÃO DE ESTRUTURA SECUNDÁRIA DE PROTEÍNA

A estrutura secundária (ES) refere-se aos arranjos regulares, como conformações locais que ocorrem frequentemente em uma cadeia de aminoácidos, formando pequenos grupos de conformações. Essas estruturas são mantidas por ligações de hidrogênio que ocorrem entre os grupos Amina e Carboxila dos aminoácidos. Como já visto na Seção 2.1.1, segundo o DSSP existem 8 tipos de ES, as quais podemos chamar de 8-classes ou Q8, sendo elas: H (α-*helix*), G (3^{10}-*helix*), I (π-*helix*), E (β-*strand*), B (β-*sheet*), T (alça), S (volta) e C (outros). Porém, na literatura é mais comum o uso de 3-classes ou Q3 ao invés de Q8, e são elas: hélices (H), folhas (E) e voltas (C). Existem diversas formas de converter a notação Q8 para Q3. A Tabela 2 mostra 5 diferentes formas de conversão, note que as classificações podem ser bem diferentes, e isto pode impactar em certo ponto a performance da predição (SMOLARCZYK et al., 2020).

Tabela 2 – 5 métodos para transformação de 8-classes para 3-classes de ES. Fonte: Adaptado de (SMOLARCZYK et al., 2020)

Classe	Método 1 (CASP)	Método 2	Método 3	Método 4	Método 5
H	H, G	H	H, G, I	H, G	H, G, I
E	E, B	E	E, B	E	E
C	S, T, I, C	G, S, T, B, I, C	S, T, C	S, T, B, I, C	S, T, B, C

Para utilizar informações de ES, se faz necessário o uso de preditores de ES. Em (SMOLARCZYK et al., 2020) os autores fazem uma revisão sobre predição de ES e trazem uma lista com os preditores de ES mais atuais, onde eles comparam a precisão de diversos preditores online através de diversos experimentos. Existem vários preditores de ES na literatura, mas é mais fácil encontrar preditores que utilizam a notação Q3 do que a Q8. Alguns até fornecem os resultados encontrados para ambas notações. Um dos preditores mais utilizados é o PSIPRED (BUCHAN; JONES, 2019), também é um dos mais antigos, desde 1999. O PSIPRED utiliza apenas a notação de 3-classes, mas é bastante utilizado em conjunto com o Rosetta, pois já fornece a entrada no formato necessário para uso no gerador de fragmentos do Rosetta. Outros preditores também se destacam, como o MUFOLD-SS (FANG; SHANG; XU, 2018), SPIDER2 (HEFFERNAN et al., 2015), SPIDER3 (HEFFERNAN et al., 2015), Porter 4.0 (MIRABELLO;

POLLASTRI, 2013) e Porter 5 (TORRISI; KALEEL; POLLASTRI, 2018), além do RaptorX (KäLLBERG et al., 2012), SSpro8 (POLLASTRI et al., 2002), PROMOTIF, PREDICT.

Dentre as características que mais pesam na escolha pelo preditor de ES ideal, está o tipo de informação que o mesmo exibe como resultado. Alguns preditores informam apenas a ES prevista para cada aminoácido da sequência, outros informam os ângulos do backbone (ϕ, ψ e ω), ainda outros informam a área acessível ao solvente e outros a probabilidade de ocorrência para cada tipo de ES, como é o caso do PSIPRED, SPIDER, MUFOLD e RaptorX. A Figura 15 exibe uma saída para o PSIPRED, por exemplo, para o aminoácido S da posição 11 significa que o preditor previu 13% de probabilidade de ocorrência da estrutura to tipo C (voltas, alças ou bobinas), 86% de probabilidade para estrutura do tipo H (hélice) e 1% de probabilidade para estrutura E (folha). Em resumo, quanto mais informações o preditor fornecer, mais útil será na previsão de estrutura terciária. Alguns preditores utilizam notações diferentes para as estruturas do tipo T (volta), S (alças) e C (bobinas). O mais comum é agrupar os 3 tipos sob uma notação C ou L (de *loop*, em inglês).

```
10 R H   0.064  0.939  0.006
11 S H   0.131  0.859  0.015
12 N C   0.678  0.270  0.011
13 F C   0.490  0.405  0.050
14 N C   0.435  0.471  0.055
15 V C   0.485  0.435  0.042
16 C C   0.545  0.394  0.041
```

Figura 15 – Exemplo de saída do preditor PSIPRED.

Com relação a previsão de estrutura terciária, a previsão de ES pode trazer informações importantes com relação aos relacionamentos entre os aminoácidos e como essas estruturas se formam. Além disso, estes preditores são essenciais na previsão *ab initio*, sendo insumo para a geração de fragmentos. É com base nos resultados emitidos pelos preditores de ES que o gerador de fragmentos gera os arquivos de pontuação, que serão utilizados na seleção dos fragmentos que irão compor as bibliotecas. Porém, vale notar que definições rigorosas como o DSSP não existem de fato na natureza, uma vez que não existem hélices e folhas ideais, como também não há limites claros entre uma estrutura e outra, nem quando uma começa e outra termina. Por isso, estima-se que o limite de precisão da previsão de ES seja em torno de 88%, considerando a classificação de 3 classes (SMOLARCZYK et al., 2020). Logo, é possível concluir que este tipo de informação pode sim auxiliar na conformação da estrutura 3D, mas não pode ser um fator limitante.

De maneira geral, a utilização de ES visa melhorar a exploração do espaço de busca conformacional, bem como melhorar a estrutura das conformações previstas, uma vez que um baixo valor de energia nem sempre está relacionado a uma boa estrutura, mas uma boa estrutura de proteína está relacionada a uma baixa energia. Conhecendo a ES é possível controlar onde as perturbações na conformação devem ocorrer, ou seja, preservando ou alterando estruturas específicas. Segundo (GARZA-FABRE et al., 2016) alterações nas regiões de voltas, alças e

bobinas, podem ter forte relação com o arranjo tridimensional da proteína. Além disso, a previsão de ES aliada a informação do DSSP que é fornecida pelo Rosetta, serve para saber o quanto uma conformação está próximo de sua estrutura prevista. Neste trabalho, a informação dos preditores de ES, é utilizada em dois momentos: (i) para gerar a biblioteca de fragmentos, e neste caso, para a utilização com o protocolo Quota é necessário que os preditores informem a probabilidade de ocorrência para cada tipo de ES; (ii) para avaliar a estrutura da conformação, aqui a informação do DSSP, que é a estrutura da conformação atual, é então comparada com a previsão do preditor de ES.

2.5 MAPAS DE CONTATO

Mapas de contato são representações bidimensionais de estruturas tridimensionais de proteínas. Em outras palavras, um mapa de contato é uma matriz que representa os contatos resíduo-resíduo de uma proteína, fornecendo informações a respeito de sua forma tridimensional. Ou mesmo, que se trata de uma "impressão digital"estrutural da proteína, e que a partir dela é possível identificar uma proteína e sua forma 3D (EMERSON; AMALA, 2017). Ou seja, os contatos identificados no mapa fornecem informações importantes a respeito de sua conformação, desde as estruturas secundárias até sua forma terciária.

Os mapas de contato nada mais são do que matrizes de duas dimensões onde as células representam a distância entre pares de resíduos de proteínas. Para gerar um mapa de contato, o processo é simples. Dada uma matriz quadrada, onde as linhas e colunas são a quantidade de aminoácidos de uma proteína, calcula-se a distância euclidiana entre cada par de resíduos e armazena o valor na célula correspondente. Obviamente a diagonal da matriz irá conter todos os valores zero, uma vez que é a distância entre o próprio aminoácido. Feito isso, define-se um valor de corte, geralmente 8Å. As células que possuírem um valor abaixo do valor de corte representam um possível contato entre o par de resíduos e devem assumir valor 1, caso contrário, valor 0. A Figura 16 apresenta a construção de um mapa de contatos. A imagem a esquerda representa uma matriz com as distâncias euclidianas. A imagem a direita sua representação no mapa de contatos, onde as células pretas representam um possível contato entre os resíduos.

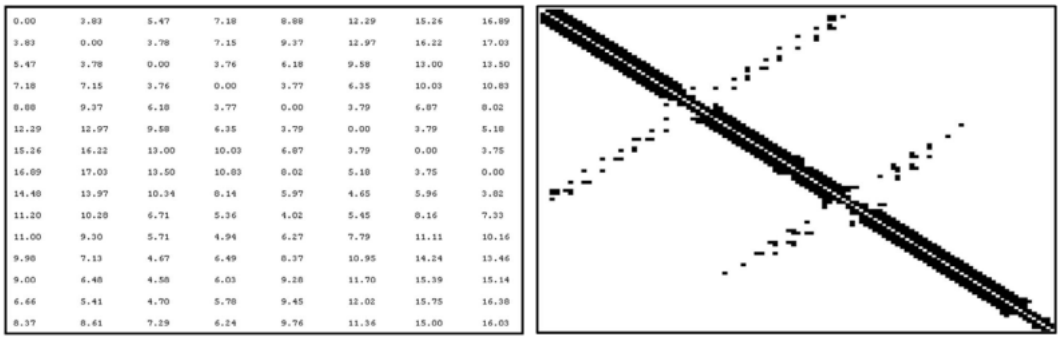


Figura 16 – Construção de mapa de contato. Fonte: (EMERSON; AMALA, 2017)

Segundo (EMERSON; AMALA, 2017), é possível extrair diversas informações de mapas

de contato, como por exemplo, clusters de contatos que representam estruturas secundárias e possíveis estruturas terciárias. Recentemente, o uso de mapas de contato aplicado ao PSP tem se tornado cada vez mais popular. A partir da 10 edição do CASP, abordagens *de novo* combinando o uso de informações de contato mostraram resultados promissores. Segundo (SANTOS et al., 2017), a partir da 11 edição do CASP, as técnicas que aplicaram mapas de contato ao PSP permitiram trabalhar com proteínas muito maiores do que era possível anteriormente.

A previsão dos mapas de contato se dá pela análise de mutações correlacionadas obtidas a partir de múltiplos alinhamentos de sequência de proteínas homólogas (ROCHA et al., 2018). Atualmente existem diversos preditores de mapas de contato, entre eles: CCMpred, GREMLIN, MetaPSICOV e RaptorX. De maneira geral, a informação fornecida pelos preditores de contato se resume aos pares de contatos possíveis e suas probabilidades de contato. No PSP, é comum se utilizar apenas os pares com maior probabilidade de contato. Atualmente, se considera os L, L/2, L/5 e L/10 como os principais contatos, sendo L o tamanho da sequência de aminoácidos (PENG; ZHOU; ZHANG, 2020).

Apesar do sucesso em sua utilização, muitos pesquisadores ainda questionam se as previsões dos mapas são confiáveis, por isso utilizam mais de um mapa em seus métodos. Devido a grande quantidade de pares de contatos, a dificuldade está em saber quais pares realmente devem ser considerados. Em (PENG; ZHOU; ZHANG, 2020) os autores reclamam a falta de informação de distância, uma vez que a informação é binária (ou seja, a distância é menor que 8Å, ou não), pois mesmo sabendo que um contato existe, não se tem certeza da distância exata. Neste trabalho, a informação fornecida por mapas de contato é utilizada para calcular o potencial de contato de uma conformação. Este valor, por sua vez, é utilizado apenas durante a etapa de seleção, onde o indivíduo que tiver o maior potencial de contato é selecionado para a nova geração.

3 TRABALHOS RELACIONADOS

Neste capítulo, são apresentados os trabalhos relacionados. Considera-se como trabalho relacionado apenas os trabalhos FM, que utilizam alguma meta-heurística como algoritmo de otimização e que também utilizam informação do problema, ou seja, trabalhos que se encaixam na classificação *de novo*. As técnicas consideradas para incluir informação do problema no PSP são: montagem de fragmentos, estrutura secundária, mapas de contato e baseados em física. Os trabalhos são apresentados em ordem crescente por data de publicação.

Em (LEE; KIM; LEE, 2005) é proposto um método baseado em montagem de fragmentos e otimização global. Utiliza um algoritmo de otimização chamado Recozimento do Espaço Conformacional (*Conformational Space Annealing*) (CSA, sigla em inglês), que emprega Monte Carlo Metropolis e annealing para a busca no espaço conformacional. Neste trabalho são usados fragmentos de tamanho 15 para alterar as conformações. A biblioteca de fragmentos é gerada usando o PROFESY, e as informações de estrutura secundária são geradas pelo preditor PREDICT.

Um modelo híbrido para previsão *ab initio* é proposto por (ZHOU; SKOLNICK, 2007). O modelo usa conceitos do Rosetta e do TASSER (ZHANG; ARAKAKI; SKOLNICK, 2005), que é um método baseado em modelos. A integração entre os 2 conceitos é realizada dobrando pedaços de proteína por meio da técnica de inserção de fragmentos, combinando estes modelos com modelos identificados que tem baixa similaridade. O modelo inclui o uso de informação de estrutura supersecundária. E também cria uma biblioteca de fragmentos de tamanhos 3 e 9 resíduos.

Os fragmentos também são utilizados na proposta de (DORN; SOUZA, 2008), chamada CReF. Nessa proposta não são utilizados fragmentos inteiros, apenas as informações de Phi e Psi do resíduo central. Utiliza técnicas de agrupamento e a previsão de ES para orientar a busca e inserção de fragmentos na conformação. São utilizados 3 preditores de ES: DSC, PHD, PREDATOR. A função de energia utilizada é AMBER. Segundo os autores, este método é muito rápido, e tem o objetivo de gerar conformações iniciais para servir de entrada em métodos de refinamento. Em 2011, (DORN; BURIOL; LAMB, 2011) propos um novo modelo que utiliza o CReF para obter informações estruturais de modelos de proteínas do PDB. O modelo inclui um AG híbrido baseado em uma população estruturada e relink de caminho para escapar de mínimos locais.

Em (KAZMIER et al., 2011) é proposto um algoritmo baseado em Monte Carlo para otimizar os termos de restrição. Em resumo, utilizam informação de estrutura secundária e esposição ao solvente para criar uma métrica de restrição de distância EPR (Ressonância Paramagnética Eletrônica), essas restrições são usadas como penalidade na função de energia do Rosetta. É usado o preditor PSIPRED para previsão da estrutura secundária, e o programa NetSurfP para previsão de exposição ao solvente.

Informação de estrutura secundária também foi utilizado por (WEINER et al., 2013). Os

autores reunem os elementos de estrutura secundária em topologias usando o algoritmo de Monte Carlo. Os métodos de previsão da estrutura secundária, PSIPRED e JUFO, foram combinados para alcançar um consenso de previsão da estrutura secundária de 3-classes, e evitar erros de previsão de apenas 1 preditor. O método é projetado em cima das proteínas de membrana, e por isso as regiões de voltas e alças não são consideradas, os modelos são criados sobre estruturas do tipo hélice e folhas.

Já em (GARZA-FABRE et al., 2016), ES e fragmentos começam a ser utilizados em conjunto. Neste método é proposto um algoritmo memético baseado em um AG e na inserção de fragmentos, com o objetivo de tentar melhorar a exploração de dobras de proteínas e alcançar energias mais baixas. Para isso, o algoritmo proposto passa por um processo de seleção de pais, onde todos os indivíduos da população são selecionados apenas 1 vez, o que preserva a diversidade. Além disso os autores utilizam operadores genéticos especializados que se concentram nas regiões de voltas e alças, incentivando a exploração do espaço conformacional. Um algoritmo memético também é proposto em (CORREA; DORN, 2020). Neste modelo, a população é organizada em uma estrutura em árvore. O algoritmo memético age como otimizador principal, enquanto que um algoritmo de Colônia Artificial de Abelhas age como busca local em cada nodo da árvore. Além disso, o método utiliza informação de fragmentos e estrutura secundária.

Um novo método para melhorar a exploração do espaço de busca conformacional é proposto em (HAO et al., 2016). Os autores utilizam uma técnica chamada *Abstract Convex Underestimation* (ACUE, sigla em inglês) que converte o espaço conformacional original de alta dimensionalidade em espaço de recursos, cuja dimensão é reduzida pela técnica de abstração de recursos. A proposta é feita com base na estrutura de algoritmo evolutivo, usando a técnica ACUE, inserção de fragmentos e Monte Carlo. Foram utilizados fragmentos de tamanho 3 e 9 do Rosetta na geração da população inicial, e para a geração da biblioteca de fragmentos foi usado o servidor PISCES (ref). Os resultados dos testes demonstraram que este método pode ser mais rápido e eficaz. Posteriormente, os autores propuseram uma adaptação ao modelo, com a inclusão de uma técnica chamada *Lipschitz Underestimation* (LUE, sigla em inglês) (HAO; ZHANG; ZHOU, 2018). Com o objetivo de excluir antecipadamente as regiões e conformações inválidas, e com isso economizar os tempos de avaliação conduzindo a exploração a uma área com mais potencial.

Ainda na linha de melhorar a exploração do espaço de busca, um DE é proposto em (ZHANG et al., 2016) chamado *Replica Exchange based Differential Evolution* (REDE). Os autores utilizam a técnica de montagem de fragmentos na geração da população inicial, para melhorar a exploração do espaço conformacional e evitar o efeito da entropia na pesquisa. O Rosetta é utilizado tanto pela função de energia, quanto pela montagem de fragmentos. A biblioteca de fragmentos é gerada usando o servidor PISCES. Segundo os autores, a estratégia de *Replica Exchange* aumenta a diversidade da população e a capacidade do algoritmo em ultrapassar o mínimo local, aumentando assim a eficiência da busca.

Cálculos de probabilidade são utilizados em (HAO; ZHANG, 2017), em um algoritmo

chamado de distribuição por estimativa dupla (DED, sigla em inglês). As trajetórias de MC são lançadas simultaneamente, e duas distribuições de probabilidade são projetadas. Uma é chamada de probabilidade atual que é calculada com base na energia da última conformação em cada trajetória. A outra é chamada de probabilidade de aceitação histórica, que é calculada com base em quantas vezes uma conformação é aceita em cada trajetória. A ideia disso é que conformações com menor energia e mais vezes aceitas tem maior chance de serem selecionadas. Utiliza fragmentos para atualizar as conformações e a função de energia `score3` do Rosetta.

Outro trabalho que utiliza probabilidade para orientar os operadores de seleção é (ZHANG et al., 2017). Que propõem uma evolução diferencial guiada por perfil de distância, chamado de DPDE (sigla em inglês), onde uma estratégia de seleção é usada com o objetivo de orientar a amostragem no espaço conformacional. Ou seja, no processo de seleção do DE, além da energia, também é considerado um perfil de distância para selecionar as conformações. Isso é feito calculando-se uma probabilidade de aceitação da distância resíduo-resíduo dos aminoácidos. O objetivo em fazer isto é manter as conformações com energias mais altas, mas estruturas mais razoáveis. Segundo os autores, essa técnica aumenta a capacidade de escapar de mínimos locais e a eficiência da pesquisa é aprimorada. É utilizada a função de energia `score3` do Rosetta, bem como os fragmentos durante a mutação e o crossover.

Em (SANTOS et al., 2017) é proposto um algoritmo genético com aglomeração fenotípica (GAPF) em uma abordagem FM. O objetivo dos autores é validar a utilização de mapas de contato para o PSP. Para isso, são gerados 2 mapas de contato: um mapa de contato nativo obtido da estrutura experimental e outro mapa de contato filtrado com apenas os contatos nativos presentes em um mapa previsto. Estes termos são incorporados na função de adequação do algoritmo. Os resultados confirmam uma importante melhoria na precisão das previsões.

Uma estratégia baseada em inteligência de enxames é proposta em (CORREA; DORN, 2018), onde são desenvolvidas duas versões do algoritmo colônia artificial de abelhas (Artificial Bee Colony) (ABC, sigla em inglês). A primeira é uma variação do ABC padrão, a segunda é uma versão modificada chamada de Mod-ABC. A segunda versão foca em realizar atualizações apenas nas regiões de estrutura secundária irregular, ou seja, voltas e bobinas. O objetivo é aumentar a diversidade e melhorar a exploração do espaço de busca. Além disso, é proposto uma função de energia que soma 3 termos: um termo baseado na função de energia do Rosetta, outro termo baseado na área de superfície acessível ao solvente e um termo baseado na estrutura secundária, este último tem a ideia de melhorar a formação de estruturas secundárias corretas.

Em (SILVA, 2019), é proposto uma algoritmo evolutivo baseado na evolução diferencial SaDE (Self Adaptive Differential Evolution), com foco na auto adaptação dos parâmetros do DE, porém sem utilizar a parte das diferenças no operador mutação. O método proposto utiliza a inserção de fragmentos para promover diversidade e melhorar a busca. Também é utilizado o preditor PSIPRED para a geração da biblioteca de fragmentos.

Uma nova função de energia é proposta por (MISHRA; HOQUE, 2019), chamada de 3DIGARS. O preditor de ES SPIDER2 é utilizado para fornecer informações da estrutura

secundária da proteína, como ângulos de torção (PHI e PSI) e área acessível ao solvente (ASA). Estas informações são utilizadas para construir a função de energia (chamada 3DIGARS), que é formada pela informação de ASA, ângulos de torção e propriedades hidrofóbico-hidrofílico dos aminoácidos. Como algoritmo de otimização é utilizado um Algoritmo Genético (chamado KGA). Além da função de energia, as informações de ES também são utilizadas no processo de crossover. Os pontos de crossover são selecionados de acordo com a ES, onde são preservadas as estruturas do tipo folha (E e B), estas não são consideradas como pontos de cruzamento. O objetivo disso é preservar as regiões de estrutura tipo folha durante a operação de cruzamento e realizar alterações mais cuidadosas nessas regiões durante a operação de mutação.

Outro algoritmo que promove alteração de acordo com a ES é proposto em (ZHANG et al., 2019), e é chamado de algoritmo baseado em população guiado por entropia de informação (PAIE, sigla em inglês). É construído um algoritmo populacional dividido em 2 estágios, de *exploration* e *exploitation*. O primeiro estágio é focado na exploração das conformações no espaço de busca, utilizando fragmentos de tamanho 9 para fazer as perturbações nos indivíduos. Após a fase de exploração, cada conformação passa por uma perturbação do ângulo de torção selecionado aleatoriamente, desde que sua ES seja uma volta ou alça. No estágio 2, são utilizados fragmentos de tamanho 3 para fazer o refinamento dos indivíduos.

Outro trabalho que propõem melhorar a função de energia é de (PENG; ZHOU; ZHANG, 2020). Neste trabalho é proposto um método de previsão *de novo* chamado CoDiFold, onde os contatos previstos e os perfis de distância são usados para melhorar a precisão da função de energia, e a estratégia de multi-mutação é utilizada para aprimorar a eficiência da amostragem. A função de energia utilizada é o `score3` do Rosetta em conjunto com um termo de perfil de distância e um termo de contatos previstos por 2 mapas de contato. O objetivo em acrescentar estes 2 termos a função de energia do Rosetta é aliviar a imprecisão da função e melhorar a correlação energia-RMSD. Eles utilizam 2 mapas de contato para amenizar erros de previsão de um único mapa. Além disso, são usadas 3 estratégias de mutação baseada em fragmentos, com o objetivo de evitar a convergência prematura e equilibrar a intensificação e exploração. Os fragmentos também são usados na população inicial.

Em (ZHANG et al., 2020) é proposto um algoritmo de evolução diferencial que utiliza informação de estrutura secundária e contato resíduo-resíduo, chamada de (SCDE). A proposta inclui duas estratégias de seleção projetadas para orientar a busca. A primeira é baseada na estrutura secundária e tem o objetivo de gerar conformações com estruturas mais ajustadas. A segunda é baseada em contato, visando encontrar estruturas mais razoáveis no espaço de conformação. São usados fragmentos de tamanho 3 e 9 na população inicial e na mutação, e a função de energia `score3` do Rosetta. O PSIPRED é utilizado para prever a ES. O programa RaptorX-Contact é usado para prever o mapa de contato. Uma estratégia interessante deste trabalho é o crossover de 2 pontos baseado na ES. Acredita-se que ele pode aumentar a exploração do espaço de possíveis regiões da estrutura secundária e descobrir diferentes dobras de proteínas.

Uma evolução diferencial multiobjetiva é proposta em (CHEN et al., 2020). Chamado

MODE-K, a função de energia é decomposta em 2 termos como função multiobjetivo, sendo um termo dependente da distância e o outro dependente da orientação. Além disso, o preditor PSIPRED é utilizado para prever a ES da proteína, essa informação é utilizada para gerar a população inicial com base nos limites definidos pela ES. Um modelo multiobjetivo também é proposto por (MARCHI; PARPINELLI, 2021), chamado de MO-BRKGA. O método utiliza 3 objetivos: pontuação de energia, informações de estrutura secundária e informações de mapas de contato. O preditor PSIPRED é utilizado para fornecer a informação de ES, e o RaptorX é usado para gerar os mapas de contato.

Também é interessante citar alguns trabalhos baseados em física. Em (JOSHI; JYOTHI, 2003) fazem a previsão de estrutura terciária através de diversos cálculos estatísticos, os quais utilizam informações como a propensão a rotação *beta* e previsões de *turns*. Já em (RUIZ-BLANCO et al., 2014) é utilizado termos como efeito hidrofóbico, interações eletrostáticas e de Van der Waals, além do potencial de torção para compor uma função de pontuação em um algoritmo populacional. Em (DORN; BURIOL; LAMB, 2013) os autores utilizam a saída prevista pelo algoritmo A3N (DORN; SOUZA, 2010) e submetem a uma simulação de dinâmica molecular (DM), com a utilização dos programas GENBOX e MDRUN da GROMACS, e o campo de força GROMOS96. As simulações levaram cerca de 600 horas de tempo de processamento. Os autores relatam diversas melhorias identificadas, incluindo a correção de desvios na estrutura do polipeptídeo e aprimoramentos em termos de RMSD.

Com base nessa revisão de literatura é possível fazer algumas observações. O uso de estratégias baseadas puramente em física eram mais utilizadas em anos anteriores, porém devido aos requisitos computacionais foi ficando para trás. Nesta categoria se encaixam os trabalhos que utilizam informações como campos de força, potencial de torção, interações eletrostáticas e de Van der Waals, além de simulações de dinâmica molecular (DM). Em geral, estas técnicas podem prever novas dobras mais facilmente, mas precisam de mais poder computacional. Porém, com a evolução computacional atualmente, estima-se que aos poucos voltem a ser utilizadas.

A montagem de fragmentos está relacionada a conformação em si, ou seja, os ângulos de torção, logo a utilização de fragmentos implica a alteração da conformação. Já a utilização de ES e mapas de contato, está relacionado a estrutura da conformação, não há uma alteração direta na conformação, mas sim uma orientação pelo espaço de busca. Por isso, seu impacto é menor e sua utilização é mais ampla, servindo tanto para orientação quanto para avaliação. É possível avaliar uma conformação segundo a sua estrutura.

Os desafios também são diferentes. Quando se fala em fragmentos, se está alterando a conformação. Logo, deve-se ter cuidado com os tamanhos de fragmentos utilizados, visto que cada tamanho tem objetivos diferentes e devem ser utilizados em momentos diferentes. Mas quando a questão é ES ou mapas de contato, o problema é a falta de confiança. Previsões incorretas podem causar grande impacto nas conformações. Neste caso, o desafio é incorporar estocasticidade ao algoritmo ou múltiplos preditores, pois nem sempre a previsão fornecida por um preditor é o melhor caminho.

Com o passar dos anos, os trabalhos passaram a utilizar cada vez mais tipos de informação do problema e de diversas formas, como por exemplo, orientando nas operações de *crossover*, mutação e até de seleção. A função de energia nem sempre representa a conformação da melhor maneira possível, por isso, considerar termos de estrutura secundária ou mapas de contato trazem a possibilidade de novas conformações. Essas mudanças mostram que trazer informação do problema para dentro da modelagem *ab initio*, podem auxiliar na construção de conformações com estruturas melhores e melhorar a exploração do espaço de busca. A Tabela 3 reune os trabalhos apresentados neste capítulo em uma ordem cronológica. São listados apenas os trabalhos dos últimos 10 anos, a partir de 2011. Os trabalhos são comparados quanto a incorporação de informação do problema no método proposto e estruturação da população. Na tabela, **IF** se refere a utilização de inserção de fragmentos, **SS** se refere ao uso de estruturas secundárias, **MC** ao uso de mapas de contato, **NP** é o número de preditores utilizados, **PE** se o trabalho utilizou algum tipo de população estruturada e, **AS** se o método alterou a função de avaliação ou se propôs alguma estratégia de seleção.

Tabela 3 – Comparação com trabalhos da literatura.

Fonte	Algoritmo	IF	ES	MC	NP	PE	AS
(DORN; BURIOL; LAMB, 2011)	-	-	S	-	-	S	-
(KAZMIER et al., 2011)	-	-	S	-	1	-	-
(WEINER et al., 2013)	-	-	S	-	2	-	-
(GARZA-FABRE et al., 2016)	RMA	S	S	-	1	-	seleção
(HAO et al., 2016)	ACUE	S	-	-	-	-	-
(ZHANG et al., 2016)	REDE	S	-	-	-	-	-
(HAO; ZHANG, 2017)	DED	S	-	-	-	-	-
(ZHANG et al., 2017)	DPDE	S	-	-	-	-	seleção
(SANTOS et al., 2017)	GAPF	-	S	S	1	-	avaliação
(CORREA; DORN, 2018)	ABC	-	S	-	-	-	avaliação
(HAO; ZHANG; ZHOU, 2018)	LUE	S	-	-	-	-	-
(MISHRA; HOQUE, 2019)	3DIGARS	-	S	-	1	-	avaliação
(ZHANG et al., 2019)	PAIE	S	S	-	-	-	-
(SILVA, 2019)	PPF-MC	S	-	-	-	-	-
(PENG; ZHOU; ZHANG, 2020)	CoDiFold	S	-	S	-	-	avaliação
(ZHANG et al., 2020)	SCDE	S	S	S	1	-	seleção
(CHEN et al., 2020)	MODE-K	-	S	-	1	-	avaliação
(CORREA; DORN, 2020)	MABC	S	S	-	-	S	avaliação
(MARCHI; PARPINELLI, 2021)	MO-BRKGA	S	S	S	1	-	avaliação
Método proposto	DSEA-FSC	S	S	S	3	S	seleção

Analisando a tabela, nota-se que ainda são poucos os trabalhos que utilizam mais de um preditor de ES, e como já foi discutido neste trabalho, os preditores podem trazer certa limitação na predição de estruturas 3D, não apenas por previsões incorretas de ES, mas simplesmente pelo fato de que não existe um limite claro entre as estruturas em sua forma natural. Logo, utilizar

mais de um preditor de ES se faz necessário para evitar este tipo de limitação. Outro ponto que vale destacar é o uso de mapas de contato, ainda são poucos os trabalhos que arriscaram seu uso, mas pode se tornar um grande aliado na previsão de estruturas melhores. Também vale notar que em nenhum dos trabalhos foi utilizada uma técnica de população baseada em especiação dinâmica, em (CORREA; DORN, 2020) a população é dividida em subpopulações de tamanho fixo e em (DORN; BURIOL; LAMB, 2011) é usada uma população estruturada. Como já discutido na Seção 2.2.3.1, a especiação dinâmica pode encontrar grupos de indivíduos potencialmente úteis em outras regiões, o que pode ajudar a espacar de mínimos locais e trazer diversidade a população. Portanto, esta é uma técnica de pesquisa ainda pouco explorada.

4 MÉTODO PROPOSTO

O presente trabalho apresenta uma proposta para o problema de predição de estrutura de proteínas, o qual utiliza uma abordagem *de novo*. Neste Capítulo, são apresentados os métodos envolvidos e a estruturação da abordagem proposta. Como o principal objetivo deste trabalho é incluir diversas formas de informação do problema para melhorar a exploração do espaço de busca, propõem-se o uso dos 3 tipos de informação mais comuns atualmente, a inserção de fragmentos, estrutura secundária e mapas de contato. Todos são utilizados pelos operadores genéticos do algoritmo de otimização. Os fragmentos são utilizados na operação de mutação para melhorar as conformações e inserir diversidade. A estrutura secundária orienta a operação de crossover na troca de informação genética. Mapas de contato e estrutura secundária são utilizados para compor uma estratégia de seleção.

A partir do trabalho de (DENG et al., 2019), foi desenvolvido um Algoritmo Evolutivo baseado na Técnica de Especiação Dinâmica. O objetivo desta técnica é organizar a população em grupos de indivíduos mais próximos entre si. Cada grupo só tem indivíduos que compartilham da mesma região do espaço, porém com valores de aptidão diferentes. No algoritmo proposto, apenas indivíduos da mesma espécie compartilham informação genética. Isto faz com que cada grupo evolua em direção ao seu ótimo local. A cada nova geração, os indivíduos são reorganizados e novas espécies são geradas, promovendo assim a diversidade da população. O uso de uma população estruturada favorece o aumento da capacidade de exploração do algoritmo. Uma vez que se tem os indivíduos separados por suas características, é possível explorar estas características de forma mais orientada.

Com base no trabalho de (SILVA, 2019), o projeto foi estruturado em 3 partes. Na primeira parte são obtidas as informações necessárias ao processo de otimização. Na segunda parte é executado o algoritmo de otimização e na terceira parte o melhor resultado encontrado é convertido para uma representação completa da proteína. O protocolo Quota do Rosetta foi selecionado para geração das bibliotecas de fragmentos por promover maior diversidade. Para isso, 3 preditores de ES foram selecionados. A execução dos preditores e geração das bibliotecas não é um processo trivial, por isso a necessidade de uma fase inicial para preparação destes materiais.

Em resumo, esta proposta é composta por um algoritmo de otimização com potencial em organizar a população para que os indivíduos possam ser explorados da melhor maneira possível. Além disso, o uso de informações do problema aumenta a capacidade exploratória e orienta a busca por melhores soluções. De acordo com o que foi visto na Seção 2.2, 3 aspectos devem ser esclarecidos quanto ao método que se propõem: a representação computacional, a função de energia e o método de busca. Cada um destes será apresentado em suas respectivas Seções 4.1, 4.2 e 4.3.

4.1 REPRESENTAÇÃO COMPUTACIONAL

A representação computacional da proteína, neste trabalho, segue a representação apresentada por (SILVA, 2019), que é uma representação no modelo de centroide, fornecido pelo Rosetta. Neste tipo de representação, os átomos da cadeia lateral são representados por um centroide localizado no centro de massa da cadeia. Cada átomo do *backbone* é representado pela tupla de ângulos diédricos (ϕ, ψ, ω), como pode ser visto na Figura 3 da Seção 2.1. A Figura 17 compara as duas formas de representação da proteína no Rosetta, por centroide e *all-atom*. Claramente, a representação por centroide é mais simplificada.

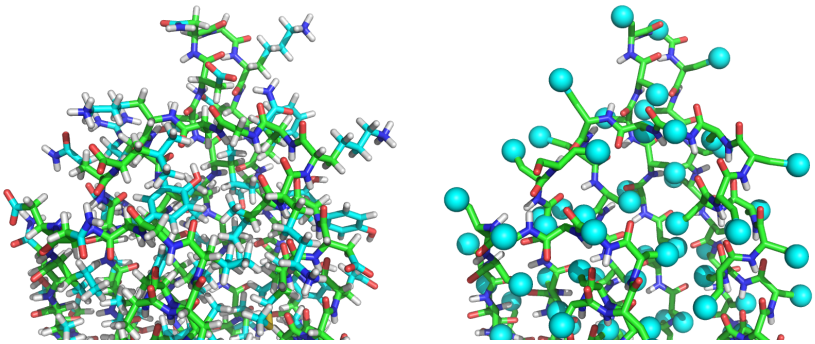


Figura 17 – Representação da proteína 1QYS no modelo *all-atom* (esquerda) e centroide (direita). Fonte: (Rosetta Commons, 2021)

Da mesma forma que no trabalho de (SILVA, 2019), neste trabalho o modelo da proteína é composto por mais 2 componentes além dos ângulos diédricos. A Figura 18 representa o modelo para a proteína 1PLW. O primeiro componente, é um objeto do Rosetta, chamado de objeto Pose. Este objeto armazena a conformação de uma proteína com todos os seus átomos, e é o responsável por atualizar os átomos a qualquer alteração nos ângulos. O segundo componente é um vetor que representa os ângulos do *backbone*, ou seja, a tupla dos ângulos diédricos (ϕ, ψ, ω). Cada posição do vetor armazena um ângulo, a cada 3 ângulos o conjunto representa um aminoácido. Logo, o tamanho do vetor possui 3 vezes o tamanho da sequência de aminoácidos.

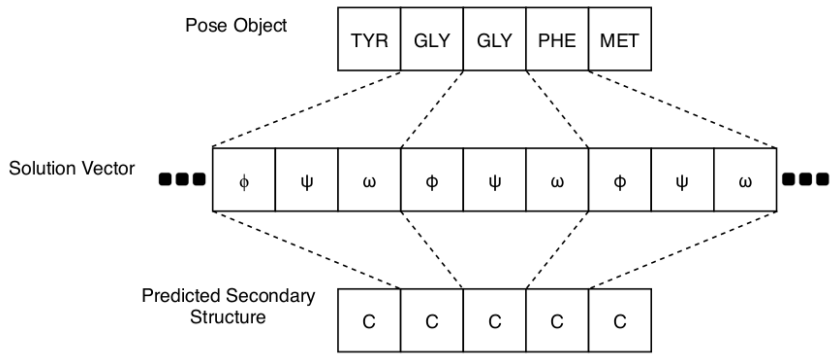


Figura 18 – Representação computacional da proteína. Fonte: (SILVA, 2019)

Vale observar que o algoritmo de otimização não pode agir sobre o objeto Pose, mas apenas sobre o vetor de ângulos, por isso o vetor serve como uma interface para o otimizador. Uma vez que todas as funções do Rosetta atuam sobre o objeto Pose, a cada alteração que o vetor de ângulos sofre pelo otimizador, o objeto Pose deve ser atualizado, e vice versa. Da mesma forma, o objeto Pose ao sofrer uma inserção de fragmentos, por exemplo, deve informar o vetor de ângulos sobre a atualização. O terceiro componente, também é um vetor que possui o tamanho da cadeia de aminoácidos e armazena as estruturas secundárias previstas pelo preditor para cada aminoácido da sequência. Este vetor serve para orientar o procedimento de busca nas operações de cruzamento e seleção.

Comumente, é usado na literatura, representações do tipo centroide para o modelo da proteína durante o processo de otimização, isto porque, como já foi dito anteriormente, uma representação *all-atom* requer muito mais recursos computacionais. Porém, também é comum, ao final do processo de otimização, representar as soluções encontradas no modelo *all-atom*. Neste trabalho, também não será diferente. Ao final do processo de otimização, é utilizada uma função chamada *Repacking* do Rosetta, a qual substitui todos os elipsoides da conformação pelas cadeias laterais. E como resultado, tem-se uma representação completa da proteína prevista.

4.2 FUNÇÃO DE ENERGIA

Este trabalho utiliza 3 funções de energia disponíveis no Rosetta, e são usadas em diferentes etapas durante o processo de previsão, de acordo com sua funcionalidade. A primeira função é chamada `score0` e ela considera apenas as forças de *Van der Waals*, onde um `score0` com valor 0 indica que uma conformação não tem conflitos entre as diferentes partes da proteína (SILVA, 2019). Esta função é destinada ao uso com inserção de fragmentos, onde os próprios fragmentos fornecem a maioria das informações. Ela é indicada para uso na primeira fase do protocolo *ab initio* do Rosetta, e neste trabalho será usada na geração da população inicial.

A segunda função é chamada `score3`, é uma função bastante usada e inclui a maioria dos termos comuns de pontuação de centroide. Um centroide utiliza uma representação completa dos átomos do *backbone* e um elipsoide para representar as cadeias laterais. Neste trabalho, a função `score3` é utilizada durante o ciclo evolutivo do processo de otimização.

A terceira função é chamada `scorefxn`, e abrange as mesmas informações que a função `score3`, porém considera a representação completa dos átomos do backbone e da cadeia lateral. Esta função é utilizada na última etapa do método proposto, ou seja, após todo o ciclo evolutivo e após o *Repacking*, com o objetivo de produzir uma representação completa da conformação.

4.3 MÉTODO DE OTIMIZAÇÃO PROPOSTO

A metodologia proposta por este trabalho consiste em três etapas: Inicialização, Otimização e Pós-processamento. A fase de Inicialização ocorre apenas uma vez por proteína, e gera insumos para a fase de otimização. Na fase de Otimização, como o próprio nome sugere, é onde

ocorre a otimização, ou seja, é onde o algoritmo de busca percorre o espaço conformacional em busca de conformações para a proteína. A saída desta etapa são as possíveis soluções para o problema. Na etapa de Pós-processamento, a conformação prevista é convertida para a representação de todos os átomos da proteína. A Figura 19 mostra o fluxograma do método proposto, envolvendo todas as fases do processo. As subseções a seguir explicam estas etapas em maiores detalhes.

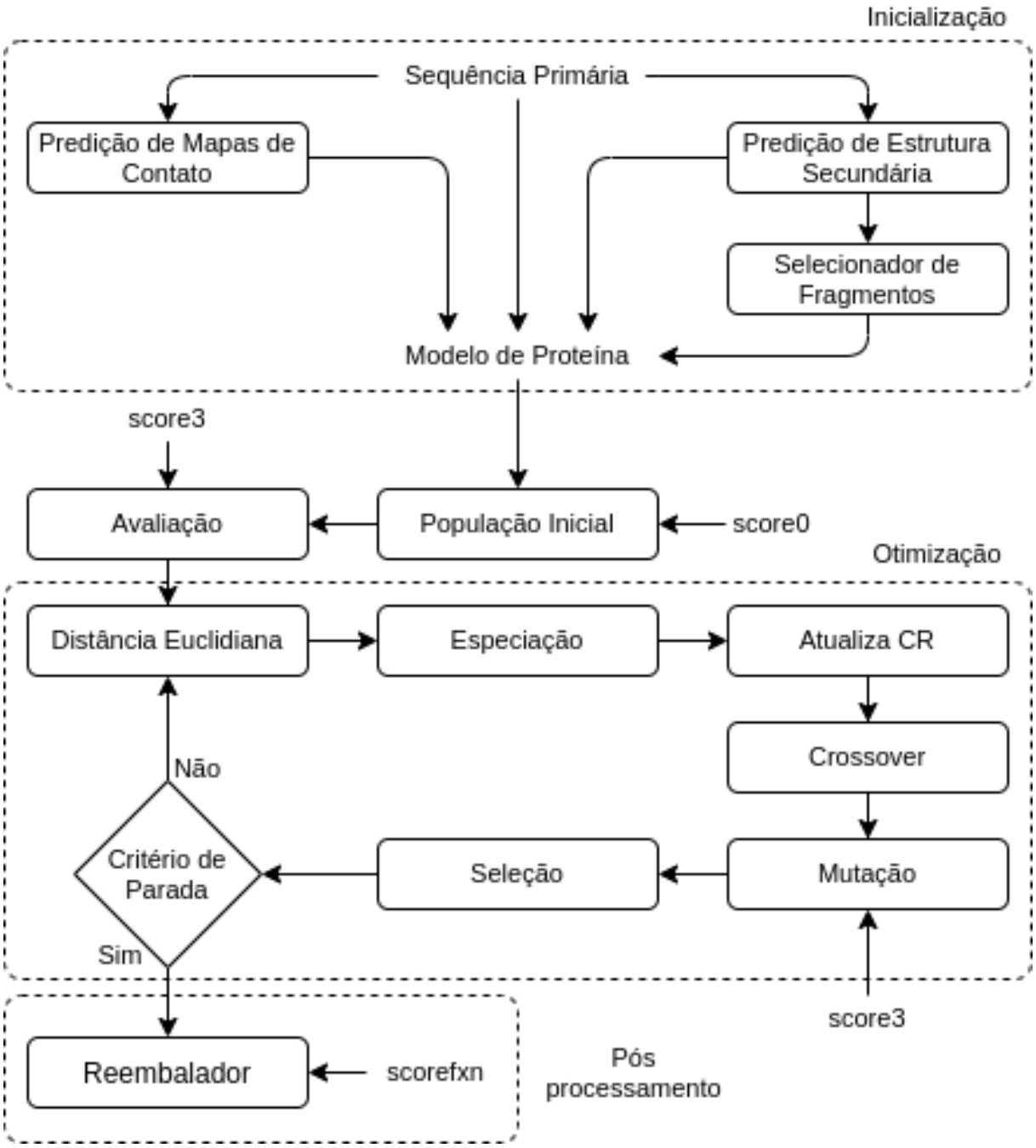


Figura 19 – Fluxograma da metodologia proposta.

4.3.1 Etapa 1: Inicialização

O grande objetivo da fase de inicialização é gerar os insumos para a fase de otimização, que são 4: sequência primária da proteína, ou seja, sua sequência de aminoácidos; estrutura secundária; mapas de contato; e uma biblioteca de fragmentos. Estas entradas só precisam ser geradas uma vez e, a partir disso, o algoritmo de otimização pode executar quantas vezes for necessário. Por isso, o tempo de execução desta etapa não é considerado nos experimentos, além de que o tempo de execução de um preditor de ES para outro pode variar muito. A etapa de inicialização teve como base o trabalho de (SILVA, 2019).

O primeiro passo é selecionar as proteínas para teste e obter sua sequência FASTA. A sequência FASTA é a sequência de aminoácidos representados pela sua sigla de 1 letra. Por exemplo, "VSCEDCPEHCSTQKAQAKCDNDKCVCEPI"é a sequência FASTA para a proteína 1ACW. Essa sequência pode ser facilmente obtida através de um banco de dados de proteínas, como por exemplo o PDB. Neste trabalho, o FASTA é utilizado pelo preditor de ES, o preditor de mapa de contato, pelo selecionador de fragmentos e pelo algoritmo de otimização.

O próximo passo, é a submissão da sequência FASTA ao preditor de ES para obtenção da previsão de ES da proteína, que por sua vez servirá de entrada para o selecionador de fragmentos. Neste ponto, se faz necessário algumas observações. A escolha do preditor de ES deve servir ao seu propósito, que neste caso é produzir insumo para o selecionador de fragmentos. Neste trabalho, será utilizado o selecionador do Rosetta, que necessita como entrada um arquivo contendo uma matriz de probabilidades de ES, conforme visto em 2.3.1. Os preditores PSIPRED, SPIDER2 e RaptorX foram selecionados por possuírem os requisitos necessários e por obterem boas precisões de previsão como apontado em trabalhos na literatura (DHINGRA et al., 2020). O método proposto inclui os 3 preditores como forma de aumentar a diversidade dos fragmentos e evitar erros de previsões incorretas. Outra observação, é que alguns preditores de ES tem certas limitações de tamanho de proteínas, por exemplo, alguns preditores só realizam previsões em proteínas com pelo menos 30 aminoácidos.

O terceiro passo envolve o selecionador de fragmentos, o qual tem o objetivo de gerar uma biblioteca de fragmentos que serão usados durante todo o ciclo evolutivo da etapa de otimização. Como já foi dito anteriormente, o Rosetta dispõem de 2 protocolos para seleção de fragmentos, o protocolo Best e o Quota. Neste trabalho, os 2 protocolos serão utilizados para fins de experimentos e comparação. Porém, o foco será no protocolo Quota, o qual utiliza 3 preditores de ES. Devem ser informados a sequência FASTA, os arquivos com a ES de cada preditor, o arquivo `quota.def` e o arquivo `quota_protocol.wghts`. As saídas dos preditores de ES devem ser adaptadas para o formato de entrada aceito pelo Rosetta, conforme visto na Seção 2.3.1. As Figuras 20 e 21 mostram um exemplo de configuração para os arquivos `quota.def` e `quota_protocol.wghts`, que é explicado em detalhes por (GRONT et al., 2011). O arquivo `quota.def` deve especificar a participação de cada preditor de ES na função de pontuação, como mostra a Figura 20. O arquivo `quota_protocol.wghts`, define os pesos para cada

componente de pontuação. Os componentes são diversos, como por exemplo, **ProfileScoreL1**, **FragmentCrmsd**, **SecondarySimilarity** e **RamaScore**. Os 2 últimos dependem das previsões feitas pelos preditores de ES (GRONT et al., 2011).

#pool_id	pool_name	fraction
1	psipred	0.6
2	jufo	0.2
3	sam	0.2

Figura 20 – Exemplo de configuração do arquivo `quota.def`. Fonte: (GRONT et al., 2011)

# score name	priority	wght	max	extras
SecondarySimilarity	350	0.5	-	psipred
SecondarySimilarity	300	0.5	-	sam
SecondarySimilarity	250	0.5	-	jufo
RamaScore	150	1.0	-	psipred
RamaScore	150	1.0	-	jufo
RamaScore	150	1.0	-	sam
ProfileScoreL1	200	1.0	-	
FragmentCrmsd	30	0.0	-	

Figura 21 – Exemplo de configuração do arquivo `quota_protocol.wghts`. Fonte: (GRONT et al., 2011)

Neste trabalho, a participação dos preditores foi definida de forma empírica, sendo 40% para o RaptorX, 30% para o SPIDER2 e 30% para o PSIPRED. A última etapa, mas que também pode ser executada em paralelo as demais por ser uma etapa independente, é a geração dos mapas de contato. Neste trabalho, foi selecionado o preditor RaptorX para os mapas de contato. Basta submeter a sequência fasta ao preditor e o mesmo retornará um arquivo contendo todos os possíveis pares de contato e suas respectivas probabilidades. Todos os resultados desta etapa são armazenados para uso na etapa de otimização.

4.3.2 Etapa 2: Otimização

Ao término da fase de inicialização, é gerada a população inicial. Os indivíduos da população inicial são conformações aleatórias geradas a partir dos fragmentos da fase anterior. Para cada indivíduo, é realizada uma busca MC (Monte Carlo) que para quando atinge um limite pré-definido de 100 iterações ou quando a função `score0` chega a zero. Quando toda a população inicial estiver gerada, é feita uma avaliação dos indivíduos utilizando a função `score3`, e então é iniciada a fase de otimização.

Como visto na Seção 2.2.3, diversas meta-heurísticas podem ser utilizadas para o problema do PSP. Os AEs se destacam por apresentarem uma população de soluções, além de sua

capacidade em convergir naturalmente para um único ótimo global (CORREA; DORN, 2020). Neste trabalho é utilizado um algoritmo evolutivo para o problema do PSP, e para aumentar o seu desempenho, principalmente no que diz respeito a diversidade, é utilizada a técnica de especiação dinâmica com base na estrutura do DSM-DE (2.2.3.1). Porém, como este trabalho não utiliza duas estratégias de mutação, nem utiliza a mutação com base em operações de diferenças, não cabe aqui chamá-lo de DSM-DE nem de DE, mas sim um algoritmo evolutivo baseado em especiação dinâmica. Portanto, a fase de otimização se dá pela execução de um algoritmo evolutivo com base em especiação dinâmica.

Neste trabalho, a Técnica de Especiação Dinâmica (DST, sigla em inglês) é utilizada com o objetivo de estruturar a população, dividindo-a em grupos, de forma a orientar a busca pelo espaço conformacional e aumentar a diversidade. Os grupos são formados por indivíduos com diferentes valores de aptidão, mas que estão próximos entre si quanto a sua posição espacial, isto é chamado de vizinhança baseada em distância. Assim, cada espécie ou grupo representa uma região no espaço conformacional, e a semente de cada espécie é o indivíduo com melhor aptidão. Dessa forma, a semente de cada espécie representa um ótimo local. Nesta ideia, o DST entra como uma estratégia de diversificação, que vai reagrupar a população a cada iteração para localizar novos ótimos locais. Enquanto que o AE tem a função de intensificação, através das operações de cruzamento e mutação agindo localmente em cada espécie. O Algoritmo 2 apresenta o pseudocódigo para o algoritmo evolutivo utilizado neste trabalho.

As linhas de 1 a 9 representam a inicialização das variáveis. O tamanho da população *NP* (linha 5) e *C_max* (linha 9) foram definidos conforme orientação dos autores do DSM-DE (DENG et al., 2019). *S_num* e *m* são definidos conforme a equação (1). A variável *C_min* é definida empiricamente como a raiz quadrada de *NP*. Esta variável é utilizada no cálculo do tamanho das espécies. Através de diversos testes foi observado que um valor menor poderia prejudicar a especiação. As demais variáveis são definidas de acordo com as características do problema, ou seja, da proteína em questão. A partir da linha 12 o algoritmo entra no ciclo evolutivo, o ciclo termina após alcançar um limite pré-definido de funções de avaliação. A saber, uma função de energia representa uma função de avaliação, logo, a cada função de energia (`score0` ou `score3`) executada, soma-se um valor a variável FES.

Então, é calculado o valor de PSC, que deve auxiliar na etapa de seleção. Antes de realizar a especiação, deve ser calculada a distância euclidiana de todos os indivíduos para todos os indivíduos (linha 14), isto porque a especiação é realizada com base nesse valor. A distância euclidiana é calculada usando a seguinte equação:

$$dist(\vec{X}^{(i)}, \vec{X}^{(j)}) = \sqrt{\sum_{k=1}^{D} (x_k^{(i)} - x_k^{(j)})^2} \tag{2}$$

Onde $\vec{X}^{(i)}$ e $\vec{X}^{(j)}$ são indivíduos da população, de forma que para cada indivíduo X

```
1  FES: Funções de Avaliação;
2  D: Dimensões do problema;
3  FES_max = NP * 10000: Máximo de Funções de Avaliação;
4  g: Geração atual;
5  NP = 5 * L: Tamanho da população (L=número de aminoácidos da proteína);
6  m: Tamanho da especiação;
7  S_num: Número de especiações;
8  C_min = sqrt(NP): Tamanho mínimo da especiação;
9  C_max = NP/10: Tamanho máximo da especiação;
10 Inicialização
11 Gera a população inicial P_0 = X_{1,0}, X_{2,0}, ..., X_{NP,0};
12 while FES < FES_max do
13     Calcula a probabilidade baseda na estrutura secundária P_SS;
14     Calcula a distância euclidiana entre os indivíduos da pop. P;
15     Executa DST no Algoritmo 1 para obter S = S_1, S_2, ..., S_{S_num} com S_num espécies;
16     for i = 1 : S_num do
17         for j = 1 : m(i) do
18             Atualiza CR_{i,j,g};
19             Gera vetor trial como cópia do vetor atual: U_{i,j,g} ← X_{i,j,g};
20             Crossover
21             Realiza cruzamento entre o vetor trial U_{i,j,g} e a semente da espécie atual
                X_{seed,j,g} conforme (4.3.2.1);
22             Mutação
23             Realiza mutação no vetor trial U_{i,j,g}, a partir da inserção de fragmento,
                conforme (4.3.2.2);
24             Seleção
25             Gera número aleatório r entre (0, 1);
26             if r ≤ P_SS then
27                 Realiza a seleção baseada em contato;
28             else
29                 Realiza a seleção baseada em estrutura;
30             end
31         end
32     end
33 end
```

Algoritmo 2: Pseudocódigo do Algoritmo Evolutivo baseado em DST

tem-se:

$$X = (aa_1, aa_2, ..., aa_D) \tag{3}$$

Onde aa_i representa os aminoácidos da sequência, e D representa a quantidade de aminoácidos na sequência. E para cada aminoácido aa tem-se uma tupla com os 3 ângulos

diédricos, conforme:

$$aa_i = (\phi_i, \psi_i, \omega_i) \tag{4}$$

Em seguida, a população é dividida em espécies (linha 15), conforme explicado na Seção 2.2.3.1 (ver Algoritmo 1, considerando os valores para *C_min* e *C_max* atualizados conforme Algoritmo 2). Os demais estágios do AE, cruzamento, mutação e seleção, ocorrem sobre cada indivíduo, considerando apenas a espécie atual. O cruzamento ocorre entre o indivíduo atual (`target`) e a semente da espécie, de acordo com uma taxa de CR, que é definida na Equação 5 (BREST et al., 2006).

$$CR_{i,g+1} = \begin{cases} r_1 & \text{if } r_2 < T \\ CR_{i,g} & \text{otherwise} \end{cases} \tag{5}$$

Onde r_1 e r_2 são valores aleatórios $\in \{0.1\}$, e $T = 0.1$.

Nota-se que o vetor trial é uma cópia do indivíduo atual que, de acordo com uma taxa de CR, pode receber informação genética da semente da espécie. Uma vez que a semente é o indivíduo da espécie com melhor aptidão, espera-se, com isto, melhorar a aptidão do novo indivíduo. Assim, pode-se dizer que este procedimento de cruzamento busca acelerar a convergência para o ótimo local. Em contrapartida, a operação de mutação deve inserir um fragmento no novo indivíduo gerado com o intuito de causar uma perturbação. Este procedimento deve trazer diversidade e evitar a convergência prematura. Vale observar que, caso o target seja a própria semente, o cruzamento é desnecessário, mas a mutação vai fazer a semente competir com uma variação dela mesma.

4.3.2.1 Crossover de 2 pontos baseado na estrutura secundária

Neste trabalho, o cruzamento sempre ocorre entre o indivíduo atual e a semente da espécie, e é semelhante a um cruzamento de 2 pontos. Neste tipo de cruzamento, são compartilhadas apenas as informações que estão entre estes 2 pontos. Aqui, os 2 pontos foram definidos conforme o trabalho de (ZHANG et al., 2020). O primeiro ponto é selecionado aleatoriamente entre as posições dos aminoácidos, o segundo ponto é definido de acordo com a ES. Desde o primeiro ponto, a ES de cada aminoácido é verificada de acordo com a previsão do preditor RaptorX. Quando a ES do aminoácido em questão é diferente da ES do primeiro ponto, a verificação é finalizada e o segundo ponto é definido. O segundo ponto é o último aminoácido na sequência cuja ES é igual ao primeiro ponto.

A troca de informação entre o primeiro e o segundo ponto é feito de acordo com a probabilidade CR. Para cada aminoácido no trecho entre os 2 pontos, é selecionado um valor aleatório entre 0 e 1. Se o valor for menor que CR, o indivíduo Trial recebe a informação genética

da semente para o aminoácido em questão. Neste trabalho, a operação de cruzamento resultará em apenas 1 filho. Chamado de vetor trial, este filho é uma cópia do vetor target, ou seja, o indivíduo atual, que também é um dos pais. Assim, de acordo com a verificação descrita, o filho pode receber um gene do pai semente, ou pode simplesmente permanecer com o gene do pai target, já que o trial é uma cópia do target. A saber, um gene representa um aminoácido, que é uma tupla dos ângulos diédricos. A Figura 22 mostra o comportamento dessa operação.

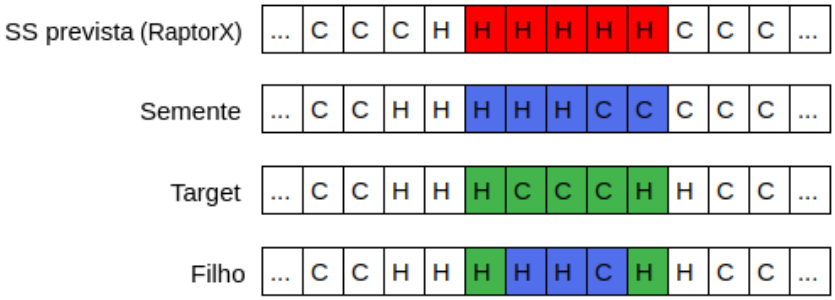


Figura 22 – Exemplo do operador de cruzamento.

O cruzamento uniforme, como geralmente é feito no DE, envolve muito tempo de processamento, pois exige a verificação de todos os aminoácidos da sequência, o que aumenta proporcionalmente ao tamanho da proteína. Já o cruzamento apenas no trecho entre 2 pontos, reduz consideravelmente a quantidade de verificações dos aminoácidos da sequência, o que reduz também o tempo de processamento. Além disso, o cruzamento de 2 pontos procura manter a diversidade das conformações, além de possibilitar encontrar novas formas de ES.

4.3.2.2 Mutação baseada em inserção de fragmentos

A mutação baseada em fragmentos sempre ocorre, mesmo se não houver cruzamento. O objetivo é melhorar as conformações e trazer diversidade. Neste trabalho, a mutação é feita pela inserção do fragmento do Rosetta. Considerando os 4 operadores vistos na Seção 2.3.1, sendo 2 operadores clássicos de tamanho 3 e 9, e 2 operadores suaves de tamanho 3 e 9. Ao iniciar a operação de mutação, um dos 4 operadores é selecionado aleatoriamente, e uma busca MC (Monte Carlo) é iniciada e executada num máximo de 100 iterações ou até que a conformação tenha melhorado. Isto se faz necessário, uma vez que a inserção de fragmentos pode "bagunçar"a conformação e piorar seu valor de aptidão. É comum, que nas primeiras iterações do ciclo evolutivo, quando as conformações ainda estão muito longe do esperado, sejam necessárias menos inserções de fragmentos até melhorar a conformação. Com o passar dos ciclos, torna-se cada vez mais difícil melhorar as conformações, logo, o número de funções de avaliação gastas por ciclo aumenta com o passar do tempo.

4.3.2.3 Seleção baseada em estrutura e contato

São utilizadas duas estratégias de seleção inspiradas em (ZHANG et al., 2020). Uma estratégia de seleção baseada em ES e outra estratégia baseada em contato. Na estratégia de seleção baseada em ES, é calculado um termo que representa a ES do indivíduo. Quanto maior o valor deste termo significa que o indivíduo possui a ES mais próxima do previsto. Para isso, é verificado a ES de cada aminoácido da conformação, segundo o DSSP. Então, é somado a probabilidade prevista para esta estrutura. Ao final, a soma de todas as probabilidade representa o termo de ES para a conformação. Nessa estratégia de seleção, a conformação (Target ou Trial) que tiver o valor mais alto para o termo de ES é selecionado para a próxima geração.

Na estratégia de seleção baseada em mapa de contatos, é calculado um termo que representa a probabilidade de contatos do indivíduo. Quanto maior o valor deste termo significa que o indivíduo possui mais contatos, e está mais próximo do predito. Para o cálculo do termo P_{CM}, são considerados os L primeiros contatos previstos pelo RaptorX, sendo L a quantidade de aminoácidos da proteína alvo. Para cada contato previsto, é calculado a distância d_i entre o par de resíduos para a conformação. Se a distância d_i for menor ou igual a 8Å, soma-se o valor da probabilidade prevista p_i. Caso contrário, é aplicada uma penalidade, como mostra a Equação 6. Ao final, a soma das probabilidades representa o termo de contato para a conformação. Nessa estratégia, a conformação que tiver o valor mais alto para este termo é selecionado para a próxima geração.

$$P_{CM} = \sum_{i}^{L} \begin{cases} p_i & \text{if } d_i \leq 8 \\ p_i/d_i & \text{otherwise} \end{cases} \tag{6}$$

Toda vez que o ciclo evolutivo chega na etapa de seleção, o algoritmo deve escolher qual das duas estratégias deve ser utilizada. Isto é feito de forma adaptativa, com base no valor calculado para P_{SS}. P_{SS} utiliza a informação de ES de toda a população para calcular o quanto esta população está estruturalmente ajustada. Desta forma, um valor alto para P_{SS} demonstra que a seleção baseada em contato pode ser mais adequada. Então, um número aleatório distribuído uniformemente r entre 0 e 1 é gerado. Se r for maior que P_{SS}, a seleção baseada em estrutura é utilizada. Caso contrário, a seleção baseada em contato é utilizada. A Equação 7 é fornecida em (ZHANG et al., 2020) e apresenta o cálculo para o P_{SS}.

$$P_{SS} = exp\left(-c * \hat{E}_{SS} * (\frac{\bar{E}_{SS}}{L} - 1)^2\right) \tag{7}$$

Onde L é o comprimento da sequência, $\bar{E}_{SS}$ é a média de E_{SS} para todas as conformações na população atual, $\hat{E}_{SS}$ é a variação correspondente e c é uma constante. A Equação 8 apresenta

o cálculo para E_{SS} (ZHANG et al., 2020).

$$E_{SS} = \sum_{t=1}^{L} \begin{cases} 1 & \text{if } S_t^{predicted} = S_t^{trial} \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

O valor de P_{SS} deve ser calculado a cada nova geração. O objetivo em utilizar duas estratégias de seleção é evitar conformações "problemáticas"devido a informação incorreta dos preditores. Em resumo, um valor alto de P_{SS} indica que a média das conformações está com a ES mais ajustada. Logo, aumentam as chances em utilizar a estratégia baseada em contato. A troca entre as estratégias de seleção muda o valor de P_{SS}, pois mudam as conformações. Isso provoca um refinamento das conformações ao longo das gerações.

4.3.3 Etapa 3: Pós-processamento

Ao final do ciclo evolutivo, o RMSD de todos os indivíduos da população é calculado com base na conformação nativa. O indivíduo com o menor valor RMSD é selecionado para se submeter ao procedimento de reembalagem (*repacking*) para ter uma representação completa da conformação. Após a reembalagem, são calculados o `scorefxn` do indivíduo e o RMSD para a conformação nativa.

5 EXPERIMENTOS, RESULTADOS E ANÁLISES

Neste capítulo, são apresentados os resultados e análises dos experimentos realizados. Os experimentos realizados tiveram 2 objetivos principais. Primeiro, comparar o desempenho dos 2 protocolos do Rosetta (Best e Quota) para geração de bibliotecas de fragmentos. Segundo, analisar o desempenho do método proposto em comparação com outros trabalhos na literatura. Na Seção 5.1 são explicadas as configurações necessárias para a execução dos testes, bem como a configuração de hardware, definição dos parâmetros do algoritmo e os preditores selecionados para teste. Na Seção 5.2 é feita uma análise dos valores do RMSD e GDT das proteínas. É realizado um teste ANOVA sobre estes valores para verificar se as diferenças sao estatisticamente significativas. Na Seção 5.3 é feita a análise de convergência da função de energia `score3`. Na Seção 5.4 é feita uma comparação dos tempos de processamento e na Seção 5.6 é feita uma representação visual das proteínas-alvo em comparação com a proteína nativa. Das seções 5.2 a 5.6 o objetivo é comparar o desempenho dos protocolos Best e Quota. Ao final do capítulo, na Seção 5.5 é apresentada uma comparação dos resultados obtidos com o protocolo Quota e outros métodos da literatura, em termos de RMSD e GDT.

5.1 CONFIGURAÇÃO DOS EXPERIMENTOS

Todos os experimentos foram executados na mesma máquina. A Tabela 4 apresenta as configurações de hardware da máquina utilizada. O algoritmo foi implementado na linguagem Python. A experimentação consistiu na execução do método proposto, considerando as três etapas. Primeiramente, foi executada a fase de Inicialização para cada proteína, onde foram geradas as bibliotecas de fragmentos. O resultado, é uma grande base de dados com todos os insumos necessários para a realização da próxima etapa. O tempo de processamento da primeira etapa não é considerado nos resultados da experimentação. Em seguida é executada a fase de Otimização seguida pelo Pós-processamento, onde são feitos registros de tempo de processamento, RMSD, GDT e energia, para posterior análise. Cada experimentação de teste consiste em um programa serial que é executado sem interrupção.

Tabela 4 – Configurações de teste

Nome	Valor
Sistema Operacional	Ubuntu 18.04 x86_64
CPU	Intel Core i9-9900k CPU 3,60 GHz
Número de núcleos	8 núcleos e 16 threads
RAM	64 GB

Uma quantidade de 9 proteínas do PDB com diferentes tamanhos e tipos de ES foram selecionadas para teste. Estas proteínas foram selecionadas a partir da revisão de literatura apresentada no Capítulo 3. Optou-se por proteínas que apareceram em mais de um trabalho

e que forneceram o seu melhor RMSD. Além disso, também era essencial que as proteínas não apresentassem problemas na execução pelos 3 preditores de ES. A Tabela 5 apresenta as 9 proteínas ordenadas por tamanho (L), da menor para a maior. Na primeira coluna, é identificado o ID da proteína conforme se encontra no PDB. A coluna Tamanho, informa a quantidade de aminoácidos presentes na proteína. A coluna Estrutura, informa a quantidade e o tipo de ES presente na proteína, sendo α para representar as hélices e β para representar as folhas.

Tabela 5 – Conjunto de teste de proteínas-alvo

ID-Proteína	Tamanho (L)	Estrutura
1ACW	29	$1\alpha/2\beta$
1ZDD	34	2α
1I6C	39	3β
2MR9	44	3α
2P81	44	2α
1CRN	46	$2\alpha/2\beta$
1ENH	54	3α
1ROP	63	2α
1AIL	73	3α

Na fase de Inicialização, 4 bibliotecas de fragmentos foram geradas para cada proteína. Uma biblioteca para o protocolo Quota e 3 bibliotecas para o protocolo Best. Como o protocolo Quota utiliza 3 preditores em conjunto, só é necessário 1 biblioteca para este protocolo. Como o objetivo é comparar o desempenho dos 2 protocolos, e como o protocolo Best utiliza apenas 1 preditor, é necessário gerar uma biblioteca para cada preditor. Assim, tem-se 1 biblioteca para o protocolo Quota com os 3 preditores juntos, e 3 bibliotecas para o protocolo Best, sendo uma para cada preditor individualmente. Neste trabalho, foram utilizados os preditores PSIPRED, SPIDER2 e RaptorX. Para o protocolo Quota, também é necessário definir a participação de cada preditor. Como os 3 preditores tiveram desempenho semelhante, de acordo com (SMOLARCZYK et al., 2020), foi definida a participação de 40% para RaptorX, 30% para SPIDER2 e 30% para PSIPRED. Esses valores foram escolhidos empiricamente.

Na fase de Otimização, uma quantidade de 30 execuções independentes são realizadas para cada biblioteca de fragmentos em cada proteína, é calculada a média e desvio padrão destas execuções e são apresentados nos resultados a seguir. Quanto as configurações do algoritmo de otimização, o tamanho da população NP é definido de acordo com o tamanho da proteína. Assim, $NP = 5 * L$, onde 5 é um fator definido de acordo com (DENG et al., 2019) e L é a quantidade de aminoácidos em uma proteína. Por exemplo, para a proteína 1CRN, que tem 46 aminoácidos, $NP = 5 * 46 = 230$ indivíduos. FES_{max} foi definido com um máximo de 1000000 de avaliações. C da Equação 7 foi definido como 2, de acordo com (ZHANG et al., 2020).

Foram aplicadas 2 métricas para avaliar os resultados encontrados, RMSD e GDT. The *Root Mean Square Deviation* (RMSD) é umas das métricas mais comuns utilizadas para medir a

similaridade entre duas proteínas. Um valor de RMSD mais baixo, significa maior similaridade entre as sequências. A distância é calculada considerando os átomos C_α de cada aminoácido. A fórmula do RMSD pode ser expressa como na Equação 9. Onde A e B são duas proteínas a serem comparadas, e n é a quantidade de aminoácidos das sequências.

$$RMSD(A,B) = \sqrt{\frac{\sum_{i=1}^{n}(A_i - B_i)^2}{n}} \tag{9}$$

Normalmente, o RMSD é calculado entre a proteína-alvo, que se deseja identificar, e a proteína nativa. Quanto menor o valor do RMSD, mais próximo da conformação nativa está a previsão. A métrica *Global Distance Test* (GDT), assim como o RMSD, avalia a semelhança entre duas estruturas. Porem, pode ser considerada mais robusta, por não ser tão sensível as regiões irregulares da proteína. Diferente do RMSD, para o GDT, um valor mais alto, indica maior similaridade entre as sequências. O GDT pode ser calculado conforme a Equação 10. Onde GDT_n representa o número de resíduos até o limite de $\leq n$.

$$GDT = \frac{(GDT_1 + GDT_2 + GDT_4 + GDT_8)}{4} \tag{10}$$

5.2 ANÁLISE DO SCOREFXN, RMSD E GDT

A Tabela 6 faz uma comparação entre os resultados alcançados pelo protocolo Quota e pelo protocolo Best. Para isso, são considerados os resultados obtidos com o uso de cada biblioteca de fragmentos. Na primeira coluna está o PDB-ID da proteína. Na segunda coluna, as bibliotecas para cada protocolo. Na terceira e quarta colunas, o melhor valor de RMSD nas 30 execuções, média e desvio padrão. Na quinta e sexta colunas, o melhor valor de GDT nas 30 execuções, média e desvio padrão. As proteínas estão organizadas em ordem lexicográfica.

Para determinar se os resultados da Tabela 6 são estatisticamente diferentes, deve-se realizar uma análise estatística. Para isso, foi realizado um teste estatístico utilizando ANOVA com o Tukey HSD post-hoc para cada proteína. Foi determinado um nível de confiança de 95%. ANOVA é realizado utilizando os valores de média e desvio padrão, para RMSD e GDT separadamente. Os resultados estatisticamente significativos são destacados em negrito preto. Os melhores resultados gerais são destacados em negrito vermelho. As duas últimas linhas apresentam um resumo dos resultados obtidos. Na linha B/S/W, B é o número de instâncias do problema onde Quota foi melhor estatisticamente que as demais bibliotecas do protocolo Best, S ele foi semelhante e W ele foi pior. Essa linha considera apenas a média e desvio padrão. Na linha B* é apresentado o número de instâncias do problema onde cada biblioteca obteve o melhor resultado geral.

Para complementar a análise da Tabela 6 é apresentado um gráfico em formato Boxplot nas Figuras 23 e 24. A Figura 23 apresenta o Boxplot para o RMSD das previsões, e a Figura 24 apresenta o Boxplot para o GDT. No eixo x se encontram as proteínas em ordem lexicográfica. No

Tabela 6 – Comparação dos resultados

PDB-ID	Algorithm	RMSD		GDT	
		Best	Média±DP	Best	Média±DP
1ACW	Quota	**1.74**	**3.15 ± 0.77**	**86.90**	**70.80 ± 9.70**
	Best-PSIPRED	3.41	4.79 ± 0.78	63.45	46.37 ± 6.51
	Best-SPIDER2	3.50	4.36 ± 0.55	71.03	52.23 ± 7.67
	Best-RaptorX	2.72	**3.40 ± 0.46**	76.55	**67.49 ± 6.33**
1AIL	Quota	**2.53**	4.28 ± 1.48	**83.14**	58.26 ± 11.18
	Best-PSIPRED	3.40	4.80 ± 1.88	68.29	53.76 ± 8.72
	Best-SPIDER2	3.25	5.42 ± 1.86	66.29	48.71 ± 8.46
	Best-RaptorX	3.17	5.21 ± 2.96	76.57	54.52 ± 8.26
1CRN	Quota	**0.85**	**1.58 ± 0.62**	**97.39**	**88.94 ± 7.22**
	Best-PSIPRED	1.92	3.55 ± 0.81	86.52	66.97 ± 10.33
	Best-SPIDER2	1.09	2.15 ± 0.89	95.65	**83.52 ± 9.97**
	Best-RaptorX	1.45	2.76 ± 0.69	91.30	78.23 ± 7.35
1ENH	Quota	1.45	2.36 ± 0.47	96.30	86.94 ± 6.91
	Best-PSIPRED	**1.00**	1.95 ± 0.55	**97.78**	86.73 ± 7.17
	Best-SPIDER2	1.18	2.17 ± 0.46	94.44	85.99 ± 6.11
	Best-RaptorX	1.14	2.25 ± 0.83	95.93	83.42 ± 8.45
1I6C	Quota	2.97	4.42 ± 1.08	74.36	61.95 ± 7.28
	Best-PSIPRED	3.10	4.29 ± 0.87	75.38	63.03 ± 5.97
	Best-SPIDER2	**2.45**	4.12 ± 0.96	76.41	63.98 ± 6.82
	Best-RaptorX	2.93	4.14 ± 1.07	**78.46**	64.72 ± 5.98
1ROP	Quota	**0.94**	2.11 ± 0.55	92.86	77.20 ± 8.47
	Best-PSIPRED	1.79	2.46 ± 0.43	80.71	72.10 ± 5.76
	Best-SPIDER2	0.98	2.12 ± 0.53	**93.21**	76.79 ± 7.95
	Best-RaptorX	1.63	2.50 ± 0.69	86.07	72.82 ± 8.01
1ZDD	Quota	0.85	1.19 ± 0.24	97.06	91.65 ± 3.33
	Best-PSIPRED	**0.79**	1.23 ± 0.24	97.06	91.39 ± 3.68
	Best-SPIDER2	0.87	1.17 ± 0.20	96.47	91.96 ± 3.00
	Best-RaptorX	0.89	1.12 ± 0.13	**97.65**	93.41 ± 2.31
2MR9	Quota	**1.41**	2.25 ± 0.51	87.73	83.15 ± 4.90
	Best-PSIPRED	1.50	2.15 ± 0.43	91.82	82.82 ± 5.03
	Best-SPIDER2	1.77	2.35 ± 0.53	89.09	82.15 ± 5.65
	Best-RaptorX	1.49	2.20 ± 0.55	**93.18**	81.95 ± 6.93
2P81	Quota	**3.66**	4.98 ± 0.53	**73.64**	**61.64 ± 3.87**
	Best-PSIPRED	3.91	5.06 ± 0.83	70.00	**62.70 ± 5.53**
	Best-SPIDER2	4.00	5.37 ± 0.61	64.09	57.20 ± 4.11
	Best-RaptorX	4.37	5.47 ± 0.52	63.18	57.26 ± 3.91
B/S/W	Quota		-		-
	Best-PSIPRED		2/6/1		2/7/0
	Best-SPIDER2		2/7/0		3/6/0
	Best-RaptorX		3/6/0		2/7/0
B*	Quota		6		4
	Best-PSIPRED		2		1
	Best-SPIDER2		1		1
	Best-RaptorX		0		3

eixo y, o valor de RMSD e GDT, respectivamente. As bibliotecas estão agrupadas horizontalmente por proteína. No caso do RMSD, quanto menor o valor, melhor. No caso do GDT, o valor mais

alto é melhor.

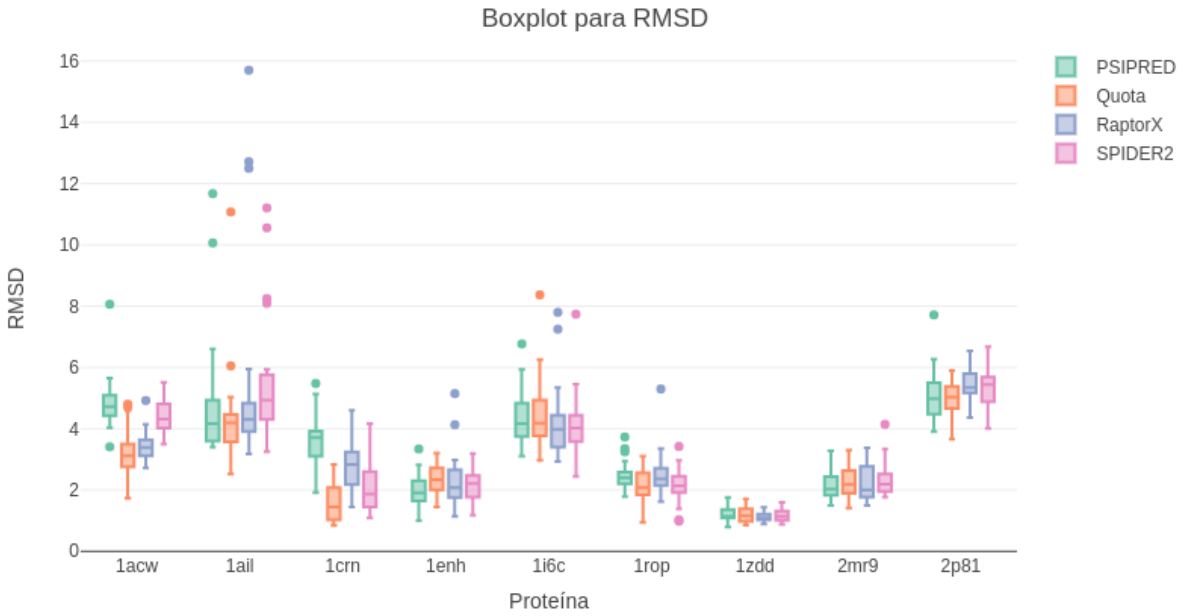


Figura 23 – Boxplot do RMSD para as previsões das proteínas com as 4 bibliotecas.

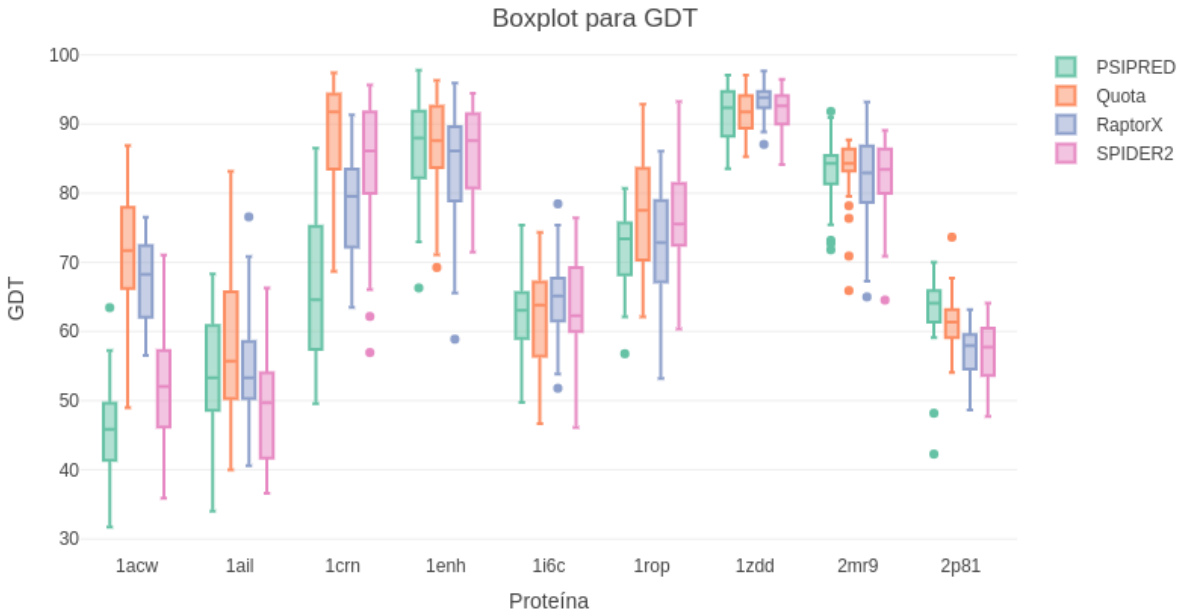


Figura 24 – Boxplot do GDT para as previsões das proteínas com as 4 bibliotecas.

Os gráficos em Boxplot facilitam a análise e confirmam os resultados estatísticos dos testes ANOVA da Tabela 6. Analisando a Figura 23, com relação ao RMSD, é possível notar que a maioria das proteínas apresentaram resultados muito semelhantes, com exceção para as proteínas 1ACW e 1CRN. Os testes ANOVA indicam que, das 9 proteínas, 4 delas (1AIL, 1I6C, 1ZDD e 2MR9) não apresentaram nenhuma diferença significativa entre as bibliotecas. Porém, para 3 proteínas (1ENH, 2P81 e 1ROP) houve diferenças significativas, mas apenas entre 2 bibliotecas, por isso, não se encontram em negrito preto na Tabela 6. A proteína 1ENH apresentou diferença entre Quota e PSIPRED, sendo que o PSIPRED obteve o melhor resultado,

mas se manteve semelhante aos demais. A proteína 2P81 apresentou diferença entre Quota e RaptorX, sendo que o Quota obteve melhor resultado, porém se manteve semelhante aos demais. Já para a proteína 1ROP, tanto Quota quanto SPIDER2 apresentaram diferenças com relação ao RaptorX, porém ambos são semelhantes entre si e ao PSIPRED. Assim, nestes 3 casos, não é possível dizer qual é o melhor estatisticamente, por isso não foram destacados na tabela.

Para a proteína 1CRN houve diferenças significativas entre todas as bibliotecas, sendo que o Quota se mostrou melhor que todos estatisticamente, por isso o seu destaque na tabela. Com relação a proteína 1ACW, tanto Quota quanto RaptorX se sobressaíram aos demais, porém sendo semelhantes entre si. A linha B/S/W da Tabela 6 resume esta análise, onde é possível ver que o Quota foi pior em apenas 1 proteína (a 1ENH) apenas para o PSIPRED, sendo o melhor em pelo menos 2 proteínas com relação a todas as bibliotecas, e apresentando semelhanças nas demais. Sobre o melhor resultado geral, a linha B* da tabela mostra que o Quota alcançou os melhores resultados para 6 proteínas. Com isso, é possível dizer que o protocolo Quota pode obter boas médias ou até melhores que o protocolo Best, mas além disso pode obter os melhores resultados para a maioria das proteínas.

Analisando a Figura 24, é possível notar que o GDT segue a mesma tendência mostrada pelo RMSD. Não apresentando diferenças significativas para 1ENH, 1I6C, 1ROP, 1ZDD e 2MR9. A proteína 1AIL apresentou diferença significativa apenas entre Quota e SPIDER2, se mantendo semelhante entre os demais. Diferente do RMSD, o GDT apresentou diferenças significativas para 3 proteínas, sendo elas 1ACW, 1CRN e 2P81. Porém, nos 3 casos não é possível afirmar que apenas uma biblioteca tenha sido a melhor estatisticamente, mas o Quota aparece nos 3 casos. O que sugere que o Quota pode alcançar melhores médias para mais proteínas. Na linha B/S/W nota-se que o Quota não obteve a pior média em nenhuma situação. Com relação ao melhor resultado geral apresentado na linha B*, das 9 proteínas, Quota obteve o melhor resultado para 4 delas.

A Figura 25 apresenta um gráfico em formato Boxplot para o `scorefxn`, semelhante as Figuras 23 e 24. Comparando este Boxplot com os anteriores para RMSD e GDT, nota-se que os 3 gráficos apresentam comportamentos semelhantes para as proteínas, sendo mais visível em alguns casos. Como por exemplo, a proteína 1AIL que apresenta um resultado melhor para o Quota com relação as outras bibliotecas, tanto para RMSD e GDT, quanto para o `scorefxn`. Igualmente para a proteína 1CRN e 1ROP, que também apresentaram melhores resultados para o Quota nos 3 gráficos. Já para a proteína 2P81, fica mais claro a semelhança nos resultados para o GDT e `scorefxn`. Esta análise visual mostra a relação entre o `scorefxn` e os resultados alcançados para RMSD e GDT.

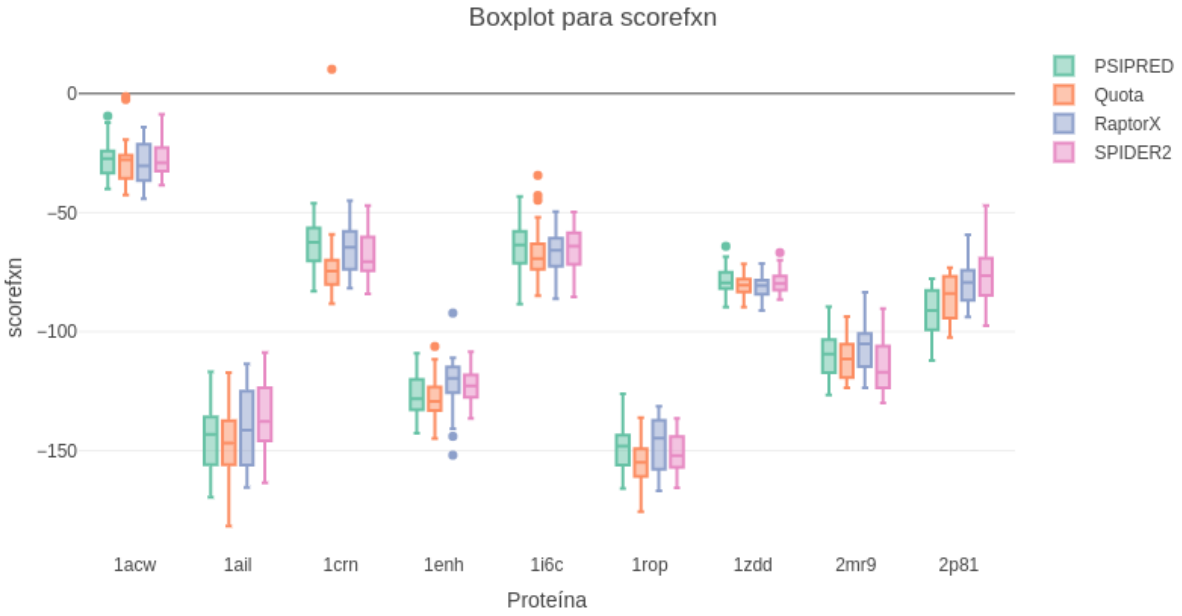


Figura 25 – Boxplot do `scorefxn` para as previsões das proteínas com as 4 bibliotecas.

5.3 ANÁLISE DE CONVERGÊNCIA DE ENERGIA SCORE3

As Figuras 26 e 27 apresenta o gráfico de convergência de energia para a função `score3`, para cada proteína do conjunto de teste. O gráfico é gerado calculando a média das 30 execuções do melhor indivíduo. As proteínas estão organizadas em ordem lexicográfica da esquerda para a direita, de cima para baixo. De maneira geral, o gráfico tem um comportamento similar para todas as proteínas. Convergindo rapidamente entre 200000 e 400000 avaliações, mas o melhor indivíduo continua evoluindo até o final das execuções. Outra observação, é que a medida que o tamanho da proteína aumenta, a curva de convergência também aumenta, isto pode ser observado principalmente nas proteínas 1ROP e 1AIL. Nestas proteínas, nota-se que a energia diminui a um passo mais lento, o que pode explicar a necessidade de mais iterações para proteínas maiores, para alcançar uma energia mais baixa.

Outra observação interessante, mais fácil de observar na proteína 1ACW, mas que também ocorre nas outras proteínas, são pequenos aumentos no valor da energia ao longo da evolução. Uma vez que o algoritmo proposto não é elitista e a estratégia de seleção não é baseada no valor da energia. A troca entre as duas estratégias de seleção pode causar esse impacto no valor do `score3`. Até que haja um equilíbrio entre contato e estrutura para que o algoritmo convirja para a melhor solução.

Comparando as bibliotecas, das 9 proteínas, 5 (1ACW, 1I6C, 1CRN, 1ENH, 1ROP) apresentaram a menor média do `score3` para o Quota, que é a linha em verde no gráfico. Mostrando que o protocolo Quota pode alcançar energias mais baixas. Para as proteínas 1ZDD e 2MR9, as linhas do gráfico apresentam pouquíssima diferença. Vale notar que para estas proteínas não houveram quaisquer diferenças estatisticamente significativas para RMSD e GDT. Ainda em comparação aos resultados da Tabela 6, falando estatisticamente, é interessante notar

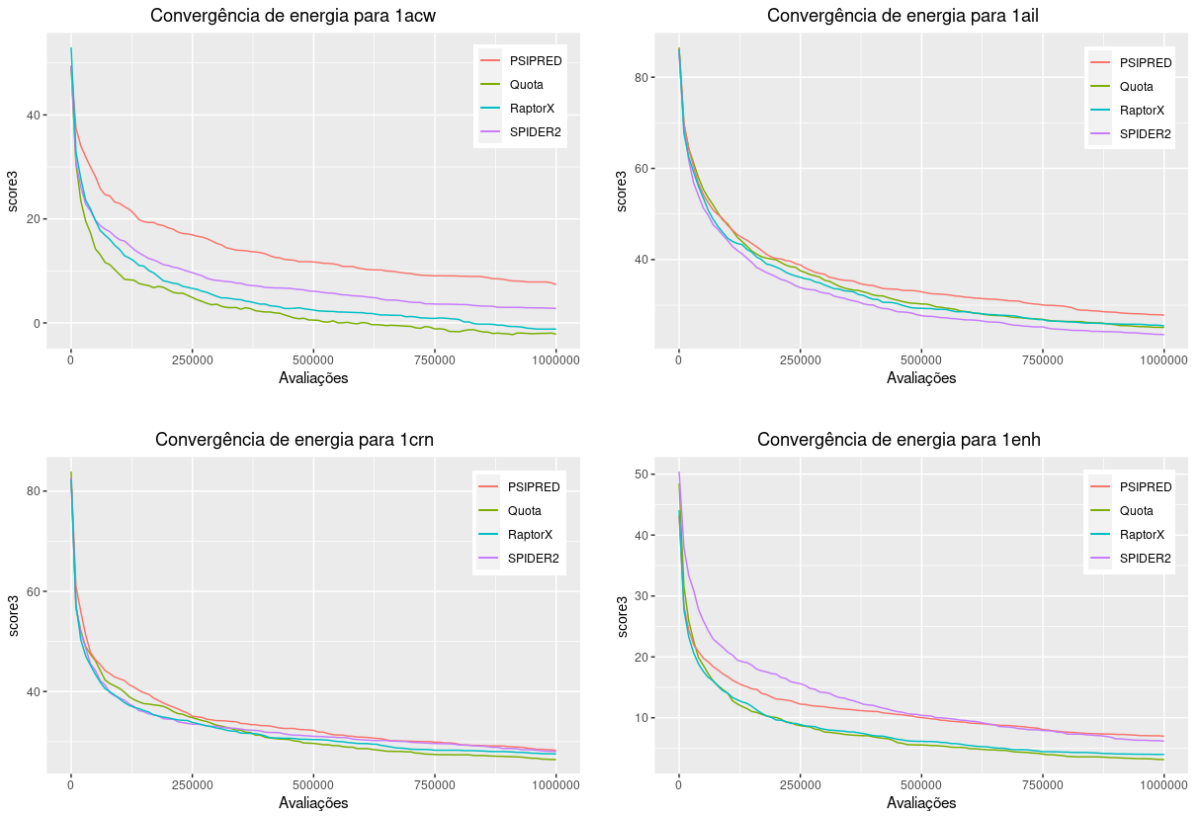


Figura 26 – Análise de convergência para `score3` das proteínas 1ACW, 1AIL, 1CRN e 1ENH

que os comportamentos no gráfico são bem diferentes para as duas proteínas que tiveram maiores diferenças significativas, 1ACW e 1CRN. Para a proteína 1ACW é notável a diferença entre as linhas que representam as bibliotecas no gráfico da Figura 26. Já para 1CRN as diferenças no gráfico são muito pequenas. Isso mostra que as proteínas não se comportam da mesma forma com relação ao valor de energia.

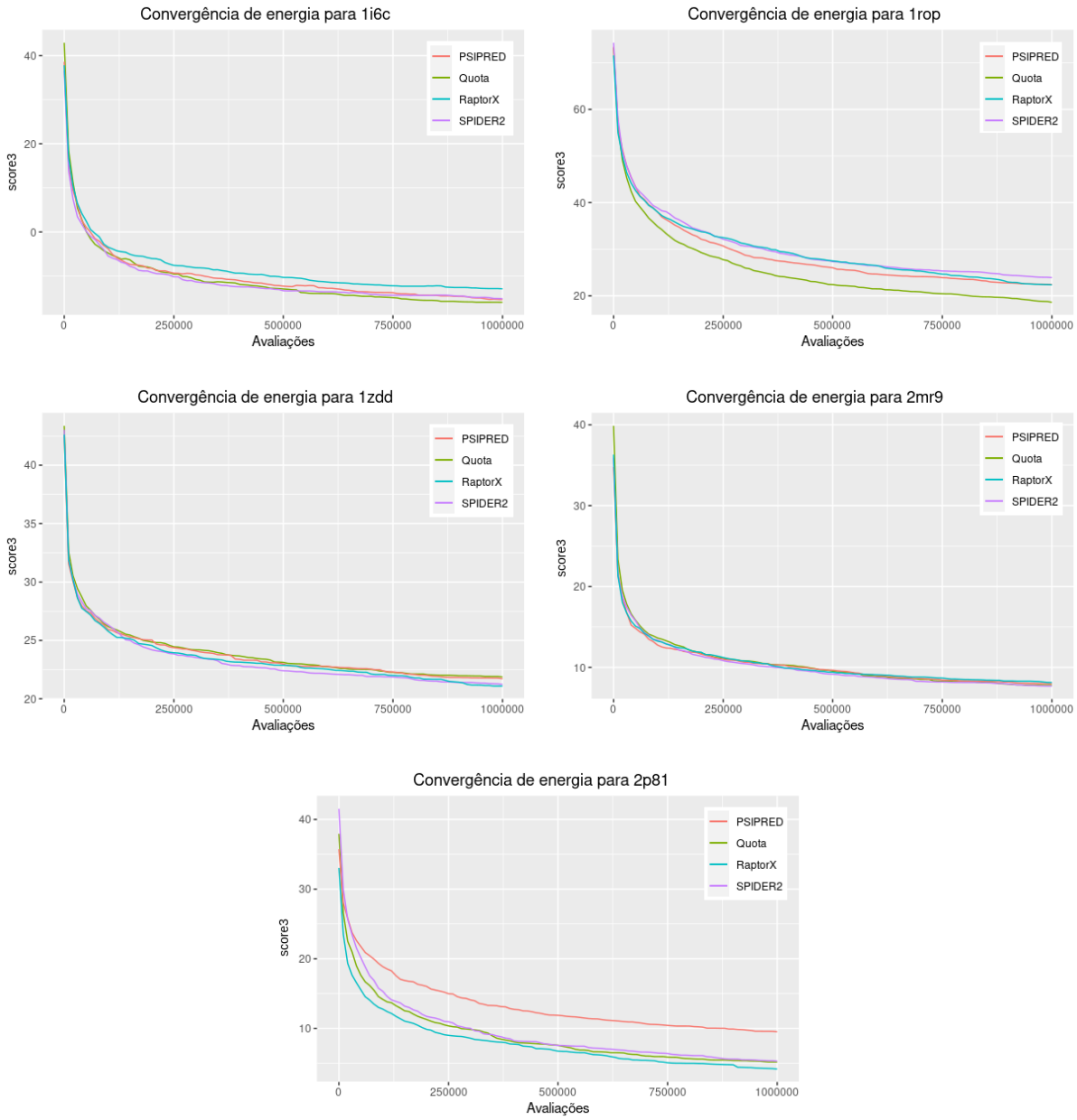


Figura 27 – Análise de convergência para `score3` das proteínas 1I6C, 1ROP, 1ZDD, 2MR9 e 2P81

5.4 ANÁLISE DO TEMPO DE PROCESSAMENTO

Em relação ao tempo de processamento do método proposto, não foi considerado o tempo da fase de Inicialização, pois é uma fase de pré-processamento, sendo executado apenas uma vez. O tempo calculado considera a geração da população inicial, a execução do ciclo evolutivo até o limite de iterações e a execução do procedimento de Repacking ao final do ciclo. A Tabela 7 mostra o tempo médio em segundos e o desvio padrão das 30 execuções para cada proteína. Os números em negrito representam o melhor tempo de processamento entre todas as bibliotecas. Como esperado, o tempo de processamento aumenta conforme o tamanho da proteína. Isso devido ao aumento da complexidade, o que impacta principalmente nas funções de avaliação e

na quantidade de comparações do crossover.

Tabela 7 – Tempo de processamento, em segundos, para os protocolos Best e Quota

PDB-ID	Tam.	Quota	PSIPRED	SPIDER2	RaptorX
1ACW	29	789.77 ± 61.31	$\mathbf{654.32 \pm 55.19}$	715.11 ± 56.30	713.31 ± 56.58
1ZDD	34	$\mathbf{1161.37 \pm 101.86}$	1241.61 ± 113.96	1214.00 ± 95.42	1273.68 ± 93.91
1I6C	39	$\mathbf{1197.06 \pm 88.35}$	1262.36 ± 93.90	1285.28 ± 95.71	1229.23 ± 86.17
2MR9	44	$\mathbf{2482.13 \pm 192.99}$	2664.29 ± 198.52	2567.04 ± 175.26	2627.04 ± 199.77
2P81	44	1816.43 ± 102.35	$\mathbf{1701.30 \pm 127.86}$	1754.23 ± 140.66	2244.46 ± 129.18
1CRN	46	1665.22 ± 125.85	$\mathbf{1572.26 \pm 116.95}$	1791.84 ± 131.47	1764.92 ± 137.86
1ENH	54	3190.09 ± 188.75	3796.91 ± 266.94	$\mathbf{2575.16 \pm 186.19}$	3699.46 ± 259.65
1ROP	63	$\mathbf{3294.89 \pm 225.82}$	3554.77 ± 232.23	3582.30 ± 247.02	3736.52 ± 240.23
1AIL	73	$\mathbf{4255.97 \pm 283.90}$	4700.93 ± 323.00	5140.04 ± 328.34	5029.89 ± 329.84

Outra observação, é a diferença nos tempos das proteínas 2MR9, 2P81 e 1CRN. As 3 proteínas tem tamanhos praticamente idênticos, mas a proteína 2MR9 teve um tempo de processamento bem maior que 2P81 e 1CRN. Isso significa que a proteína 2MR9 precisou de mais iterações para gastar 1000000 de avaliações. Sabendo que o número de avaliações é fixo para todas as proteínas, e que a etapa de mutação tem até 100 avaliações disponíveis por indivíduo a cada geração. Quanto menos avaliações forem gastas na mutação, mais iterações são necessárias para gastar o orçamento. A etapa de especiação demanda certo tempo de processamento, logo, quanto mais iterações tiver, esta etapa acabará influenciando no tempo total.

Com relação ao desempenho dos protocolos, ambos tiveram médias muito próximas. Das 9 proteínas, o protocolo Quota obteve a menor média para 5 delas. Em comparação com os resultados alcançados para RMSD e GDT, o tempo de processamento não parece ter relação com o que foi mostrado na Tabela 6. Por exemplo, para a proteína 1ACW, o PSIPRED obteve o menor tempo, enquanto que para RMSD e GDT ele obteve as piores médias. Já para as proteínas 1ZDD, 1I6C, 2MR9, 1ROP e 1AIL, que apresentaram o menor tempo para o Quota, foram as proteínas que não apresentaram diferenças significativas para RMSD e/ou GDT. Assim, acredita-se que o tempo de processamento está mais relacionado a quantidade de iterações necessárias para gastar o número de avaliações.

5.5 COMPARAÇÃO COM MÉTODOS DA LITERATURA

A Tabela 8 apresenta uma comparação do método proposto com diversos trabalhos competitivos da literatura. Os números em negrito representam os melhores resultados absolutos entre todos os métodos. Visto que o protocolo Quota obteve os melhores resultados gerais em comparação com o protocolo Best (ver Tabela 6), então decidiu-se por colocar na tabela apenas os resultados obtidos com o protocolo Quota. As linhas representam as proteínas, organizadas em ordem lexicográfica. A primeira coluna representa o ID da proteína no PDB. A segunda coluna representa os resultados obtidos pelo método proposto com o protocolo Quota. As demais colunas representam os métodos da literatura, ordenados por ano de publicação, sendo a esquerda

os métodos mais recentes. Na terceira coluna estão os valores obtidos pelo método Rosetta *ab initio*. Da quarta a décima coluna são métodos de trabalhos da literatura, sendo MO-BRKGA de (MARCHI; PARPINELLI, 2021), MABC de (CORREA; DORN, 2020), SCDE de (ZHANG et al., 2020), PPF-MC de (SILVA, 2019), LUE de (HAO; ZHANG; ZHOU, 2018), DPDE de (ZHANG et al., 2017) e ACUE de (HAO et al., 2016).

Tabela 8 – Comparação do RMSD com métodos da literatura dos últimos 5 anos

PDB-ID	Quota	Rosetta	MO-BRKGA	MABC	SCDE	PPF-MC	LUE	DPDE	ACUE
1ACW	1.74	1.77	2.70	**1.43**	-	4.45	-	-	-
1AIL	2.53	6.09	2.52	2.72	2.67	4.26	3.03	**1.58**	3.39
1CRN	**0.85**	5.07	1.36	2.31	-	4.18	-	-	-
1ENH	1.45	3.17	3.20	1.79	**1.12**	2.65	1.43	1.32	1.30
1I6C	2.97	6.92	2.90	-	2.69	-	3.96	**2.82**	3.52
1ROP	**0.94**	6.80	1.02	1.51	-	2.18	-	-	-
1ZDD	**0.85**	2.45	1.31	1.34	-	1.07	-	-	-
2MR9	**1.41**	2.35	1.76	1.55	-	1.66	-	-	-
2P81	3.66	6.60	5.77	**3.30**	-	-	-	-	-

Analisando a Tabela 8, observa-se que o RMSD tem diminuido com o passar do tempo, se tornando cada vez mais competitivo. Comparando com o que foi apresentado no Capítulo 3, nota-se que os métodos tem incorporado cada vez mais informações do problema, o que também justifica a melhora nos resultados do RMSD. Na tabela, destacam-se os métodos DPDE, SCDE, MABC e o método proposto neste trabalho, que apresentaram os melhores resultados em negrito. Com relação ao método proposto com o protocolo Quota, das 9 proteínas, o método obteve o RMSD mais baixo para 4 proteínas, con centrando a maior quantidade de melhores resultados. Isso mostra que o método proposto tem potencial para competir com outros métodos da literatura.

Para uma análise mais aprofundada dos resultados obtidos pelo método proposto, convém analisar os resultados estatisticamente com relação a RMSD e GDT. Para isto, a Tabela 9 mostra uma comparação do método proposto com o protocolo Quota a outros 3 métodos da literatura. O protocolo Rosetta *ab initio* é uma abordagem de objetivo único e *benchmark* para teste de proteínas. Já os métodos MO-BRKGA e MABC são os métodos mais recentes apresentado na Tabela 8 e trazem o mesmo conjunto de proteínas para teste, além de serem os mais competitivos. O MO-BRKGA utiliza uma abordagem multi-objetivo com informação de ES, fragmentos e mapas de contato. O MABC utiliza uma algoritmo memético com população estruturada e uso de informação de ES. Todos os algoritmos foram avaliados seguindo o mesmo protocolo com 30 execuções independentes e um orçamento fixo de 1000000 de avaliações. Os resultados do Rosetta e MO-BRKGA vem do artigo de (MARCHI; PARPINELLI, 2021). Os resultados do MABC vem do artigo de (CORREA; DORN, 2020). Na Tabela 9 são apresentados os valores de RMSD e GDT, sendo o melhor valor encontrado nas 30 execuções, a média e o desvio padrão.

Para determinar se os resultados da Tabela 9 são estatisticamente diferentes, foi realizada uma análise estatística utilizando ANOVA com o Tukey HSD post-hoc para cada proteína. Foi

determinado um nível de confiança de 95%. ANOVA é realizado utilizando os valores de média e desvio padrão, para RMSD e GDT separadamente. Na Tabela 9, os resultados com diferença significativa estão em negrito preto. Os melhores resultados gerais estão em negrito vermelho. As duas últimas linhas apresentam um resumo dos resultados obtidos. Na linha B/S/W, B é o número de instâncias do problema onde Quota foi melhor estatisticamente que aos demais algoritmos, S ele foi semelhante e W ele foi pior. Essa linha considera apenas a média e desvio padrão. Na linha B* é apresentado o número de instâncias do problema onde cada algoritmo obteve o melhor resultado geral.

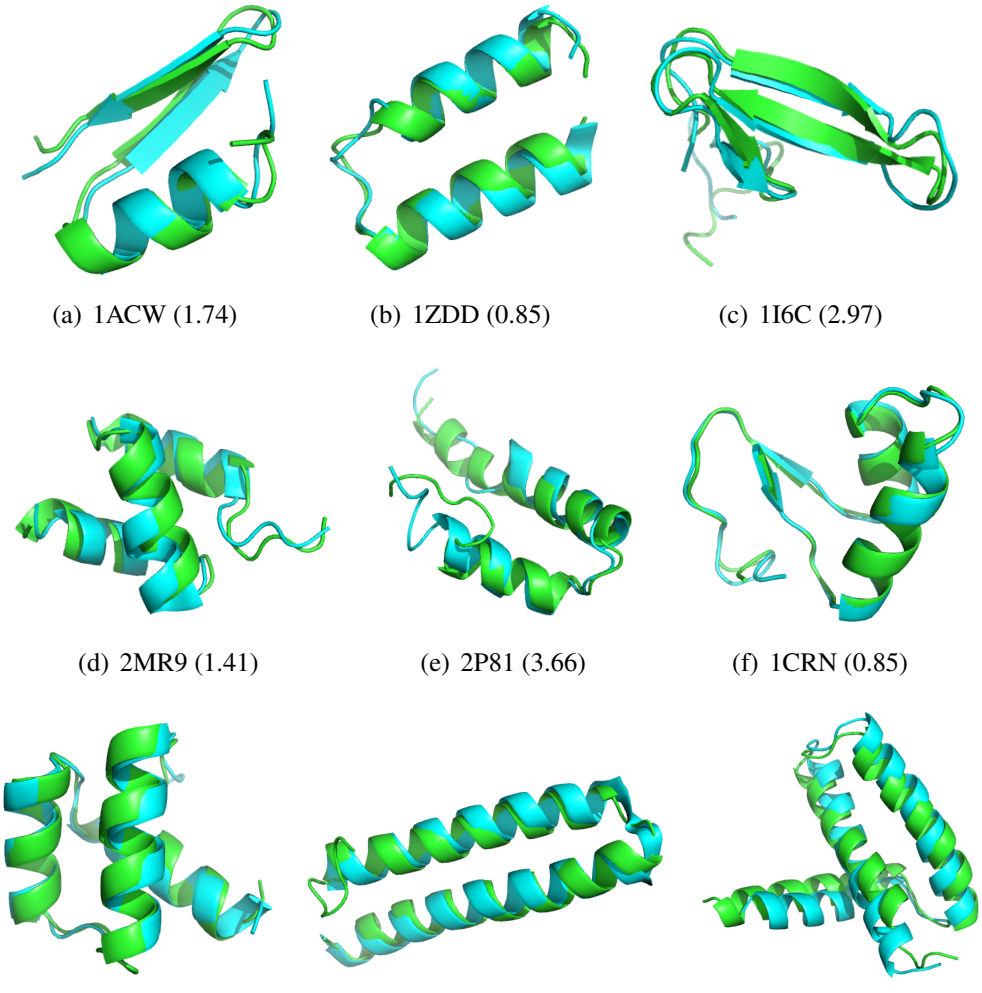
Em relação ao protocolo Rosetta *ab initio*, o método proposto foi superior tanto na média quanto na melhor conformação para todas as proteínas, tanto para RMSD quanto para GDT. Em relação ao MO-BRKGA e MABC, considerando o RMSD, o método proposto obteve a melhor média para pelo menos 3 proteínas, a 1ZDD, 1CRN e 1AIL; a pior média para no máximo 3 proteínas, a 1ACW, I16C e 2P81, sendo que destas ele é superior ao MO-BRKGA para 1ACW e 2P81; estatisticamente equivalente para as demais proteínas. Considerando o GDT, o método MABC não obteve a melhor média em nenhuma das proteínas, sendo no máximo estatisticamente equivalente aos demais métodos. Dessa forma, o método proposto teve a pior média para GDT apenas para as proteínas 1I6C e 2P81, sendo superior ou equivalente para as demais proteínas. Note que na linha B/S/W, os somatórios referentes ao MABC não somam um total de 9 proteínas. Isto porque o método MABC não realizou o teste para 1I6C, logo esta proteína não foi considerada na comparação com o método proposto. Considerando apenas os melhores resultados gerais na linha B* da tabela, o método obteve o RMSD mais baixo em 5 proteínas (1ZDD, 2MR9, 1CRN, 1ENH e 1ROP), e o GDT mais alto para 6 proteínas (1ACW, 1ZDD, 1I6C, 1CRN, 1ENH e 1AIL). Os resultados indicam que o método proposto é competitivo com métodos da literatura e pode alcançar bons ou até melhores resultados. Porém, os valores referentes ao desvio padrão não são tão baixos quando comparados aos outros métodos, o que indica maior dispersão dos dados.

Tabela 9 – Comparação com métodos da literatura

PDB-ID	Algoritmo	RMSD		GDT	
		Best	Média±DP	Best	Média±DP
1ACW	Rosetta	1.77	5.91 ± 1.43	84.83	44.76 ± 13.65
	MABC	**1.43**	**1.90 ± 0.23**	81.03	**75.43 ± 2.80**
	MO-BRKGA	2.70	3.95 ± 0.44	72.41	60.37 ± 4.71
	Método proposto	1.74	3.15 ± 0.77	**86.90**	**70.80 ± 9.70**
1ZDD	Rosetta	2.45	4.95 ± 0.90	80.59	56.97 ± 9.39
	MABC	1.34	1.95 ± 0.34	45.59	44.36 ± 0.67
	MO-BRKGA	1.31	1.65 ± 0.24	96.47	**90.83 ± 2.74**
	Método proposto	**0.85**	**1.19 ± 0.24**	**97.06**	**91.65 ± 3.33**
1I6C	Rosetta	6.92	8.43 ± 0.85	58.46	50.46 ± 4.63
	MABC	-	-	-	-
	MO-BRKGA	**2.90**	**3.71 ± 0.41**	73.85	**67.71 ± 3.13**
	Método proposto	2.97	4.42 ± 1.08	**74.36**	61.95 ± 7.28
2MR9	Rosetta	2.35	4.06 ± 1.94	80.45	66.88 ± 11.45
	MABC	1.55	2.08 ± 0.29	85.23	73.92 ± 3.90
	MO-BRKGA	1.76	2.22 ± 0.28	**91.82**	**85.75 ± 2.50**
	Método proposto	**1.41**	2.25 ± 0.51	87.73	**83.15 ± 4.90**
2P81	Rosetta	6.60	8.02 ± 0.91	62.73	54.94 ± 4.70
	MABC	**3.30**	**4.45 ± 0.57**	39.20	37.58 ± 0.76
	MO-BRKGA	5.77	6.95 ± 0.68	**73.64**	**66.25 ± 2.24**
	Método proposto	3.66	4.98 ± 0.53	**73.64**	61.64 ± 3.87
1CRN	Rosetta	5.07	7.59 ± 1.16	47.83	39.51 ± 5.25
	MABC	2.31	3.69 ± 0.73	76.09	67.68 ± 4.41
	MO-BRKGA	1.36	2.59 ± 0.79	93.04	**87.12 ± 2.88**
	Método proposto	**0.85**	**1.58 ± 0.62**	**97.39**	**88.94 ± 7.22**
1ENH	Rosetta	3.17	5.50 ± 1.51	79.26	57.99 ± 10.65
	MABC	1.79	**2.76 ± 0.38**	50.46	46.90 ± 1.96
	MO-BRKGA	3.20	3.59 ± 0.24	81.11	76.40 ± 2.84
	Método proposto	**1.45**	**2.36 ± 0.47**	**96.30**	**86.94 ± 6.91**
1ROP	Rosetta	6.80	10.25 ± 1.78	46.07	37.46 ± 4.16
	MABC	1.51	1.85 ± 0.23	83.48	78.23 ± 2.96
	MO-BRKGA	1.02	2.14 ± 0.75	**93.93**	81.56 ± 7.95
	Método proposto	**0.94**	2.11 ± 0.55	92.86	77.20 ± 8.47
1AIL	Rosetta	6.09	9.70 ± 1.63	40.86	28.84 ± 4.49
	MABC	2.72	5.33 ± 1.44	68.57	**57.42 ± 5.05**
	MO-BRKGA	**2.52**	5.42 ± 1.13	69.14	52.51 ± 6.62
	Método proposto	2.53	**4.28 ± 1.48**	**83.14**	**58.26 ± 11.18**
B/S/W	Rosetta		9/0/0		9/0/0
	MABC		3/3/2		5/3/0
	MO-BRKGA		6/2/1		3/4/2
	Método proposto		-		-
B*	Rosetta		0		0
	MABC		2		0
	MO-BRKGA		2		2
	Método proposto		5		6

5.6 REPRESENTAÇÃO VISUAL DAS PROTEÍNAS

A Figura 28 mostra a conformação visual da melhor solução encontrada para cada proteína, considerando o valor de RMSD. A análise visual das conformações é uma etapa importante na avaliação dos resultados, pois mostra onde estão as maiores dificuldades do método de previsão. A partir dela, é possível ver se as ES estão sendo previstas corretamente e sua posição no espaço 3D. Na imagem, em azul a conformação nativa e em verde a conformação prevista. Em parênteses, o valor do RMSD da melhor conformação. Com relação a estrutura secundária, em todos os casos, as estruturas foram previstas. As proteínas 1I6C, 2P81 e 1AIL tiveram os piores valores de RMSD entre as 9 proteínas. Isso é visível também em sua representação 3D. Na proteína 1I6C, tanto alças quanto folhas estão mal posicionadas e com tamanhos incorretos. Na proteína 2P81, as hélices estão com tamanhos bem diferentes da estrutura nativa. A proteína 1AIL também apresenta tamanhos incorretos para as hélices.



(a) 1ACW (1.74) (b) 1ZDD (0.85) (c) 1I6C (2.97)

(d) 2MR9 (1.41) (e) 2P81 (3.66) (f) 1CRN (0.85)

(g) 1ENH (1.45) (h) 1ROP (0.94) (i) 1AIL (2.53)

Figura 28 – Comparação entre as conformações previstas (em verde) e nativa (em azul)

As proteínas 1ZDD, 1CRN e 1ROP apresentaram os melhores valores de RMSD entre as 9 proteínas, todas abaixo de 1Å. Com destaque para as proteínas 1ZDD e 1CRN que apresentaram

o menor valor de RMSD, ambas com 0.85Å. Suas estruturas estão muito bem ajustadas, inclusive as do tipo voltas, que são mais difíceis de serem ajustadas. De maneira geral, todas as proteínas analisadas apresentaram uma boa conformação visual. Isso indica que o método proposto tem potencial para encontrar soluções bem estruturadas para proteínas menores.

6 CONCLUSÕES

As proteínas estão presentes em todos os seres vivos e realizam funções essenciais a vida. Estas funções estão diretamente relacionadas a sua estrutura tridimensional. Prever a estrutura 3D de uma proteína pode fornecer informações a respeito de sua funcionalidade, o que é fundamental no desenvolvimento de medicamentos e tratamentos para a saúde. Nos últimos anos, muitos esforços têm sido dedicados ao desenvolvimento de métodos computacionais para a previsão de estrutura de proteína. Mas apesar de todo o avanço tecnológico e dos métodos computacionais, ainda não existe uma solução ótima para o problema do PSP. Neste sentido, uma classe de métodos tem ganhado destaque nos últimos anos, por conseguir reduzir a complexidade envolvida através da utilização de informações do problema, são chamados de métodos *de novo*.

Estes métodos costumam utilizar informações do problema através de fragmentos de proteínas já conhecidas, previsões de estrutura secundária, mapas de contato e dinâmica molecular. Claramente, a utilização destas técnicas tem se tornado cada vez mais comum no PSP, pela redução da complexidade, auxiliando na compreensão do problema e orientação na busca de conformações mais adequadas. Mas apesar de toda essa evolução ainda existem problemas. A utilização de fragmentos de proteína requer uma biblioteca de fragmentos, que por sua vez utiliza informação de ES para seleção dos fragmentos. Para isto é necessário o uso de preditores de ES. Mas não só pelo uso de fragmentos, como também pela própria ES em si. Isso tem tornado estes métodos dependentes da previsão dos preditores de ES. Assim, uma vez que uma ES incorreta seja prevista, ela pode levar a má formações de estruturas tridimensionais. Além disso, existe o problema da relação entre função de energia x RMSD, pois nem sempre uma baixa energia representa a melhor conformação, ou seja, o RMSD mais baixo. Por isso, orientar a busca apenas com base na função de energia pode ser muito insatisfatório.

Em vista disso, neste trabalho é proposto um método de otimização para o problema do PSP, o qual se encaixa na abordagem *de novo*. O método proposto implementa um algoritmo evolutivo que usa uma técnica de especiação dinâmica, chamada DST, para agrupar indivíduos da população de acordo com a distância euclidiana. O método utiliza 3 técnicas para inserir informação do problema, são elas: fragmentos, estrutura secundária e mapas de contato. O uso da DST e a mutação baseada em fragmentos, auxiliam na manutenção da diversidade da população. Além disso, a biblioteca de fragmentos foi gerada com base no protocolo Quota do Rosetta, que tem por objetivo gerar fragmentos com maior diversidade. Para isso, foram utilizados 3 preditores de ES, PSIPRED, SPIDER2 e RaptorX. O que ajuda a minimizar a dependência dos preditores de ES e evitar possíveis erros de previsões incorretas.

Já as informações de ES e mapas de contato, foram fornecidas pelo preditor RaptorX, e utilizadas para orientar a busca pelo espaço conformacional. No crossover, uma estratégia baseada na ES orienta a troca de informação genética entre os indivíduos. Duas estratégias de seleção, uma baseada em estrutura e outra baseada em contato, são utilizadas para selecionar indivíduos mais aptos para a próxima geração. Um mecanismo auto-adaptativo, baseado na ES

da população atual, fornece informações para a escolha entre as duas estratégias. Isso favorece um equilibrio entre contato e estrutura, e evita uma possível dependência dos preditores. Esta ideia tenta minimizar os efeitos do segundo problema citado acima, de tentar relacionar um baixo RMSD com uma baixa energia.

O método proposto foi analisado a partir de um experimento realizado com 9 proteínas: 1ACW, 1AIL, 1CRN, 1ENH, 1I6C, 1ROP, 1ZDD, 2MR9 e 2P81. Para comparar o desempenho do protocolo Quota com relação ao protocolo Best, foram geradas bibliotecas de fragmentos para ambos os protocolos. Ao total foram 4 bibliotecas de fragmentos para cada proteína, sendo 1 biblioteca para o protocolo Quota considerando os 3 preditores de ES juntos, e 3 bibliotecas para o protocolo Best, sendo 1 para cada preditor individualmente. Assim, os resultados para cada proteína puderam ser comparados quanto a biblioteca utilizada. Para os experimentos foram utilizados os preditores de ES: PSIPRED, SPIDER2 e RaptorX. Para o protocolo Quota, foi considerado a participação de 40% para o RaptorX, 30% para o SPIDER2 e 30% para o PSIPRED. Os protocolos foram comparados em termos de RMSD, GDT, convergência de energia e tempo de processamento. Além disso, os resultados de RMSD e GDT foram analisados estatisticamente com o teste ANOVA.

Tanto para RMSD quanto GDT, o teste ANOVA indicou que o protocolo Quota obteve a melhor média estatisticamente para pelo menos 2 das 9 proteínas, sendo semelhante nas demais. Com relação ao melhor resultado geral, o protocolo Quota obteve o melhor resultado para 6 das 9 proteínas, considerando o RMSD, e 4 proteínas considerando o GDT. Isso mostra que o protocolo Quota pode obter boas médias ou até melhores que o protocolo Best, mas além disso pode alcançar melhores resultados para proteínas pequenas. A análise de convergência de energia mostrou que o Quota obteve a melhor média do `score3` para 5 das 9 proteínas. Da mesma forma, a análise do tempo de processamento mostrou que o Quota obteve a menor média de tempo para 5 proteínas. Como o tempo de processamento está relacionado a quantidade de iterações do algoritmo e a forma como são gastas as avaliações, isso mostra que o protocolo Quota necessita de menos iterações, conseguindo chegar aos mesmos resultados ou até melhores que o protocolo Best.

O desempenho do método proposto também foi analisado com relação a outros trabalhos da literatura. Os resultados obtidos para o RMSD e GDT com o protocolo Quota, foram comparados estatisticamente através do teste ANOVA com o método Rosetta `ab initio`, MO-BRKGA e MABC. Em termos de RMSD, o método proposto alcançou melhores resultados estatisticamente para pelo menos 3 proteínas, e piores resultados para no máximo 3 proteínas, sendo equivalente para as demais. Em termos de GDT, o método foi melhor estatisticamente para pelo menos 3 proteínas, sendo pior apenas para 2 proteínas e equivalente para as demais. Com relação ao melhor resultado geral, o método proposto obteve o RMSD mais baixo para 5 proteínas, e o GDT mais alto para 6 proteínas. Isso mostra que o método proposto é competitivo com outros métodos da literatura. Uma análise visual das conformações 3D mostra que o método foi eficaz nos resultados encontrados para proteínas pequenas, gerando conformações com estruturas

secundárias bem formadas. O uso de 3 preditores aumentou a possibilidade de identificar diferentes estruturas. Logo, é possível supor que o uso de 3 preditores para geração dos fragmentos aliado a previsão de ES e aos mapas de contato, reduziram a dependência dos preditores de ES, impactando diretamente nas conformações 3D previstas.

Como trabalhos futuros, é interessante avaliar o comportamento do método para proteínas maiores e proteínas do CASP. Mas para isto, talvez seja necessário reduzir o tempo de processamento. Uma opção seria o uso de arquiteturas de computação de alto desempenho, como GPUs (*Graphics Processing Units*). Outra opção seria adaptar o mecanismo de DST para que não seja necessário sua execução a cada geração. Outra direção de pesquisa, seria analisar o impacto na energia mediante as estratégias de seleção empregadas. Também pode ser interessante a implementação de um mecanismo para medir a diversidade da população, isso seria útil para saber se o protocolo Quota realmente proporciona maior diversidade com relação ao Best. Como alternativa para trabalho futuro, uma possível solução seria realizar a participação auto-adaptativa dos preditores empregados.

O tabalho desenvolvido foi publicado no Hybrid Intelligent Systems, 2020 (WILL; PARPINELLI, 2021).

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

REFERÊNCIAS

ALMEIDA, Alexandre Barbosa de. **Predição de Estrutura Terciária de Proteínas com Técnicas Multiobjetivo no Algoritmo de Monte Carlo**. Dissertação (Mestrado) — Universidade Federal de Goiás - UFG, Goiânia, 2016. Citado na página 31.

ANFINSEN, Christian B. Principles that govern the folding of protein chains. **Science**, American Association for the Advancement of Science, v. 181, n. 4096, p. 223–230, 1973. ISSN 0036-8075. Citado 4 vezes nas páginas 16, 24, 28 e 30.

BOIANI, Mateus. **A GPU-Based Hybrid jDE Algorithm Applied to the Protein Structure Prediction Problem**. Dissertação (Mestrado) — Universidade do Estado de Santa Catarina - UDESC, 2019. Citado na página 30.

BORGUESAN, Bruno et al. Apl: An angle probability list to improve knowledge-based metaheuristics for the three-dimensional protein structure prediction. **Computational Biology and Chemistry**, v. 59, p. 142 – 157, 2015. ISSN 1476-9271. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1476927115301250>. Citado na página 23.

BRANDEN, C.; TOOZE, J. **Introduction to protein structure**. 2. ed. New York, USA: Garland Science, 1999. Citado na página 21.

BREST, Janez et al. Self-adapting control parameters in differential evolution: A comparative study on numerical benchmark problems. **IEEE Transactions on Evolutionary Computation**, v. 10, n. 6, p. 646–657, 2006. Citado 2 vezes nas páginas 32 e 56.

BROOKS, B. R. et al. Charmm: The biomolecular simulation program. **Journal of Computational Chemistry**, v. 30, n. 10, p. 1545–1614, 2009. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.21287>. Citado na página 30.

BUCHAN, Daniel W A; JONES, David T. The PSIPRED Protein Analysis Workbench: 20 years on. **Nucleic Acids Research**, v. 47, n. W1, p. W402–W407, 04 2019. ISSN 0305-1048. Disponível em: <https://doi.org/10.1093/nar/gkz297>. Citado na página 37.

CALLAWAY, Ewen. The revolution will not be crystallized: a new method sweeps through structural biology. **Nature 525**, Springer Nature, p. 172–174, 2015. Citado na página 26.

CARDOSO, Marcos Borba. **Uma proposta para a predição computacional da estrutura terciária de polipeptídeos**. Dissertação (Mestrado) — Pontifícia Universidade Católica do Rio Grande do Sul, 2007. Citado na página 14.

CHEN, Xingqian et al. Incorporating a multiobjective knowledge-based energy function into differential evolution for protein structure prediction. **Information Sciences**, v. 540, p. 69 – 88, 2020. ISSN 0020-0255. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0020025520305703>. Citado 2 vezes nas páginas 44 e 46.

COMBS, Steven et al. Small-molecule ligand docking into comparative models with rosetta. **Nature protocols**, v. 8, p. 1277–98, 06 2013. Citado na página 17.

CORREA, Leonardo de Lima; DORN, Márcio. A knowledge-based artificial bee colony algorithm for the 3-d protein structure prediction problem. In: **2018 IEEE Congress on Evolutionary Computation (CEC)**. [S.l.: s.n.], 2018. p. 1–8. Citado 5 vezes nas páginas 15, 26, 31, 43 e 46.

CORREA, Leonardo de Lima; DORN, Márcio. A multi-population memetic algorithm for the 3-d protein structure prediction problem. **Swarm and Evolutionary Computation**, v. 55, p. 100677, 2020. ISSN 2210-6502. Citado 12 vezes nas páginas 16, 21, 23, 24, 26, 28, 30, 42, 46, 47, 54 e 70.

CORREA, Leonardo de Lima; INOSTROZA-PONTA, Mario; DORN, Márcio. An evolutionary multi-agent algorithm to explore the high degree of selectivity in three-dimensional protein structures. In: **2017 IEEE Congress on Evolutionary Computation (CEC)**. [S.l.: s.n.], 2017. p. 1111–1118. Citado 2 vezes nas páginas 30 e 31.

DAS, S.; SUGANTHAN, P. N. Differential evolution: A survey of the state-of-the-art. **IEEE Transactions on Evolutionary Computation**, v. 15, n. 1, p. 4–31, Feb 2011. ISSN 1941-0026. Citado na página 32.

DENG, Libao et al. Dsm-de: a differential evolution with dynamic speciation-based mutation for single-objective optimization. **Memetic Computing**, 01 2019. Citado 7 vezes nas páginas 8, 32, 33, 34, 48, 54 e 61.

DHINGRA, Surbhi et al. A glance into the evolution of template-free protein structure prediction methodologies. **Biochimie**, p. 85–92, 02 2020. Citado 8 vezes nas páginas 8, 15, 16, 26, 27, 28, 35 e 52.

DORN, Márcio; BURIOL, Luciana S.; LAMB, Luis C. A hybrid genetic algorithm for the 3-d protein structure prediction problem using a path-relinking strategy. In: **2011 IEEE Congress of Evolutionary Computation (CEC)**. [S.l.: s.n.], 2011. p. 2709–2716. ISSN 1089-778X. Citado 5 vezes nas páginas 14, 31, 41, 46 e 47.

DORN, Márcio; BURIOL, Luciana S.; LAMB, Luis C. A molecular dynamics and knowledge-based computational strategy to predict native-like structures of polypeptides. **Expert Systems with Applications**, v. 40, n. 2, p. 698 – 706, 2013. ISSN 0957-4174. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0957417412009645>. Citado na página 45.

DORN, Márcio et al. Three-dimensional protein structure prediction. **Comput. Biol. Chem.**, Elsevier Science Publishers B. V., NLD, v. 53, n. PB, p. 251–276, dez. 2014. ISSN 1476-9271. Citado 7 vezes nas páginas 16, 20, 21, 24, 28, 29 e 31.

DORN, Márcio; SOUZA, Osmar Norberto de. Cref: A central-residue-fragment-based method for predicting approximate 3-d polypeptides structures. In: **Proceedings of the 2008 ACM Symposium on Applied Computing**. New York, NY, USA: Association for Computing Machinery, 2008. (SAC '08), p. 1261–1267. ISBN 9781595937537. Citado 5 vezes nas páginas 8, 25, 34, 35 e 41.

DORN, Márcio; SOUZA, Osmar Norberto de. A3n: An artificial neural network n-gram-based method to approximate 3-d polypeptides structure prediction. **Expert Systems with Applications**, v. 37, n. 12, p. 7497 – 7508, 2010. ISSN 0957-4174. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0957417410003969>. Citado 2 vezes nas páginas 34 e 45.

DUNKER, A.Keith et al. Intrinsically disordered protein. **Journal of Molecular Graphics and Modelling**, v. 19, n. 1, p. 26 – 59, 2001. ISSN 1093-3263. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1093326300001388>. Citado na página 24.

EICHENBERGER, Andreas P. et al. Gromos++ software for the analysis of biomolecular simulation trajectories. **Journal of Chemical Theory and Computation**, v. 7, n. 10, p. 3379–3390, 2011. Citado na página 30.

EMERSON, Isaac Arnold; AMALA, Arumugam. Protein contact maps: A binary depiction of protein 3d structures. **Physica A: Statistical Mechanics and its Applications**, v. 465, p. 782 – 791, 2017. ISSN 0378-4371. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0378437116305507>. Citado 2 vezes nas páginas 8 e 39.

FANG, Chao; SHANG, Yi; XU, Dong. Mufold-ss: New deep inception-inside-inception networks for protein secondary structure prediction. **Proteins: Structure, Function, and Bioinformatics**, v. 86, n. 5, p. 592–598, 2018. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.25487>. Citado na página 37.

GABRIEL, Paulo H. R.; MELO, Vinícius V. de; DELBEM, Alexandre C. B. Algoritmos evolutivos e modelo HP para predição de estruturas de proteínas. **Sba: Controle Automação Sociedade Brasileira de Automatica**, scielo, v. 23, p. 25 – 37, 02 2012. ISSN 0103-1759. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-17592012000100003&nrm=iso>. Citado na página 30.

GARRETT, R.; GRISHAM, C. M. **Biochemistry**. New York, USA: Saunder's College Publishing, 1999. Citado 2 vezes nas páginas 8 e 23.

GARZA-FABRE, Mario et al. Generating, maintaining, and exploiting diversity in a memetic algorithm for protein structure prediction. **Evol. Comput.**, MIT Press, Cambridge, MA, USA, v. 24, n. 4, p. 577–607, dez. 2016. ISSN 1063-6560. Citado 7 vezes nas páginas 8, 19, 20, 34, 38, 42 e 46.

GEOURJON, C.; DELéAGE, G. SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. **Bioinformatics**, v. 11, n. 6, p. 681–684, 12 1995. ISSN 1367-4803. Disponível em: <https://doi.org/10.1093/bioinformatics/11.6.681>. Citado na página 22.

GOES, Andréa Carla de Souza; OLIVEIRA, Bruno Vinicius Ximenes de. Projeto genoma humano: um retrato da construção do conhecimento científico sob a ótica da revista ciência hoje. **Ciência e Educação (Bauru)**, scielo, v. 20, p. 561 – 577, 09 2014. ISSN 1516-7313. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1516-73132014000300561&nrm=iso>. Citado na página 26.

GRONT, Dominik et al. Generalized fragment picking in rosetta: Design, protocols and applications. **PLOS ONE**, Public Library of Science, v. 6, n. 8, p. 1–10, 08 2011. Disponível em: <https://doi.org/10.1371/journal.pone.0023294>. Citado 5 vezes nas páginas 8, 35, 36, 52 e 53.

HAO, Xiao-Hu; ZHANG, Gui-Jun. Double estimation of distribution guided sampling algorithm for de-novo protein structure prediction. In: **2017 36th Chinese Control Conference (CCC)**. [S.l.: s.n.], 2017. p. 9853–9858. Citado 2 vezes nas páginas 42 e 46.

HAO, Xiao-Hu; ZHANG, Gui-Jun; ZHOU, Xiao-Gen. Guiding exploration in conformational feature space with lipschitz underestimation for ab-initio protein structure prediction. **Computational Biology and Chemistry**, v. 73, p. 105 – 119, 2018. ISSN 1476-9271. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1476927117302451>. Citado 3 vezes nas páginas 42, 46 e 70.

HAO, Xiao-Hu et al. A novel method using abstract convex underestimation in ab-initio protein structure prediction for guiding search in conformational feature space. **IEEE/ACM Transactions on Computational Biology and Bioinformatics**, v. 13, n. 5, p. 887–900, 2016. Citado 6 vezes nas páginas 17, 34, 35, 42, 46 e 70.

HEFFERNAN, Rhys et al. Highly accurate sequence-based prediction of half-sphere exposures of amino acid residues in proteins. **Bioinformatics**, v. 32, n. 6, p. 843–849, 11 2015. ISSN 1367-4803. Disponível em: <https://doi.org/10.1093/bioinformatics/btv665>. Citado na página 37.

JOSHI, Rajani R.; JYOTHI, S. Ab-initio prediction and reliability of protein structural genomics by propainor algorithm. **Computational Biology and Chemistry**, v. 27, n. 3, p. 241 – 252, 2003. ISSN 1476-9271. Computers and Chemistry. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0097848502000748>. Citado na página 45.

JUMPER, John et al. High accuracy protein structure prediction using deep learning. **In Fourteenth Critical Assessment of Techniques for Protein Structure Prediction**, p. 1–1, 2020. Citado na página 17.

KAZMIER, Kelli et al. Algorithm for selection of optimized epr distance restraints for de novo protein structure determination. **Journal of Structural Biology**, v. 173, n. 3, p. 549 – 557, 2011. ISSN 1047-8477. Combining computational modeling with sparse and low-resolution data. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1047847710003400>. Citado 2 vezes nas páginas 41 e 46.

Khan Academy. **Introdução às proteínas e aos aminoácidos**. [S.l.], 2021. Accessed: fev-2021. Disponível em: <https://pt.khanacademy.org/science/biology/macromolecules/proteins-and-amino-acids/a/introduction-to-proteins-and-amino-acids>. Citado 3 vezes nas páginas 8, 21 e 22.

KäLLBERG, Morten et al. Template-based protein structure modeling using the raptorx web server. Nature Protocols, p. 1511–1522, 2012. Citado na página 38.

LEE, Julian; KIM, Seung-Yeon; LEE, Jooyoung. Protein structure prediction based on fragment assembly and parameter optimization. **Biophysical Chemistry**, v. 115, n. 2, p. 209 – 214, 2005. ISSN 0301-4622. BIFI 2004 International Conference Biology after the Genoma: A Physical View. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0301462204003801>. Citado 2 vezes nas páginas 34 e 41.

LODISH, Harvey et al. **Molecular cell biology**. Berlin, Germany: Macmillan, 2008. Citado 4 vezes nas páginas 8, 19, 22 e 24.

LOPES, Jhennefer N.; VENSKE, Sandra M. Predição da estrutura de proteínas utilizando algoritmo evolutivo adaptativo. In: Bastos Filho, C. J. A.; POZO, A. R.; LOPES, H. S. (Ed.). **Anais do 12 Congresso Brasileiro de Inteligência Computacional**. Curitiba, PR: ABRICOM, 2015. p. 1–6. Citado na página 14.

MARCHI, Felipe; PARPINELLI, Rafael Stubs. A multi-objective approach to the protein structure prediction problem using the biased random-key genetic algorithm. **CEC 2021 - Congress on Evolutionary Computation**, 2021. To be published. Citado 4 vezes nas páginas 16, 45, 46 e 70.

MIRABELLO, Claudio; POLLASTRI, Gianluca. Porter, PaleAle 4.0: high-accuracy prediction of protein secondary structure and relative solvent accessibility. **Bioinformatics**, v. 29, n. 16, p. 2056–2058, 06 2013. ISSN 1367-4803. Disponível em: <https://doi.org/10.1093/bioinformatics/btt344>. Citado na página 38.

MISHRA, Avdesh; HOQUE, Md Tamjidul. 3digars-psp: A novel statistical energy function and effective conformational search strategy based ab initio protein structure prediction. In: **2019 22nd International Conference on Computer and Information Technology (ICCIT)**. [S.l.: s.n.], 2019. p. 1–7. Citado 3 vezes nas páginas 27, 43 e 46.

NAMBA, A. M.; SILVA, V. B. da; SILVA, C. H. T. P. da. Dinâmica molecular: teoria e aplicações em planejamento de fármacos. **Eclética Química**, scielo, v. 33, p. 13 – 24, 12 2008. ISSN 0100-4670. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-46702008000400002&nrm=iso>. Citado na página 31.

NARLOCH, Pedro; PARPINELLI, Rafael. Diversification strategies in differential evolution algorithm to solve the protein structure prediction problem. In: . [S.l.: s.n.], 2017. p. 125–134. ISBN 978-3-319-53479-4. Citado na página 32.

NARLOCH, Pedro; PARPINELLI, Rafael. The protein structure prediction problem approached by a cascade differential evolution algorithm using rosetta. In: . [S.l.: s.n.], 2017. p. 294–299. Citado na página 29.

NELSON, D. L.; LEHNINGER, A. L.; COX, M. M. **Lehninger principles of biochemistry**. Berlin, Germany: Macmillan, 2008. Citado na página 26.

OpenStax. **Proteins**. [S.l.], 2021. Accessed: fev-2021. Disponível em: <https://cnx.org/contents/GFy_h8cu@9.85:2zzm1QG9@7/Proteins>. Citado 3 vezes nas páginas 8, 24 e 25.

PARPINELLI, Rafael et al. A review of techniques for online control of parameters in swarm intelligence and evolutionary computation algorithms. **International Journal of Bio-Inspired Computation**, v. 13, p. 1, 01 2019. Citado na página 17.

PENG, Chun-Xiang; ZHOU, Xiao-Gen; ZHANG, Gui-Jun. De novo protein structure prediction by coupling contact with distance profile. **IEEE/ACM Transactions on Computational Biology and Bioinformatics**, p. 1–1, 2020. Citado 4 vezes nas páginas 16, 40, 44 e 46.

POLLASTRI, Gianluca et al. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. **Proteins: Structure, Function, and Bioinformatics**, v. 47, n. 2, p. 228–235, 2002. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.10082>. Citado na página 38.

QIN, A. K.; SUGANTHAN, P. N. Self-adaptive differential evolution algorithm for numerical optimization. In: **2005 IEEE Congress on Evolutionary Computation**. [S.l.: s.n.], 2005. v. 2, p. 1785–1791 Vol. 2. Citado na página 32.

RCSB PDB. [S.l.], 2021. Accessed: fev-2021. Disponível em: <https://www.rcsb.org/>. Citado 2 vezes nas páginas 8 e 25.

ROCHA, Gregório K et al. A multiobjective approach for protein structure prediction using a steady-state genetic algorithm with phenotypic crowding. In: **2015 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)**. [S.l.: s.n.], 2015. p. 1–8. ISSN null. Citado na página 31.

ROCHA, Gregório K. et al. Inserting co-evolution information from contact maps into a multiobjective genetic algorithm for protein structure prediction. In: **2018 IEEE Congress on Evolutionary Computation (CEC)**. [S.l.: s.n.], 2018. p. 1–8. Citado na página 40.

Rosetta Commons. [S.l.], 2021. Accessed: fev-2021. Disponível em: <https://www.rosettacommons.org/>. Citado 2 vezes nas páginas 8 e 49.

RUIZ-BLANCO, Yasser B. et al. A physics-based scoring function for protein structural decoys: Dynamic testing on targets of casp-roll. **Chemical Physics Letters**, v. 610-611, p. 135 – 140, 2014. ISSN 0009-2614. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0009261414005892>. Citado na página 45.

SALOMON-FERRER, Romelia; CASE, David A.; WALKER, Ross C. An overview of the amber biomolecular simulation package. **WIREs Computational Molecular Science**, v. 3, n. 2, p. 198–210, 2013. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/wcms.1121>. Citado na página 30.

SANTOS, Karina B. et al. Improving de novo protein structure prediction using contact maps information. In: **2017 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)**. [S.l.: s.n.], 2017. p. 1–6. Citado 3 vezes nas páginas 40, 43 e 46.

SILVA, Renan Samuel da. **A Self Adaptive Greedy Evolutionary Algorithm Using Monte Carlo Fragment Insertion And Conformation Clustering**. Dissertação (Mestrado) — Universidade do Estado de Santa Catarina - UDESC, 2019. Citado 14 vezes nas páginas 8, 16, 19, 21, 23, 31, 35, 43, 46, 48, 49, 50, 52 e 70.

SMOLARCZYK, Tomasz et al. Protein secondary structure prediction: A review of progress and directions. **Current Bioinformatics**, v. 15, p. 90 – 107, 03 2020. Citado 4 vezes nas páginas 9, 37, 38 e 61.

STILLINGER, Frank H.; HEAD-GORDON, Teresa; HIRSHFELD, Catherine L. Toy model for protein folding. **Phys. Rev. E**, American Physical Society, v. 48, p. 1469–1477, Aug 1993. Disponível em: <https://link.aps.org/doi/10.1103/PhysRevE.48.1469>. Citado na página 30.

TANABE, R.; FUKUNAGA, A. Evaluating the performance of shade on cec 2013 benchmark problems. In: **2013 IEEE Congress on Evolutionary Computation**. [S.l.: s.n.], 2013. p. 1952–1959. Citado na página 32.

TORRISI, Mirko; KALEEL, Manaz; POLLASTRI, Gianluca. Porter 5: fast, state-of-the-art ab initio prediction of protein secondary structure in 3 and 8 classes. **bioRxiv**, Cold Spring Harbor Laboratory, 2018. Disponível em: <https://www.biorxiv.org/content/early/2018/10/05/289033>. Citado na página 38.

WANG, Guoli; DUNBRACK ROLAND L., Jr. PISCES: a protein sequence culling server. **Bioinformatics**, v. 19, n. 12, p. 1589–1591, 08 2003. ISSN 1367-4803. Disponível em: <https://doi.org/10.1093/bioinformatics/btg224>. Citado na página 35.

WEINER, Brian E. et al. Bcl::mp-fold: Folding membrane proteins through assembly of transmembrane helices. **Structure**, v. 21, n. 7, p. 1107 – 1117, 2013. ISSN 0969-2126. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0969212613001524>. Citado 2 vezes nas páginas 41 e 46.

WILL, Nilcimar Neitzel; PARPINELLI, Rafael Stubs. Comparing best and quota fragment picker protocols applied to protein structure prediction. In: ABRAHAM, Ajith et al. (Ed.). **Hybrid Intelligent Systems**. Cham: Springer International Publishing, 2021. p. 669–678. ISBN 978-3-030-73050-5. Citado na página 77.

ZHANG, Gui-Jun et al. Secondary structure and contact guided differential evolution for protein structure prediction. **IEEE/ACM Transactions on Computational Biology and Bioinformatics**, v. 17, n. 3, p. 1068–1081, 2020. Citado 8 vezes nas páginas 16, 44, 46, 56, 58, 59, 61 e 70.

ZHANG, Gui-Jun et al. Protein structure prediction using population-based algorithm guided by information entropy. **IEEE/ACM Transactions on Computational Biology and Bioinformatics**, p. 1–1, 2019. Citado 4 vezes nas páginas 16, 34, 44 e 46.

ZHANG, Gui-Jun et al. A population-based conformational optimal algorithm using replica-exchange in ab-initio protein structure prediction. In: **2016 Chinese Control and Decision Conference (CCDC)**. [S.l.: s.n.], 2016. p. 701–706. Citado 3 vezes nas páginas 35, 42 e 46.

ZHANG, Gui-Jun et al. Enhancing protein conformational space sampling using distance profile-guided differential evolution. **IEEE/ACM Transactions on Computational Biology and Bioinformatics**, v. 14, n. 6, p. 1288–1301, 2017. Citado 3 vezes nas páginas 43, 46 e 70.

ZHANG, Yang; ARAKAKI, Adrian K.; SKOLNICK, Jeffrey. Tasser: An automated method for the prediction of protein tertiary structures in casp6. **Proteins: Structure, Function, and Bioinformatics**, v. 61, n. S7, p. 91–98, 2005. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.20724>. Citado na página 41.

ZHOU, Hongyi; SKOLNICK, Jeffrey. Ab initio protein structure prediction using chunk-tasser. **Biophysical Journal**, v. 93, n. 5, p. 1510 – 1518, 2007. ISSN 0006-3495. Disponível em: <http://www.sciencedirect.com/science/article/pii/S000634950771411X>. Citado na página 41.