PEDRO HENRIQUE NARLOCH

Análise da Aplicação de Diferentes Métricas de Energia para o Problema de Enovelamentos de Proteínas

Dissertação apresentada ao Programa de Pós-Graduação em Computação Aplicada da Universidade do Estado de Santa Catarina, como requisito parcial para obtenção do grau de Mestre em Computação Aplicada.

Orientador: Dr. Rafael Stubs Parpinelli

JOINVILLE, SC
2017

Ficha catalográfica elaborada pelo(a) autor(a), com
auxílio do programa de geração automática da
Biblioteca Setorial do CCT/UDESC

Narloch, Pedro Henrique
Análise da Aplicação de Diferentes Métricas de Energia para o Problema de Enovelamentos de Proteínas / Pedro Henrique Narloch. - Joinville , 2017.
98 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, 2017.

1. Predição de Estrutura de Proteínas. 2. Evolução Diferencial. 3. Bioinformática. I. Parpinelli, Rafael Stubs. II. Universidade do Estado de Santa Catarina. Programa de Pós-Graduação. III. Título.

Análise da Aplicação de Diferentes Métricas de Energia para o Problema de Enovelamento de Proteínas

por

Pedro Henrique Narloch

Esta dissertação foi julgada adequada para obtenção do título de

Mestre em Computação Aplicada

Área de concentração em "Ciência da Computação,
e aprovada em sua forma final pelo

CURSO Mestrado Acadêmico em Computação Aplicada
CENTRO DE CIÊNCIAS TECNOLÓGICAS DA
UNIVERSIDADE DO ESTADO DE SANTA CATARINA.

Banca Examinadora:

Prof. Dr. Rafael Stubs Parpinelli
CCT/UDESC (Orientador/Presidente)

Prof. Dr. Charles Christian Miers
CCT/UDESC

Prof. Dr. Heitor Silvério Lopes
UTFPR

Joinville,SC, 22 de agosto de 2017.

Agradecimentos

Agradeço ao apoio dos meus queridos pais que, desde a minha infância, me encorajaram a estudar e buscar o conhecimento. Agradeço também aos meus colegas de laboratório que em momentos críticos me apoiaram e que propiciaram boas risadas. Ao Prof. Dr. Rafael Stubs Parpinelli por ter me aceito como orientado e ter a paciência de repassar os seus conhecimentos a mim. Sou grato também ao Departamento de Ciência da Computação (DCC) por ter liberado os laboratórios de ensino para a execução dos testes durante todo o período do mestrado nos fins de semana.

Por fim, agradeço a todos aqueles que indiretamente torceram pelo meu sucesso, ajudando a me manter motivado nos piores momentos e construindo a história de minha vida até o presente momento. Muito obrigado.

"A ignorância mais frequentemente gera confiança do que o conhecimento: são os que sabem pouco, e não aqueles que sabem muito, que afirmam de uma forma tão categórica que este ou aquele problema nunca será resolvido pela ciência" - Charles Darwin

RESUMO

A predição de estrutura de proteínas é considerada um dos mais importantes e desafiadores problemas em aberto da bioinformática estrutural. A complexidade dela está relacionada à grande quantidade de conformações que uma sequência de aminoácidos pode assumir. O objetivo dessa dissertação é aplicar e avaliar três das mais conhecidas funções de energia utilizadas para a predição de estrutura de proteínas: CHARMM, AMBER e ROSETTA. O algoritmo de Evolução Diferencial é utilizado como método de busca, aplicando também duas diferentes técnicas de diversificação conhecidas como *Generation Gap* e mutação Gaussiana. Um modelo de representação atômico, conhecido por ângulos de torções da cadeia principal e cadeia lateral, é utilizado. Para testar as abordagens foram selecionadas as proteínas 1PLW, 1ZDD e 1CRN. Além disso, uma abordagem diferente é utilizada para predizer outras duas proteínas: 1ENH e 1AIL. Essa abordagem foi chamada de $\text{DE}_{cascata}$, pois ela utiliza diferentes mecanismos de mutação conforme a observação do comportamento do algoritmo em relação às suas capacidades de diversificação e intensificação. Os resultados obtidos são comparados com alguns trabalhos da literatura para avaliar o seu desempenho. Outra contribuição presente nesse trabalho é a correlação dos resultados de energia com o monitoramento de diversidade populacional, demonstrando o impacto que funções de diversificação possuem sobre o algoritmo. Resultados competitivos com os da literatura foram obtidos, apesar da simplicidade das abordagens utilizadas.

Palavras-chaves: Predição de Estrutura de Proteínas, Evolução Diferencial, Bioinformática.

ABSTRACT

Protein Structure Prediction is considered one of the most important and challenging open problem in structural bioinformatics. Its complexity is related to the explosion of plausible shapes that an amino acids sequence could become. The objective of this dissertation is to apply and evaluate three of the most known energy functions for protein structure prediction: CHARMM, AMBER and ROSETTA. The Differential Evolution algorithm is used as a search method and two different techniques of diversification known as Generation Gap and Gaussian Mutation. An atomic representation model known as backbone and sidechain torsion angles is used. To test the approaches it were used the proteins 1PLW, 1ZDD and 1CRN . Beyond that, a different approach was developed to predict other two proteins: 1ENH and 1AIL. This approach is called $DE_{cascade}$ because it uses different mutation mechanisms according to the observation of the behavior related to the diversification and intensification capacity. The results obtained are compared to some literature works to evaluate their performance. Another contribution in this work is the correlation study among the energy results and the genotypic diversity, showing that diversifications functions have impact in the algorithm. Competitive results were obtained in spite of the simplicity of the used approaches.

Key-words: Protein Structure Prediction, Differential Evolution, Bioinformatics.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

2D	Bidimensional
3D	Tridimensional
ABC	Colônias de Abelas Artificiais
ACO	Otimização por Colônia de Formigas
ADEMO/D	*Adaptive Differential Evolution for Multi-objective Problems based on Decomposition*
AIS	*Artificial Immune Systems*
AMBER	*Amber Molecular Dynamics Package*
APL	*Angle Probability List*
Å	*Angstrom*
BFOA	Algoritmo de Otimização por Colônia de Bactérias
CASP	*Critical Assessment of protein Structure Prediction*
CHARMM	*Chemistry at HARvard Macromolecular Mechanics*
CR	Taxa de *Crossover*
DE	Evolução Diferencial
DE_{GG}	Evolução Diferencial com *Generation Gap*
DE_{GP}	Evolução Diferencial com Mutação Gaussiana
DE_{GG-GP}	Evolução Diferencial com *Generation Gap* e Mutação Gaussiana
EC	Computação Evolucionária
ECO	Coevolutivos com Inspiração Ecológica
F	Taxa de Mutação
FSS	Busca por Cardumes de Peixes
GA	Algoritmos Genéticos
GPU	*Graphical Processor Unit*
GSD	Desvio Padrão utilizado na Mutação Gaussiano

HP	Hidrofóbico-Polar
I-PAES	*Immune-Inspired Pareto Archieved Evolution Strategy*
MAX-AVAL	Máximo de Avaliações
MDF	Medida de Diversidade Fenotípica
MDG	Medida de diversidade populacional
NMR	Ressonância Nuclear Magnética
NSGA	*Non-dominated Sorting Genetic Algorithm*
PDB	Protein Data Bank
PPEP	Problema de Predição de Estrutura de Proteínas
PSO	Otimização pro Enxame de Partículas
REU	*Rosetta Energy Unit*
RMSD	*Root Mean Square Deviation*
RNA	Redes Neurais Artificiais
SA	*Simulated Annealing*
SI	Inteligência de Enxames
SSORIGA	*Simplified Self-Organizing Random Immigrants*
UB	*Urey-Bradley*

Sumário

1 INTRODUÇÃO

Proteínas são macromoléculas compostas por ligações polipeptídicas entre aminoácidos (sendo 20 diferentes tipos de aminoácidos conhecidos), que exercem importantes funções biológicas em todo organismo quando a sua forma tridimensional (3D) é encontrada (WALSH, 2014). A predição da estrutura das proteínas é um dos problemas ainda em aberto da biologia e bioinformática estrutural. Essas moléculas são formadas por uma sequência de aminoácidos, denominada de estrutura primária, que formam uma cadeia polipeptídica através de ligações químicas. Em seu estado final o polipeptídio assume a configuração tridimensional, conhecida como conformação nativa, para exercer a função biológica da proteína no organismo. O processo químico que faz com que a proteína assuma sua forma funcional é conhecida como enovelamento de proteínas.

A estrutura de uma proteína pode ser obtida por duas técnicas experimentais conhecidas como a cristalografia por difração de raios X e ressonância magnética nuclear (NMR) (DORN et al., 2014). Apesar dessas técnicas conseguirem obter a estrutura de uma proteína, elas possuem alto custo financeiro e demandam bastante tempo para a experimentação de uma única estrutura. Devido a dificuldade em definir a estrutura 3D da proteína, gerou-se uma grande diferença entre a quantidade de proteínas sequenciadas e proteínas que possuem a estrutura conhecida. Entre proteínas sequenciadas são encontrados mais de 87.846.000[1] registros na base de dados conhecida como UniProKB (BOUTET et al., 2016) enquanto que proteínas com suas estruturas conhecidas são pouco mais de 130.000[2] registradas no Protein Data Bank (PDB) (BERMAN, 2000). Devido à essa diferença, diversos pesquisadores têm atuado em diferentes áreas de conhecimento como matemática, química, física, computação e biologia para desenvolver modelos e métodos computacionais que consigam encontrar a estrutura 3D das proteínas, surgindo dessa forma o Problema de Predição de Estrutura de Proteínas (PPEP).

Em termos computacionais, o PPEP que utiliza somente a sequência de aminoácidos como informação é conhecida como *Ab Initio* (BONNEAU; BAKER, 2001). A complexidade atrelada a esse tipo de problema é devida à quatidade de possíveis combinações de conformações que a proteína pode assumir. Dessa forma, não se conhece um algoritmo que execute em tempo polinomial e encontre a conformação nativa de uma proteína em sua representação atômica. Dada essa característica, é possível classificar o PSP como um problema pertencente à classe de problemas NP-Completo (GUYEUX et al., 2014). Dessa forma, as metaheurísticas tornam-se uma possibilidade viável de explorar o espaço de busca e encontrar algumas conformações plausíveis em um tempo aceitável (PARPINELLI; LOPES, 2013).

[1] Pesquisa realizada em Junho de 2017
[2] Pesquisa realizada em Junho de 2017

Dentre os modelos de representação das proteínas existentes, é possível classificar eles em duas categorias: modelos em *lattice* e *off lattice* (BONNEAU; BAKER, 2001). A modelagem em *lattice* tem como característica a diminuição dos graus de liberdade dos átomos. Já a modelagem *off lattice* não há restrição dos graus de liberdade dos átomos de uma proteína. Por não aplicar essa restrição, a modelagem *off lattice* é associada ao paradoxo de Levinthal que, devido a quantidade de graus de liberdades de uma proteína, ela pode assumir inúmeras formas funcionais diferentes devido ao tamanho do espaço conformacional (DORN et al., 2014). Uma das representações pertencentes a essa classe de modelagem é a de ângulos de torções da cadeia principal e cadeia lateral, ao qual são levados em considerações alguns ângulos da representação atômica da cadeia polipeptídica (CUTELLO; NARZISI; NICOSIA, 2006). Esse tipo de representação tende a ter maior nível de detalhamento na sua representação e, consequentemente, é mais fiel a representação real de uma proteína quando comparada a outras representações. Existem ainda outras representações conhecidas, como: *all-atom 3D coordinates*, *all-heavy-atom coordinates*, *backbone atom coordinates + sidechain centroids* e C_α *coordinates* (CUTELLO; NARZISI; NICOSIA, 2008). O modelo adotado nessa dissertação é o modelo *off lattice* com representação da proteína em ângulos de torções da cadeia principal e cadeia lateral por se aproximar da representação de uma proteína.

Independente da representação utilizada, é necessário que uma função de energia seja associada para verificar a qualidade da proteína predita. Essa medida é importante devido a teoria de que a proteína atinge sua conformação nativa com o menor valor de energia possível, ou seja, a proteína assume sua forma funcional 3D quando atingir o menor valor de energia (ANFINSEN et al., 1961). As fórmulas mais complexas estão associadas aos modelos de representação atômicos, como a representação de ângulos e torções da cadeia principal e cadeia lateral, ao qual são utilizados conceitos de físicos e químicos para definir o valor de energia da proteína. Essas funções de energia levam em consideração as forças ligantes, ligações entre os átomos da macromolécula, e não ligantes, como a força de Van der Waals e as interações eletroestáticas. Algumas formas de calcular a energia potencial existentes na literatura são: CHARMM, AMBER, GROMOS e ROSETTA (LEE; WU; ZHANG, 2009).

A computação natural é um dos campos de pesquisa que investigam modelos e técnicas computacionais inspiradas na natureza (KARI; ROZENBERG, 2008). Dentro da computação natural existem os algoritmos que inspiram-se em comportamentos observados na natureza, esses algoritmos são conhecidos como algoritmos bioinspirados. Os algoritmos bioinspirados são metaheurísticas que abstraem as informações encontradas na natureza para a solução de problemas complexos. Dentre esses algoritmos estão as redes neurais, os algoritmos inspirados em sistemas imunológicos artificiais, a inteligência

por enxames e também a computação evolutiva. As metaheurísticas da inteligência por enxames buscam inspirações em mecanismos encontrados em diversos grupos biológicos, como as abelhas, os peixes, pássaros e formigas (KAR, 2016). Já a computação evolutiva baseia-se na teoria da evolução das espécies de Charles Darwin.

Uma metaheurística populacional é aquela que possui um conjunto de indivíduos que representem soluções em potencial para o problema. Para que uma metaheurística populacional possa ser bem sucedida na otimização de um problema, deve existir um equilíbrio entre as rotinas de diversificação e intensificação. Como rotinas de diversificação pode-se entender sobre a busca global, tendo o objetivo de identificar possíveis áreas que possuem soluções boas. Já a intensificação refere-se na busca intensiva em determinado local que seja identificado como promissor. Com o equilibrio entre esses dois mecanismos é possível explorar melhor o espaço de busca para possivelmente encontrar soluções próximas ao ótimo global (CORRIVEAU et al., 2012). Caso não haja esse controle, há a probabilidade do algoritmo ficar preso em um ponto local, configurando uma convergência prematura e, consequentemente, soluções subótimas. Para identificar esse tipo de comportamento é possível utilizar métricas que ajudam a verificar a diversidade durante o processo de otimização.

No presente trabalho é abordada a discussão sobre os elementos e ferramentas que envolvem a predição de estrutura de proteínas para o modelo *Ab Initio* com representação de ângulos e torções da cadeia lateral e principal e da importância do controle de diversidade das soluções durante o processo de otimização do PPEP. Além dessa análise, são feitos testes com três funções de energia AMBER, CHARMM e ROSETTA utilizando as abordagens que aplicam métodos de diversificação poulacional. Por fim, uma abordagem diferente de exploração do espaço de busca é definido de maneira determinista, baseando-se no comportamento dos diferentes mecanismos de mutação encontrados no algoritmo de Evolução Diferencial (*Differential Evolution* - DE), chamado de $\text{DE}_{cascata}$, e aplicado apenas a função de energia ROSETTA devido ao baixo custo computacional quando comparado às outras duas funções de energia.

1.1 MOTIVAÇÃO

A principal motivação do presente trabalho é aplicar um modelo computacional para a predição 3D de proteínas em sua representação atômica a partir da estrutura primária utilizando diferentes funções de energia: AMBER, CHARMM e ROSETTA para verificar a qualidade das soluções geradas por elas.

Outro fator motivacional é monitorar e analisar a diversidade populacional para entender o comportamento das metaheurísticas utilizadas durante o processo de otimiza-

ção e com isso identificar a necessidade da aplicação de rotinas de diversificação durante o processo de otimização. Com a análise desse comportamento é possível que novas abordagens sejam desenvolvidas para buscar esse equilíbrio.

1.2 OBJETIVOS

O objetivo geral do presente trabalho é aplicar e avaliar três das mais conhecidas funções de energia utilizadas para a predição de estrutura terciária de proteínas em um modelo computacional. Esse modelo utiliza o algoritmo bioinspirado conhecido como Evolução Diferencial, verificando o impacto que a manutenção de diversidade populacional possui em relação ao resultado final do algoritmo. Dessa forma existem duas hipóteses que são verificadas nesse trabalho: **(a)** A quantidade de informações presentes na função de energia alcança melhores conformações ?; **(b)** A manutenção de diversidade populacional, durante o processo de otimização, é um fator que deve ser considerado ?

Para atingir o objetivo geral e validar as hipóteses dessa dissertação, os seguintes objetivos específicos foram elaborados:

- Desenvolver um modelo computacional para a predição de estruturas terciárias de proteínas que utilize as funções de energia CHARMM, AMBER e ROSETTA.
- Aplicar os mecanismos de diversificação conhecidos como *Generation Gap* e mutação Gaussiana no algoritmo DE e monitorar a diversidade populacional das abordagens.
- Desenvolver uma nova abordagem que leve em consideração o comportamento dos diferentes mecanismos de mutação do algoritmo DE e aplicar essa abordagem para outras proteínas utilizando a função de energia ROSETTA.

1.3 ORGANIZAÇÃO DO TRABALHO

Este trabalho está estruturado da seguinte maneira. O Capítulo 2 apresenta os fundamentos para melhor entendimento do presente trabalho e a revisão da literatura. O Capítulo 3 descreve os métodos utilizados para desenvolver as abordagens utilizadas, além das funções e métricas empregadas. O Capítulo 4 apresenta configurações de ambiente, parâmetros e proteínas utilizadas nos experimentos e a análise dos resultados obtidos. O Capítulo 5 apresenta as considerações finais e trabalhos futuros.

2 REVISÃO DA LITERATURA

Nessa seção são introduzidos os conceitos necessários para a compreensão do trabalho desenvolvido. São expostos os conceitos sobre proteínas como suas características e importâncias biológicas, o algoritmo de evolução diferencial que é utilizado no presente trabalho, as métricas de diversidade para análise da distribuição da população no espaço de busca e o problema de predição de estrutura de proteínas, descrevendo os principais meios de representação das proteínas, os tipos de modelagem e suas funções de energia.

2.1 PROTEÍNAS

As proteínas são macromoléculas responsáveis por diversas funções biológicas nos organismos. Elas são formadas por uma sequência de aminoácidos que, em determinado momento, se enovelam em uma forma 3D. Essa forma 3D caracteriza o estado nativo de uma proteína e é quando ela começa a exercer sua função biológica. As proteínas tem diversas funções no organismo como, por exemplo, a regulagem dos níveis de glucose no sangue, o estímulo ao hormônio que produz hemáceas responsáveis por transportar oxigênio aos tecidos, entre outros (WALSH, 2014).

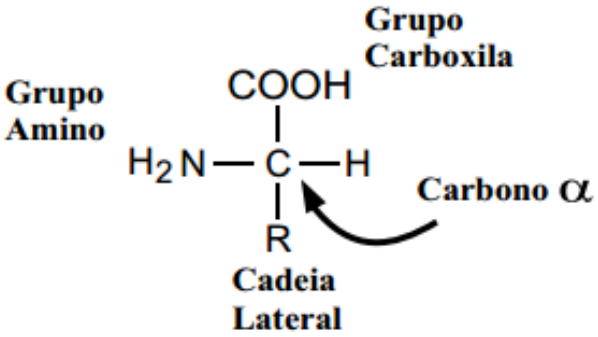
O processo de enovelamento de uma proteína é espontâneo, complexo e ainda não se conhece os detalhes biológicos de como ele acontece. Apesar de uma sequência única de aminoácidos se transformar em uma proteína com estrutura 3D única, pode ocorrer falhas no processo de enovelamento, fazendo com que as proteínas não consigam exercer sua função corretamente. Essas proteínas que possuem falhas em sua estrutura nativa podem ser descartadas pelo organismo ou assumir funções nocivas. Muitas doenças estão relacionadas a proteínas que não são enoveladas corretamente, como o mal de Parkinson, Fibrose Cística e outras doenças neurodegenerativas (MáRQUEZ-CHAMORRO et al., 2015). O que afeta a maneira como a proteína assume sua conformação nativa, segundo (ANFINSEN et al., 1961), é a sequência de aminoácidos e as condições do ambiente ao qual essa proteína está exposta. As forças ligantes e não ligantes dos átomos de cada proteína são diretamente influenciadas por esses dois fatores.

2.1.1 Aminoácidos

Os monômeros conhecidos como aminoácidos são compostos orgânicos que formam ligações polipeptídicas e, consequentemente, as proteínas. Existem vinte tipos diferentes de aminoácidos na natureza e todos eles possuem uma estrutura simples, sendo que o primeiro aminoácido identificado foi a Asparagina em 1806. Cada aminoácido é formado por um átomo de carbono central (C_α) e quatro ligações: um átomo de hidrogênio (H),

um grupo carboxila ($COOH$), um grupo amina (NH_2) e uma cadeia secundária conhecida como radical R, sendo essa cadeia secundária que diferencia cada um dos 20 aminoácidos conforme representado na Figura 1.

Figura 1 – Estrutura de um aminoácido



Fonte: Autoria Própria

Os aminoácidos comumente são identificados por apenas uma letra ou por abreviações de três letras. A Tabela 1 lista os vinte diferentes tipos de aminoácidos, sendo a primeira coluna o nome do aminoácido, a segunda coluna a identificação por três letras e a terceira coluna sendo a identificação de uma única letra (WALSH, 2014).

Tabela 1 – Listagem dos vinte aminoácidos.

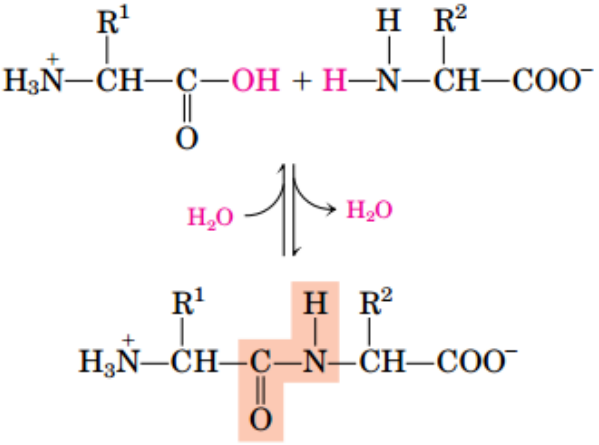
Aminoácido	Cód. 3 letras	Cód. 1 letra
Ácido Aspártico	ASP	D
Ácido Glutâmico	GLU	E
Alanina	ALA	A
Arginina	ARG	R
Asparagina	ASN	N
Cisteína	CYS	C
Fenilalanina	PHE	F
Glicina	GLY	G
Glutamina	GLN	Q
Histidina	HIS	H
Isoleucina	ILE	I
Lisina	LYS	K
Leucina	LEU	L
Metionina	MET	M
Prolina	PRO	P
Serina	SER	S
Tirosina	TYR	Y
Treonina	THR	T
Triptofano	TRP	W
Valina	VAL	V

Os aminoácidos ainda podem ser classificados pelas características de sua cadeia secundária, sendo divididos em cinco diferentes classes: grupos apolares alifáticos, que não

possuem afinidade com a água (hidrofóbicos); grupos apolares aromáticos que também são hidrofóbicos; grupos não carregados e polares que são mais solúveis em água e possuem maior afinidade com a água; grupos carregados positivamente ou básicos e grupos carregados negativamente ou ácidos (NELSON; COX; LEHNINGER, 2013).

Os aminoácidos criam cadeias peptídicas através de ligações químicas covalentes que ocorre no ribossomo de uma célula. Esse tipo de ligação ocorre quando se perde uma molécula de água e o grupo carboxila de um aminoácido se une ao grupo amina de outro aminoácido. Esse processo de perda de molécula de hidrogênio para a união de dois aminoácidos é exemplificado pela Figura 2 (NELSON; COX; LEHNINGER, 2013).

Figura 2 – Processo de formação de peptídeos.



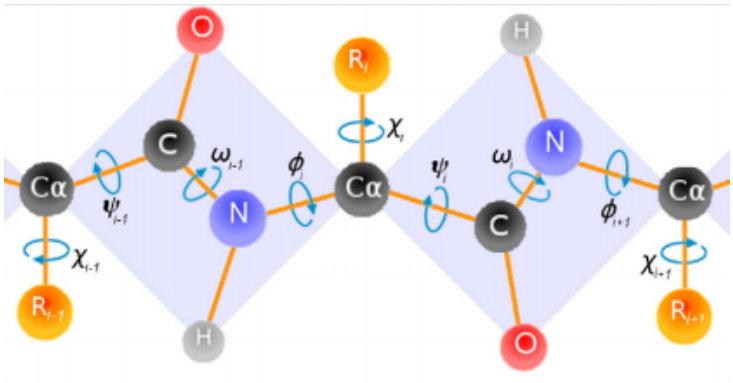
Fonte: (NELSON; COX; LEHNINGER, 2013)

Dessa forma, quando vários aminoácidos fazem esse tipo de ligação eles se tornam um polipeptídio que posteriormente é enovelado em uma forma 3D, chegando a conformação nativa da proteína. Quando há a formação de peptídeos, três ângulos são associados a essa estrutura (ϕ, ψ, ω). O ângulo ϕ está relacionado a ligação $N - C_\alpha$, enquanto que o ângulo ψ está entre a ligação $C - C_\alpha$. O ângulo ω está entre a ligação $C - N$. A Figura 3 apresenta a disposição dos ângulos existentes nos peptídeos

Esses três ângulos formam a cadeia principal (*backbone*) da proteína, sendo que os ângulos ϕ e ψ possuem rotação livre enquanto que o ângulo ω não pode rotacionar livremente, tomando uma forma planar no espaço. Essa característica se dá devido ao caráter de ligação dupla parcial entre a ligação $C - N$, pois o átomo N adquire uma carga positiva parcial enquanto que o átomo C possui uma carga parcial negativa. Dessa forma, o ângulo ω tem liberdade de torsão perto de 0° ou 180°. Além do *backbone* de uma proteína, ainda existem os ângulos χ_n que são associados a cadeia secundária de cada aminoácido na ligação $C_\alpha - R$ e que possuem quantidades diferentes conforme apresentado pela Tabela 2.

A quantidade de ângulos χ varia de acordo com o aminoácido, possuindo também

Figura 3 – Ângulos de torções de peptídeos.



Fonte: (BORGUESAN et al., 2015)

Tabela 2 – Ângulos χ para cada aminoácido.

Aminoácido	χ Ângulos
GLY, ALA, PRO	*Backbone*
SER, CYS, THR, VAL	χ_1
ILE, LEU, ASP, ASN, PHE, TYR, TRP	χ_1, χ_2
MET, GLU, GLN	χ_1, χ_2, χ_3
LYS, ARG	χ_1, χ_2, χ_3, χ_4

grau de rotação livre no espaço. Dessa forma, uma proteína pode ser representada computacionalmente com 3 ângulos pertencentes a cadeia principal (ϕ, ψ e ω) e i ângulos pertencentes a cadeia lateral (χ_i).

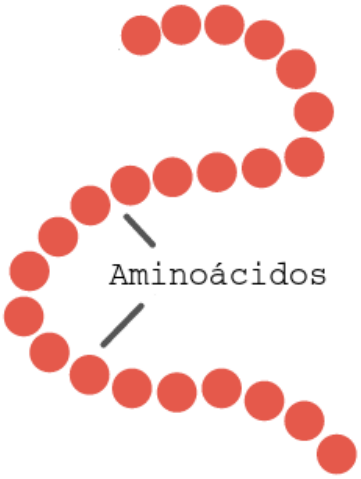
2.1.2 Classificação de Estruturas

As proteínas podem ser classificadas em quatro diferentes estruturas: estrutura primária, estrutura secundária, estrutura terciária e estrutura quartenária. A estrutura primária de uma proteína é o nível inicial de sua formação e é nela que se encontra a sequência de aminoácidos (WALSH, 2014). A Figura 4 representa a estrutura primária de uma proteína, ao qual cada aminoácido é representado por um círculo vermelho. A quantidade de aminoácidos pertencentes a uma proteína é variável.

A estrutura secundária de uma proteína possui uma forma 3D, porém ela representa somente o arranjo espacial dos componentes da cadeia principal, não levando em consideração a cadeia secundária (NELSON; COX; LEHNINGER, 2013). Geralmente a estrutura secundária de uma proteína assume formatos de α-hélices ou folhas β. Esses tipos de estruturas se estabilizam dessa forma devido à maximização das ligações de hidrogênios intramoleculares e minimização da repulsão entre as cadeias laterais de cada aminoácido (WALSH, 2014).

A formação de uma α-hélice possui aproximadamente 3.6 aminoácidos por volta,

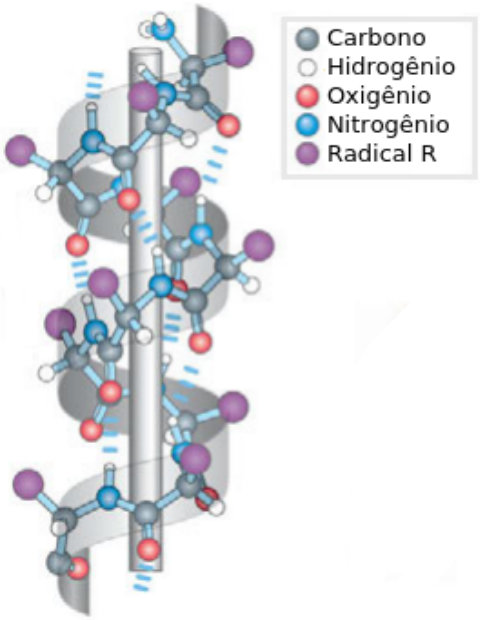
Figura 4 – Estrutura primária de uma proteína.



Fonte: Autoria Própria

sendo que a cadeia secundária de cada aminoácido fica do lado externo da estrutura. A estrutura de uma α-hélice é representada na Figura 5.

Figura 5 – Estrutura secundária - α-Hélice.



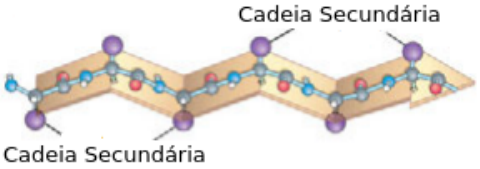
Fonte: Adaptado de (NELSON; COX; LEHNINGER, 2013)

Esse tipo de estrutura é estabilizado por ligações de hidrogênios entre o grupo $C = O$ e o grupo $N - H$. A quantidade de voltas desse tipo de estrutura varia, podendo ir de apenas uma única volta até valores superiores a dez voltas consecutivas (WALSH, 2014)

As Folhas β, assim como as hélices α, são estruturas bastante comuns nas proteínas. Esse tipo de estrutura geralmente é formada por 5 a 10 aminoácidos e raramente se encontram sozinhas e possuem uma estrutura em "zigue zague"(WALSH, 2014). A ca-

deia lateral dos aminoácidos que compõe essa estrutura se projetam a partir do "zigue zague"formado, conforme a Figura 6.

Figura 6 – Estrutura secundária - Folha-β.



Fonte: Adaptado de (NELSON; COX; LEHNINGER, 2013)

As folhas β podem ser classificadas em paralelas ou anti-paralelas, sendo o alinhamento da ligação amino-carboxílico o que define essa característica (NELSON; COX; LEHNINGER, 2013). A Figura 7 diferencia visualmente as folhas β paralelas ou anti-paralelas.

Figura 7 – Diferença entre folhas β paralelas e anti-paralelas.

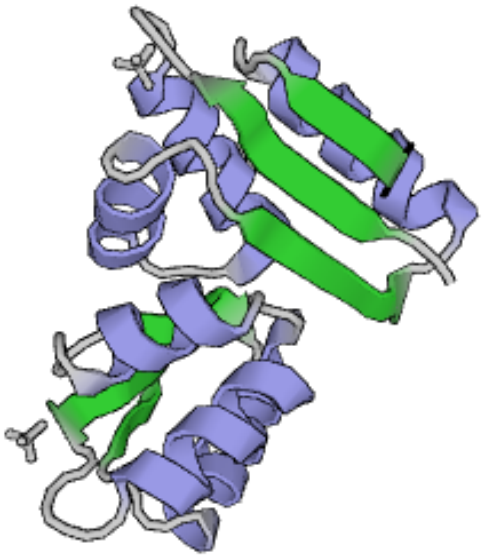
Fonte: Adaptado de (NELSON; COX; LEHNINGER, 2013)

Além das α-hélices e das folhas β, existem outras estruturas que servem para fazer a ligação entre essas estruturas secundárias, elas são conhecidas como voltas e dobras. O que diferencia uma da outra é o seu comprimento.

O terceiro tipo de estrutura é conhecido como estrutura terciária de uma proteína, conhecida também como estrutura nativa e é a partir desse nível que a proteína começa a exercer sua função biológica. Enquanto que a estrutura secundária de uma proteína referencia a organização dos aminoácidos próximos de um segmento do polipeptídeo, a estrutura terciária inclui todos os aspectos dos polipeptídeos (NELSON; COX; LEHNINGER, 2013). A Figura 8 apresenta a estrutura terciária da proteína identificada como 1CTF no banco de dados de proteínas (PDB) e possui 74 aminoácidos. Nesse tipo de proteína observa-se que há uma interação entre todas as estruturas, com ligações entre α-hélices, folhas β, voltas e dobras.

Por fim, existe a estrutura quartenária das proteínas. Esse tipo de estrutura ocorre quando uma proteína apresenta subunidades terciárias arranjadas no espaço. A proteína conhecida como Hemoglobina possui a interação entre quatro subunidades e é apresentada na Figura 9.

Figura 8 – Estrutura terciária da proteína 1CTF.



Fonte: (BERMAN, 2000)

Figura 9 – Estrutura quartenária da hemoglobina.

Fonte: (BERMAN, 2000)

Cada subunidade da Hemoglobina é identificada por uma cor diferente para representar a relação entre as macromoléculas.

2.1.3 Método de Determinação de Estruturas

Experimentalmente é possível obter a estrutura 3D de uma proteína através de técnicas de ressonância magnética nuclear (NMR) ou por cristalografia de raio X. A cristalografia por raio X tem sido mais utilizada em comparação com a técnica NMR (SCHNEIDER; FU; KEATING, 2009) devido a qualidade de seus resultados. Apesar da cristalografia por raio X ter melhor qualidade na definição da estrutura, ela é limitada

apenas a moléculas cristalizadas. A técnica NMR além de ser mais recente, tem a capacidade de identificar a estrutura de proteínas não cristalizadas, ou seja, em estado líquido (NELSON; COX; LEHNINGER, 2013).

Apesar de ambas as técnicas conseguirem identificar a forma 3D das proteínas e apresentarem um bom resultado, elas são caras e possuem um processo complexo que demanda muito tempo. Devido a essas restrições houve a motivação de formular métodos computacionais para a predição dessas estruturas de uma maneira mais barata e que demande menos tempo. Esses métodos computacionais serão abordados na Seção 2.2.

2.2 PREDIÇÃO DE ESTRUTURA DE PROTEÍNAS

As inferências feitas por Anfinsen sobre o enovelamento espontâneo de uma estrutura 3D, devido a ações do ambiente ao qual essa molécula estava envolvida (ANFINSEN et al., 1961), foram importantes para definir o problema da predição das proteínas a partir de sua estrutura primária. Esse tipo de modelagem para a descoberta da estrutura terciária das proteínas a partir de sua estrutura primária é conhecido como modelo *Ab Initio* (BONNEAU; BAKER, 2001).

Como mencionado na Seção 2.1.3, existem dois métodos que conseguem determinar a estrutura nativa de uma proteína, porém, o uso desses métodos é bastante custoso e complexo. Por esses motivos há uma grande diferença entre a quantidade de proteínas que foram sequenciadas com a a quantidade de proteínas que possuem sua estrutura 3D conhecida (DORN et al., 2014).

Nesta seção são abordadas as principais formas de representação das proteínas e também as modelagens do problema encontradas na literatura. Três funções de energia utilizadas nesta dissertação também são abordadas, assim como a base de dados de estruturas já conhecidas.

2.2.1 Representação Computacional das Proteínas

Devido a complexidade da representação de uma proteína em meios computacionais, diversas abordagens foram formuladas (KOLINSKI; SKOLNICK, 2004) e podem ser classificadas em duas categorias: representações em *lattice* e representações *off lattice*. As representações em *lattice* são, em sua maioria, mais simples que as *off lattice*, restringindo os graus de liberdade de rotação dos átomos de uma proteína. Já a modelagem *off lattice* representa uma proteína mais fielmente, tendo maiores graus de liberdade e maior complexidade.

Modelos reduzidos de representação são ferramentas importantes para o entendimento da dinâmica protéica e também dos elementos de termodinâmica envolvidos (KO-

LINSKI; SKOLNICK, 2004). Essas representações incluem os modelos em *lattice*, modelos de espaço contínuo e também modelos híbridos. As soluções encontradas para os modelos reduzidos muitas vezes se atentam a resolução da forma da cadeia principal da proteína, sem levar em consideração a cadeia secundária dos aminoácidos (WOOLEY; YE, 2007).

Um dos modelos reduzidos mais conhecido e explorado é o Hidrofóbico-Polar (HP), criado por Lau e Dill (LAU; DILL, 1989). A abordagem HP classifica os aminoácidos em dois tipos de resíduos: hidrofóbico (que não possuem afinidade com a água) e polares (que possuem maior afinidade com a água). Com essa representação uma conformação é avaliada conforme as interações de aminoácidos hidrofóbicos. Existem ainda outros modelos reduzidos como o modelo Helicoidal HP (THOMAS; DILL, 1993), Homopolímero Perturbado (SHAKHNOVICH; GUTIN, 1993), Incorporação de Gráficos Carregados (FRAENKEL, 1993) e Incorporação de Polímeros em *Lattice* (UNGER; MOULT, 1993). Uma revisão de vários modelos de representação podem ser consultados em (KOLINSKI; SKOLNICK, 2004) e variações do modelo HP em (DILL et al., 1995).

Por sua vez, os modelos *off lattice* apresentam características aproximadas com as representação real de uma proteína. Um dos modelos reduzidos e *off lattice* conhecidos da literatura é o modelo AB (STILLINGER; HEAD-GORDON, 1995), ao qual os aminoácidos são separados em duas classes conforme sua afinidade com a água (hidrofóbicos e hidrofílicos). Apesar desse tipo de representação não levar em consideração todas as características de uma proteína, ele possuem implicações importantes para a identificação de certas características (SCALABRIN et al., 2014). Para o modelo AB, uma sequência de aminoácidos (monômeros) possui uma quantidade de $n - 2$ ângulos a serem definidos (PARPINELLI et al., 2014).

As representação que buscam ser fiéis a uma proteína real são chamados de representações atômicas, sendo essas representações mais complexas e custosas computacionalmente. Algumas das representações que são utilizadas normalmente são: coordenadas 3D *all-atom*; coordenadas *all-heavy-atom*; coordenadas de átomos do *backbone* + centroides da cadeia secundária; centróides C_α e ângulos de torções do backbone e cadeia lateral (CUTELLO; NARZISI; NICOSIA, 2008).

Uma das representações atômicas mais utilizadas é a de ângulos de torções do *backbone* e cadeia lateral. Nesse tipo de representação cada aminoácido possui um determinado número de ângulos para fixar a estrutura 3D de uma proteína. A quantidade de ângulos da cadeia lateral (χ_i) é variável conforme o resíduo. O ângulo ω tem seu valor fixado em 180°, não possuindo graus de liberdade. Diferentemente do restante dos ângulos (ϕ, ψ e χ_i) que podem variar de $-180°$ a 180° (CUTELLO; NARZISI; NICOSIA, 2008). Dessa forma, cada aminoácido possui 3 ângulos que compõe sua cadeia principal somado a quantidade de ângulos de sua cadeia secundária que pode variar conforme o aminoácido.

Quando uma proteína é predita em seu formato atômico e, a estrutura nativa dessa proteína é conhecida, pode-se utilizar uma métrica conhecida como *Root Mean Square Deviation* (RMSD) para comparação entre a proteína predita e a proteína já conhecida. Essa métrica leva em consideração as posições atômicas das duas proteínas e faz uma comparação de distância com a medida em angstrom (Å). No presente trabalho foi utilizada o RMSD_α que leva em consideração o alinhamento dos átomos de carbono_α e não todos os átomos das duas estruturas. Dessa forma é possível verificar o alinhamento das principais estruturas preditas pelos algoritmos utilizados. A Equação 2.1 apresenta a maneira de como o RMSD_α é calculado.

$$RMSD(a,b) = \sqrt{\frac{\sum_{i=1}^{n} \mid r_{ai} - r_{bi} \mid^2}{n}} \tag{2.1}$$

Sendo r_{ai} e r_{bi} as estruturas e i-ésimo átomo de um conjunto de n átomos, tendo seus valores expressados como valores contínuos.

2.2.2 Modelagem Baseadas em Conhecimento

Os métodos pertecentes a classe de modelagem baseadas em conhecimento, pode-se citar a homologia e *threading*. O sucesso do método é dependente da qualidade que a estrutura possui quando foi identificada pelos métodos de determinação de estruturas e de como essa informação é utilizada para a predição de proteínas (DORN et al., 2014).

A modelagem por *threading* é baseada no fato de que as estruturas das proteínas são preservadas durante a evolução dos organismos, ou seja, as proteínas podem possuir a mesma estrutura nativa com estruturas primárias diferentes. Com esse tipo de abordagem, a quantidade de possibilidades de enovelamento é finito devido a característica das estruturas terem se mantido durante o período evolutivo, não precisando percorrer todo o espaço de busca para encontrar a conformação nativa da proteína.

O objetivo geral da modelagem por *threading* é encontrar a sequência da proteína que se alinhe com algum modelo estrutural já conhecido (DORN et al., 2014), conseguindo encontrar o enovelamento da proteína de interesse. O processo de predição pode ser separado em duas etapas, selecionar o modelo estrutural de uma biblioteca de modelos; encontrar o posicionamento entre a sequência desejada e o modelo selecionado da biblioteca para reduzir o espaço de busca.

Na modelagem por homologia, acredita-se que as proteínas relacionadas evolutivamente possuem estruturas primárias muito semelhantes, refletindo dessa forma em estruturas 3D semelhantes (proteínas homólogas). Em (MARTI-RENOM; MADHUSUDHAN;

SALI, 2004) e (SáNCHEZ; ŠALI, 1997), são definidos os quatro passos básicos que formam o processo de predição por homologia:

- Seleção de Molde: É o ponto inicial do método de modelagem comparativa. Nessa etapa é necessário identificar todas as estruturas de proteínas que possuem suas sequências relacionadas com a sequência alvo e selecionar elas como *templates*;
- Alinhamento entre proteína molde e proteína alvo: Nessa etapa são feitos os alinhamentos entre as duas proteínas, formando uma base de modelos;
- Construção do Modelo: Nessa etapa é construído o modelo de predição, geralmente a cadeia principal é alinhada com as regiões do molde, deixando por último a cadeia secundária; e
- Avaliação do Modelo: Esse passo leva em conta todas as informações disponíveis da proteína alvo para verificar a qualidade da solução gerada.

Dessa forma, o objetivo da predição de estruturas de proteínas por homologia utiliza de estruturas homólogas já conhecidas como modelo (LEACH, 2001) para encontrar estruturas bastante similares às nativas.

2.2.3 Modelagem *Ab Initio*

A predição de estrutura de proteínas que utiliza somente informações da estrutura primária é conhecida como *Ab Initio* (BONNEAU; BAKER, 2001). Esse tipo de modelagem é fundamentada somente nas interações físicas e químicas entre os aminoácidos, sejam elas interações por ligações ou por interações de não ligantes. Além disso, a modelagem *Ab Initio* baseia-se em que a estrutura nativa de uma proteína está em sua menor energia livre (DORN et al., 2014).

Para procurar pela estrutura nativa da proteína na modelagem *Ab Initio* é necessário utilizar de funções de energia que descrevam o estado conformacional da proteína. Algumas das energias em potencial existentes na literatura são: AMBER (CORNELL et al., 1995), CHARMM (BROOKS et al., 2009), ROSETTA (ROHL et al., 2004). Outras funções de energia e seus pacotes de dinâmica molecular podem ser consultados em (DORN et al., 2014).

Apesar da abordagem *Ab Initio* ter uma deficiência em obter predições de alta qualidade quando comparado com modelagens que utilizam conhecimento de outras estruturas, ela tem como característica a possibilidade de descobrir estruturas novas que ainda não são conhecidas. Com o intuito de melhorar o resultado da modelagem *Ab Initio*, trabalhos estão propondo abordagens híbridas, ao qual são utilizados princípios físico e

bioquímicos com alguma fonte de conhecimento como em (BORGUESAN et al., 2015) que aplicaram listas de probabilidades de ângulos utilizadas conforme a combinação de aminoácidos presentes na estrutura primária da proteína.

2.2.4 Funções de Energia Potencial

A avaliação de energia livre de uma proteína é um dos passos essenciais para a predição da sua estrutura. Esse tipo de avaliação, quando aplicado a representações atômicas, é feita por campos de forças que levam em consideração todos os átomos do polipeptídio. Uma função de energia potencial possui dois tipos de termos: ligantes e não ligantes (DORN et al., 2014). Existem diversas funções de energia na literatura que são utilizadas em biologia computacional. As mais conhecidas são a AMBER (CORNELL et al., 1995), CHARMM (BROOKS et al., 2009), GROMOS (GUNSTEREN; DAURA; MARK, 2002) e ROSETTA (ROHL et al., 2004). Apesar de existirem diferenças nas fórmulas de cálculo das funções de energia, todas elas levam em considerações as propriedades físicas e químicas no momento de avaliação de uma proteína (HALGREN, 1995). Como nessa dissertação são utilizadas três funções de energia CHARMM, AMBER e ROSETTA, todas estão detalhadas nessa seção. As funções de energias CHARMM e AMBER foram escolhidas por estarem presentes no mesmo pacote de dinâmica molecular conhecido como TINKER[1] e por serem duas das mais conhecidas da literatura. Já a função de energia ROSETTA tem demonstrado resultados promissores no evento (CASP[2]) que possui o intuito de ranquear os melhores algoritmos de predição de estruturas de proteínas. Uma diferença importante delas está na unidade de medida utilizada para mensuração da energia de uma proteína, sendo a CHARMM e AMBER retornando valores em kcal mol^{-1} e a ROSETTA por uma unidade de medida única chamada de *Rosetta Energy Unit* (REU).

A função de energia CHARMM (BROOKS et al., 2009) foi criada por pesquisadores da Universidade de Harvard e é uma das das mais utilizadas para a avaliação de proteínas (DORN et al., 2014). A Equação 2.2 demonstra como é feito o cálculo de energia em kcal mol^{-1} pela função de energia CHARMM.

$$\begin{aligned} E_{CHARMM} = & \sum_{ligações} K_b(b-b_0)^2 + \sum_{UB} K_{UB}(S-S_0)^2 + \sum_{ângulos} K_\theta(\theta-\theta_0)^2 + \\ & \sum_{diedrais} K_\chi(1+\cos(\eta-\delta)) + \sum_{impróprios} K_{imp}(\varphi-\varphi_0)^2 + \\ & \sum_{nãoligantes} \epsilon[(\frac{R_{minij}}{R_{ij}})^{12} - (\frac{R_{minij}}{R_{ij}})^6] + \frac{q_i q_j}{\epsilon_1 r_{ij}} \end{aligned} \tag{2.2}$$

[1] https://dasher.wustl.edu/tinker/
[2] http://predictioncenter.org/

Sendo E_{CHARMM} o valor de energia total da proteína. Esse valor de energia é constituído por seis diferentes componentes que avaliam os termos ligantes e não ligantes. O termo **ligações** mensura o valor de energia de acordo com a ligação que ocorre entre dois átomos. O termo **Urey-Bradley** (UB) contabiliza a interação entre pares de átomos, já o termo **ângulos** se refere à soma de todos os ângulos da estrutura. Os termos **diedrais** e **impróprios** estão associados à energia dos ângulos de torções e a deformação desses ângulos, respectivamente. O termo **não ligantes** leva em consideração a energia de Van der Waals e Eletrostática. Os valores de Van der Waals são mensurados pelas interações dos átomos conforme as forças de atração e repulsão. Já a Eletrostática varia de acordo com a distância dos átomos.

A função de energia AMBER, apesar de ser similar à CHARMM, possui algumas diferenças na composição de sua fórmula. A Equação 2.3 apresenta a formulação que compõe a função de energia AMBER.

$$\begin{aligned} E_{AMBER} = & \sum_{ligações} \frac{1}{2} K_b (b - b_0)^2 + \sum_{ângulos} \frac{1}{2} K_\theta (\theta - \theta_0)^2 + \\ & \sum_{torções} K_\chi (1 + \cos(\eta_\chi - \delta)) + \\ & \sum_{j=1}^{N-1} \sum_{i=j+1}^{N} \left\{ \epsilon_{ij} \left[\left(\frac{r_{ij}}{\tau_{ij}} \right)^{12} - 2 \left(\frac{r_{ij}}{\tau_{ij}} \right)^{6} \right] + \frac{q_i q_j}{e 4 \pi \tau_i j} \right\} \end{aligned} \tag{2.3}$$

Sendo E_{AMBER} a energia da proteína. Para a função de energia AMBER são utilizados somente quatro termos. O termo **ligações** mensura o valor de energia de acordo com a ligação que ocorre entre dois átomos, o termo **ângulos** se refere à soma de todos os ângulos da estrutura e o termo **torções** é relacionado a energia dos ângulos de torções. Por fim, o cálculo dos não ligantes que leva em consideração a energia de Van der Waals e Eletrostática. A função de energia AMBER não possui o termo **UB** que contabiliza a interação entre pares de átomos e nem um termo separado para os ângulos de torções **impróprios** como há na energia CHARMM. Desse modo, pode-se entender que a formulação do cálculo de energia AMBER é mais simples que a CHARMM.

Por fim, a função de energia ROSETTA é a que apresenta a maior quantidade de componentes em sua formulação, em um total de se te componentes, sendo eles: interações de Lennard-Jones (Equação 2.4), interações eletroestáticas inter-atômicas (Equação 2.8), aproximação pelo potencial de solvatação (Equação 2.7), potencial de ligação de hidrogênios (Equação 2.6), auto-energia de rotâmeros (Equação 2.9), potencial geométrico de disulfideo (Equação 2.10) e preferências de torsões de Ramachandran(Equação 2.5)

(ROHL et al., 2004).

$$LJ = \sum_i \sum_{j>i} \begin{cases} \left[\left(\frac{r_{ij}}{d_{ij}}\right)^{12} - 2\left(\frac{r_{ij}}{d_{ij}}\right)^{6}\right]_{ij}, if \frac{d_{ij}}{r_{ij}} > 0,6 \\ \left[-8.759, 2\left(\frac{r_{ij}}{d_{ij}}\right) + 5.672, 0\right] e_{ij}, else \end{cases} \tag{2.4}$$

$$RTP = \sum_i -ln[P(\phi_i, \psi_i | aa_i, ss_i)] \tag{2.5}$$

$$\begin{aligned} HB = \sum_i \sum_j \big(& -ln[P(d_{ij}|h_j ss_{ij})] \\ & -ln[P(cos\theta_{ij}|d_{ijh_j ss_{ij}})] \\ & -ln[P(cos\psi_{ij}|d_{ij}h_j ss_{ij})]\big) \end{aligned} \tag{2.6}$$

$$SOLV = \sum_i \left[\Delta G_i^{ref} - \sum_j \left(\frac{2\Delta G_i^{free}}{4\pi^{3/2}\lambda_j r_{ij}^2} e^{-d_{ij}^2} V_j + \frac{2\Delta G_i^{free}}{4\pi^{3/2}\lambda_j r_{ij}^2} e^{-d_{ij}^2} V_i\right)\right] \tag{2.7}$$

$$ELTR = \sum_i \sum_{j>i} -ln\left[\frac{P\big(aa_i, aa_j | d_{ij}\big)}{P\big(aa_i|d_{ij}\big)P\big(aa_i|d_{ij}\big)}\right] \tag{2.8}$$

$$DUN = \sum_i -ln\left[\frac{P\big(rot_i|\phi_i, \psi_i\big)P\big(aa_i|\phi_i, \psi_i\big)}{P\big(aa_i\big)}\right] \tag{2.9}$$

$$REF = \sum_{aa} \eta_{aa} \tag{2.10}$$

As explicações de cada parâmetro pertencentes a esses componentes podem ser vistas em (ROHL et al., 2004). A partir dessas equações é possível verificar que a função de energia ROSETTA é a que possui maior quantidade de componentes em sua fórmula de cálculo.

2.3 ALGORITMOS BIOINSPIRADOS

Metaheurísticas estão sendo utilizadas para poder encontrar soluções boas em um período de tempo aceitável para problemas complexos, como o PPEP. Dentro das diversas metaheurísticas existentes, as que são baseadas em comportamentos da natureza demonstram ser uma opção a ser considerada devido aos resultados obtidos (KAR, 2016). Esses algoritmos bioinspirados podem ser classificados em quatro categorias: Inteligência de Enxames (*Swam Intelligence* - SI); Computação Evolucionário (*Evolutionary Computing* - EC); Redes Neurais Artificiais (RNA) (RUSSELL; NORVIG, 2010) e Sistemas

Imunológicos Artificiais (*Artificial Immune Systems* - AIS) (CASTRO; TIMMIS, 2002) e pertencem a linha de pesquisa de Computação Natural.

Dentre essas quatro categorias, três são bastante utilizadas no PPEP: a SI e a EC. Os algoritmos pertecentes a classe de inteligência por enxames são observações do comportamento de seres vivos sociais, como os peixes, formigas, aves, cupins, entre outros (PARPINELLI; LOPES, 2011b). A característica mais marcante desses seres vivos é a interação do grupo para resolver problemas complexos. Essa característica é conhecida como comportamento emergente. Dois algoritmos muito conhecidos que pertecem a classe de SI é a Otimização por Colônia de Formigas (*Ant Colony Optimization* - ACO) (DORIGO; CARO, 1999) e a Otimização por Enxame de Partículas (*Particle Swarm Optimization* - PSO) (POLI; KENNEDY; BLACKWELL, 2007). Porém, existem muitos outros algoritmos que se encaixam nessa categoria, como o algoritmo baseado em cardumes de peixes (*Fish School Search* - FSS) (FILHO et al., 2009), Otimização por Colônias de Abelhas Artificiais (*Artificial Bee Colony* - ABC) (KARABOGA et al., 2014) e muitos outros (PARPINELLI; LOPES, 2011b), demonstrando que os comportamentos de diferentes exemplos biológicos podem ser utilizados em diferentes problemas de otimização.

A Computação Evolucionária é baseada na teoria da evolução de Darwin, tendo como princípio a sobrevivência dos mais fortes da população. O algoritmo da computação evolutiva mais conhecido foi desenvolvido por Holland, o Algoritmo Genético (*Genetic Algorithm* - GA) (HOLLAND, 1992). Além do algoritmo genético, existem outros algoritmos que são classificados como algoritmos evolutivos, como Evolução Diferencial - DE (*Differential Evolution*) (PRICE; STORN; LAMPINEN, 2005), os Algoritmos Coevolutivos (BOUSSAïD; LEPAGNOT; SIARRY, 2013), a Estratégia Evolutiva (BÄCK; SCHWEFEL, 1993) e os Coevolutivos com Inspiração Ecológica (*Ecological Optimization* - ECO) (PARPINELLI; LOPES, 2011a).

No presente trabalho o algoritmo de Evolução Diferencial (DE) foi escolhido. O DE é conhecido como um algoritmo capaz de encontrar resultados superiores aos outros algoritmos canônicos em problemas de domínio contínuo, que é o caso do PPEP. As suas características são detalhadas na próxima subseção.

2.3.1 Evolução Diferencial

O DE é uma metaheurística populacional bastante popular para otimização de problemas de domínio contínuo (BOUSSAïD; LEPAGNOT; SIARRY, 2013). O DE foi criado em 1997 por Storn e Price (STORN; PRICE, 1997) para a solução do problema de polinomiais de Chebyshev e provou ser uma estratégia interessante para a solução de vários problemas classificados na categoria de NP-Completos.

Por ser um algoritmo da Computação Evolucionária, ele possui uma população

de indivíduos (N vetores de solução) de dimensionalidade D, sendo cada indivíduo uma possível solução para o problema em questão. Cada indivíduo é avaliado por uma função objetivo que qualifica cada solução encontrada. A ideia de evolução da população é gerar novos candidatos com base na perturbação da combinação de diferentes indivíduos da população (PRICE; STORN; LAMPINEN, 2005). Caso o candidato gerado for melhor que a solução da população, ele segue para a próxima geração. Caso contrário, permanece o indivíduo da geração atual.

O mecanismo responsável por gerar os novos indivíduos é chamado de mutação, que sofre influência de outro mecanismo conhecido como *crossover*. Várias versões de mutação foram propostas por Price, Storn e Lampinen (2005) e são comumente rotuladas como DE/x/y/z, sendo DE o identificador de evolução diferencial, x representa o vetor base (que será perturbado), y a quantidade de indivíduos que participam do processo de perturbação e z o modelo de *crossover* utilizado (PRICE; STORN; LAMPINEN, 2005). Apesar de existirem diversas abordagens, a versão canônica do algoritmo é conhecido como DE/*rand*/1/bin, sendo o termo *rand* vindo de *Random* (aleatório) para o vetor base que será modificado, o parâmetro 1 referencia o uso de um indivíduo que servirá como base para a modificação dos valores e *bin* para o modelo binomial de *crossover*.

Além da versão canônica do algoritmo, diversas versões surgiram para melhorar o resultado obtido pelo algoritmo. Uma das mais conhecidas versões é a DE/*best*/1/bin, onde a diferença é que ao invés do indivíduo que irá compor o novo vetor de solução ser escolhido de maneira aleatória, será selecionado o indivíduo elitista, ou seja, o que possuir melhor valor de *fitness*. Dessa forma espera-se que o algoritmo seja guiado sempre pela melhor solução encontrada até a geração atual.

Nessa dissertação é utilizada também uma variação do DE/current-to-rand/1/bin ao qual o valor de dois indivíduos aleatórios são somados ao presente e a versão DE/current-to-best/1/bin ao qual utiliza informações do melhor indivíduo para perturbar o atual. A composição de cada uma das versões pode ser vista na Tabela 3. A utilização dessas versões é devido ao fato de serem alterações dos canônicos mais conhecidos (DE/-rand/1/bin e DE/best/1/bin) e por terem comportamentos diferentes durante o período de otimização.

Tabela 3 – DE Mecanismos de Mutação.

Abordagem	Equação
$\text{DE}_{best/1/bin}$	$\vec{w} = \vec{x}_{best} + F \cdot (\vec{x}_{r1} - \vec{x}_{r2})$
$\text{DE}_{rand/1/bin}$	$\vec{w} = \vec{x}_{r1} + F \cdot (\vec{x}_{r2} - \vec{x}_{r3})$
$\text{DE}_{curr-to-best}$	$\vec{w} = \vec{x} + F \cdot (\vec{x}_{r1} - \vec{x}_{r2}) + F \cdot (\vec{x}_{best} - \vec{x})$
$\text{DE}_{curr-to-rand}$	$\vec{w} = \vec{x} + F \cdot (\vec{x}_{r1} - \vec{x}_{r2})$

Da Rabela 3, $\vec{w}$ é o vetor a ser modificado, também conhecido como vetor mutante,

$\vec{x}$ é o vetor alvo (o atual da população), $\vec{x}_{best}$ o melhor indivíduo da população, $\vec{x}_{ri}$ um indivíduo selecionado aleatoriamente e F o parâmetro conhecido como fator de mutação. Para que a alteração ocorra em uma dimensão do indivíduo é necessário que a taxa de *crossover* seja atingida (CR).

As versões do DE possuem quatro parâmetros que devem ser informados antes de sua execução (*a priori)*: o tamanho da população (N), a probabilidade de *crossover* (CR), o fator de mutação (F) e a quantidade de avaliações (MAX-AV). O Algoritmo 1 apresenta como é o DE/*best*/1/bin.

Algorithm 1 Evolução Diferencial com *best*/1/*bin*

```
1: Configura os parâmetros : N, F, CR e MAX-AV
2: Inicializa a população com soluções candidatas aleatórias x_i
3: Avalia a população com a função objetivo f(x_i)
4: enquanto Critério de parada não sastifeito faça {Número máximo de avaliações}
5:     Pop_temp = ∅
6:     para i = 1 até n faça
7:         Seleciona o melhor indivíduo x_best
8:         Selecione aleatoriamente r1, r2 ∈ N com r1 ≠ r2 ≠ i
9:         Selecione uma dimensão aleatoriamente p ∈ D
10:        para j = 1 até D faça
11:            se (j == p ∨ rand(0,1) ≤ CR) então
12:                y_j = x_best_j + F · (x_r1,j − x_r2,j)
13:            senão
14:                y_j = x_ij
15:            fim se
16:        fim para
17:        Avalia a solução y
18:        se f(y) é melhor do que f(x_i) então
19:            Pop_temp = y
20:        senão
21:            Pop_temp = x_i
22:        fim se
23:    fim para
24:    Pop = Pop_temp
25:    Memorize a melhor solução encontrada até o momento
26: fim enquanto
27: Reportar os resultados obtidos
```

Inicialmente são configurados os parâmetros *a priori* conforme na linha 1. A inicialização da população é feita com números aleatórios na linha 2. Na linha 3 esses indivíduos iniciais são avaliados com a função de *fitness* do problema. A partir da linha 4 inicia-se a fase de evolução da população. A população irá evoluir conforme o número máximo de avaliações configuradas em MAX-AV. Uma população temporária Pop_{temp} é criada na linha 5 sem qualquer indivíduo. Para todos os indivíduos da população (linha 6) é feita uma nova geração soluções. Na linha 7 o melhor indivíduo da população é selecionado. Na linha 8 são gerados valores inteiros aleatórios que representam os índices dos indivíduos na população e na linha 9 é escolhida uma dimensão que será obrigatoriamente alterada. Para cada dimensão de cada indivíduo um novo valor será gerado com base na probabilidade de ocorrer o operador de *crossover* ou se é a dimensão selecionada anteriormente.

A linha 12 gera um novo indivíduo com base no melhor indivíduo da população x_{best} e com a aplicação do fator F na diferença dos indivíduos selecionados aleatoriamente r_1 e r_2 de acordo com a dimensão j. Quando o processo de mutação é encerrado, o novo indivíduo é avaliado e seu *fitness* é gerado na linha 17. Caso esse indivíduo seja melhor que seu antecessor, ele é adicionado na Pop_{temp}. Se o seu *fitness* for pior que o antecessor, esse indivíduo é descartado e o antecessor é adicionado na Pop_{temp}. Por fim, na linha 24, a população Pop é substituída pela Pop_{temp} e um novo ciclo é iniciado.

2.4 DIVERSIDADE EM ALGORITMOS BIOINSPIRADOS

O sucesso de um algoritmo em determinado problema depende do equilíbrio entre seus mecanismos de diversificação e intensificação (BOUSSAïD; LEPAGNOT; SIARRY, 2013). Rotinas de diversificação são aquelas que exploram globalmente o espaço de busca sendo o objetivo é identificar possíveis áreas promissoras. Já as rotinas de intensificação são responsáveis pela exploração local do espaço de busca, efetuando uma busca mais intensiva em determinados pontos, tendo como objetivo refinar a busca. As ações das rotinas de diversificação e intensificação são opostas entre elas, ou seja, quanto maior a intensificação, menor a diversificação e vice-versa. A tendência para os extremos dessas rotinas é prejudicial ao processo de otimização (CORRIVEAU et al., 2013), devido ao fato que se um algoritmo tem muita diversificação ele se assemelha a uma busca aleatória, removendo dessa forma a inteligência do algoritmo. Já o excesso de intensificação leva à perda de diversidade, aumentando a chance da otimização estagnar em um ponto local e ter uma convergência prematura.

Uma estratégia comum adotada em metaheurísticas é o uso de diversificação na etapa inicial do processo de otimização e gradualmente inserir a intensificação. Dessa forma o algoritmo consegue manter a diversidade por um certo período, diminuindo as chances de ter uma convergência prematura. A convergência prematura pode levar à estagnação da otimização para um ponto local do espaço de busca, prejudicando diretamente a qualidade da solução e impedindo que o ponto ótimo seja obtido. A perda de diversidade pode ser relacionada por dois fatores: pressão de seleção e deriva genética, fatores muito comuns em Algoritmos Evolutivos (JR.; ARAJO, 2010). A pressão de seleção é resultado do processo de seleção de indivíduos da próxima geração, substituindo soluções de baixa qualidade por soluções de qualidade superior. Dessa forma, a diversidade da população é perdida e a tendencia é convergir rapidamente para soluções semelhantes. Já a deriva genética é responsável por propagar para as próximas gerações características por meio dos mecanismos de reprodução. Em razão desses fatores é importante que rotinas de diversificação sejam exploradas com o intuito de minimizar o impacto da pressão de seleção

e da deriva genética, possibilitando a exploração de novos locais no espaço de busca antes da perda total da diversidade populacional. Dessa forma soluções de baixa qualidade não serão eliminadas precocemente, pois essas soluções podem ter informações úteis que podem ser utilizadas durante o período de otimização. Por sua vez, quando ocorre o excesso de diversificação o algoritmo pode vir a explorar superficialmente o espaço de busca ou ainda se tornar um algoritmo de busca aleatória (OLIVEIRA; FREITAS; GUIMARÃES, 2012).

Os mecanismos de manutenção de diversidade podem ajudar o processo de otimização com a exploração de diversas áreas e aumentar as chances de identificar pontos ótimos globais ou locais (GUPTA; GHAFIR, 2012). Alguns dos métodos mais conhecidos da literatura para a preservação da diversidade populacional são:

- *Fitness Sharing*, ao qual mensura-se regiões do espaço de busca que estão densamente povoadas (SARENI; KRAHENBUHL, 1998), compartilhando o valor de *fitness* com os indivíduos de uma região;
- *Crowding*, ao qual indivíduos que possuem um alto grau de similaridade são substituídos por novos indivíduos gerados;
- *Clearing*, diferentemente do *fitness sharing*, a técnica de *Clearing* não compartilha o *fitness* entre os indivíduos de uma aglomeração, mas atribui o *fitness* somente ao melhor indivíduo do grupo;
- *Restricted Mating*, ao qual o operador de seleção somente seleciona indivíduos que possuam um grau de similaridade baixo; entre outros que podem ser consultados em (SHIR, 2012).

No presente trabalho foram escolhidos duas funções de diversificação, a estratégia conhecida como *Generation Gap* analisada por Jong (1975) e a mutação Gaussiana Koenig (2002). Dentre os trabalhos relacionados encontrados na literatura, nenhuma dessas estratégias foram utilizadas, tornando-se dessa forma, a aplicação e análise delas uma pequena contribuição do presente trabalho.

A estratégia *Generation Gap* é uma das abordagens conhecidas que modifica a maneira do algoritmo evoluir a população. Normalmente, a cada nova geração de indivíduos, todos as soluções da população são descartadas e substituídas pelos seus descendentes (JONG; DE; SARMA, 1992). Com o uso do *Generation Gap*, um fator G é definido no intervalo $[0, 1]$. A cada nova geração, um percentual da população referente ao índice G é atualizada, mantendo partes da população atual para a próxima geração, promovendo dessa maneira maior diversidade, desacelerando a convergência do algoritmo. Os indivíduos que não serão substituídos são escolhidos aleatoriamente.

A segunda abordagem, conhecida como mutação Gaussiana, modifica o operador de mutação do algoritmo, gerando valores aleatórios dentro de uma distribuição normal (KOENIG, 2002), sendo o valor médio o atual valor do indivíduo e o desvio padrão é um parâmetro a ser definido. Dessa forma ameniza-se a deriva genética e diminui as chances de se obter uma convergência prematura. A Equação 2.11 apresenta a função de densidade Gaussiana

$$f_{Gaussian}(0, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\vec{x}^2}{2\sigma^2}} \tag{2.11}$$

sendo σ^2 a variância e $\vec{x}$ o indivíduo. A abordagem de mutação Gaussiana é aplicada estritamente para representações de domínio contínuo.

A diversidade pode ser mensurada de duas maneiras: mensurando a nível genotípico (MDG) e a nível fenotípico (MDF) (HERRERA; LOZANO, 1996). A MDG é mensurada através da diferença do material genético de cada indivíduo, ou seja, das variáveis que compõem uma possível solução. Dessa forma, a distância entre os indivíduos da população é calculada no espaço das variáveis do problema (CORRIVEAU et al., 2012). Com uma alta taxa de MDG pode-se assumir que há uma distribuição espacial alta, o que significa que há uma exploração global do espaço de busca. A partir do momento que a distribuição espacial diminui, a MDG também diminui, resultando em uma exploração local do espaço de busca (CORRIVEAU et al., 2012). A MDF leva em consideração o *fitness* dos indivíduos da população, ou seja, do valor da função objetivo e penalidades de uma possível solução. Quando a MDF está em níveis altos significa que o algoritmo ainda não obteve sua convergência. Caso o índice for baixo, significa que o algoritmo convergiu para uma solução caso o objetivo seja minimizar a função objetivo (CORRIVEAU et al., 2013). Caso seja uma função de maximização, o índice alto de MDF significará a convergência do algoritmo.

Apesar de ambas as medidas serem importantes para controlar a proporção entre diversificação e intensificação (HERRERA; LOZANO, 1996), a MDG é fortemente dependente do modelo de representação das soluções de uma população, que podem ser binárias, de domínio discreto ou contínuo, em forma de grafos, entre outros. Diferentes medidas podem ser encontradas em (CORRIVEAU et al., 2012). No presente trabalho foi adotada a MDG proposta por Corriveau (CORRIVEAU et al., 2013) por ter sido relatada pelo autor como uma métrica que ajuda na identificação do comportamento dos algoritmos e por ser indicada para problemas de domínio contínuo, que é o caso do PPEP.

A Equação 2.12 apresenta sua formulação. Como observado por (CORRIVEAU et al., 2013), cada problema possui uma característica diferente e a simples alteração dos valores de parâmetros pode afetar diretamente no comportamento do algoritmo. A Equação 2.12 mostra a MDF que mostrou ser uma ferramenta de análise de comportamento

do algoritmo, ajudando dessa forma na identificação das capacidades de intensificação e diversificação.

$$MDG = \frac{\sum_{i=1}^{N-1} ln\left(1 + \min_{j[i+1,N]} \frac{1}{D}\sqrt{\sum_{k=1}^{D}(x_{i,k} - x_{j,k})^2}\right)}{NMD} \tag{2.12}$$

Sendo D a quantidade de dimensões da solução, N o tamanho da população e x a solução candidata. O fator de normalização utilizado é feito pela variável NMD e corresponde ao valor máximo de diversidade até o momento. O valor 1 de MDG representa o valor máximo de diversidade enquanto que o valor 0 de MDG corresponde a convergência da população. Outras medidas genotípicas podem ser consultadas em (CORRIVEAU et al., 2012). Além da capacidade de análise de diversidade, essas medidas podem ser utilizadas como informações de *feedback* para ajustes de parâmetros e rotinas dos algoritmos.

2.5 TRABALHOS RELACIONADOS

Existem diversos trabalhos que aplicam algoritmos bio-inspirados ao PPEP, em diversos modelos de representação de proteínas e funções de energia. Apesar de existirem diferentes abordagens para a solução do PPEP, em nenhum deles foram encontradas análises de diversidade populacional ou fenotípica. Dessa forma, a tarefa de identificar uma possível convergência prematura se torna difícil, dificultando também a análise da efetividade dos parâmetros definidos.

Dentre as abordagens que fazem predições *Ab Initio* com ângulos de torções da cadeia principal e cadeia lateral, e utilizam CHARMM, podem ser destacados alguns trabalhos. O algoritmo proposto em (O; TINOS, 2010) é uma versão modificada de GA, chamado de SSORIGA. Em sua proposta são utilizadas duas técnicas de diversificação. A primeira é conhecida como Imigrantes Aleatórios ao qual uma parte da população é substituída por novos indivíduos gerados aleatoriamente, sendo a parte substituída um percentual pré-definido. O segundo mecanismo aplicado é uma modificação do primeiro, conhecido como técnica Auto-Adaptativa de Imigrantes Aleatórios, sendo que o percentual de indivíduos a serem substituídos é ajustado durante o processo de otimização de maneira auto-adaptativa. As proteínas experimentadas foram a 1PLW (5 aminoácidos), 1CRN (41 aminoácidos) e 1ENH (54 aminoácidos) retiradas do PDB.

O algoritmo baseado no movimento das bactérias conhecido como *Bacterial Foraging Optimization Algorithm* (BFOA) foi aplicado ao PPEP em (PAL, 2014) para predizer a estrutura da proteína 1PLW. Nesse trabalho o BFOA foi aplicado canonicamente, sem nenhuma rotina de diversificação ou intesificação. Em (ROMERO, 2010) foi utilizado o

algoritmo conhecido na literatura como NSGA-II em modelo de ilhas. Nessa abordagem o autor utilizou de técnicas de otimização multi-objetivo em duas formulações. A primeira separa a função de energia em termos de ligantes e não ligantes, tendo dessa forma 2 objetivos a serem otimizados. Posteriormente a força de Van der Waals foi otimizada separadamente dos não ligantes, formando três objetivos a serem otimizados. As proteínas utilizadas para testes de predição foram a 1PLW, 1ZDD (35 aminoácidos), 1CRN, 1ROP (63 aminoácidos) e 1UTG (70 aminoácidos).

Em (CUTELLO; NARZISI; NICOSIA, 2006)(CUTELLO; NARZISI; NICOSIA, 2008) também utilizaram uma abordagem multiobjetivo para predizer as proteínas 1PLW, 1ZDD, 1CRN, 1ROP, 1UTG, 1R69 (69 aminoácidos) e 1CTF (74 aminoácidos). O algoritmo aplicado foi o Algoritmo Evolucionário modificado com mecanismos inspirados em operadores imunológicos (clonagem e hiper mutação). A otimização multi-objetivo foi definida separando a função de energia em termos de ligantes e não-ligantes. O algoritmo recebeu o nome de I-PAES. Outra abordagem multi-objetivo semelhante ao I-PAES foi proposta em (VENSKE et al., 2016). O algoritmo DE foi utilizado para a predição das proteínas 1PLW, 1ZDD, 1CRN, 1ROP, 1CTF e 2MLT (27 aminoácidos). Modificações foram feitas no algoritmo canônico, transformando-o em adaptativo. A modificação foi feita no operador de mutação, sendo possível ao algoritmo escolher entre as abordagens $DE/rand/1/bin$, $DE/rand/2/bin$ e $DE/nonlinear$ ao longo do período de otimização. Essa técnica é chamada de *Probability Matching*, sendo a escolha de determinada estratégia feita de acordo com a sua taxa de sucesso durante as gerações. O algoritmo é conhecido como ADEMO/D.

Apesar de muitos dos algoritmos propostos para resolver o PPEP utilizarem técnicas de diversificação para melhor explorar o espaço de busca, em (TANTAR et al., 2007) é utilizado um GA híbrido aplicado em uma *grid* de computadores, aumentando o poder computacional aplicado ao PPEP. Junto com o GA foi utilizada a estratégia de *Hill Climbing*, conhecida por ser uma estratégia de busca local e de intensificação. Os peptídios utilizados como testes foram o 1L2Y (20 aminoácidos) e a $\alpha - ciclodextrin$.

Além da utilização da função de energia CHARMM, trabalhos interessantes foram encontrados utilizando as funções de energia AMBER, GROMOS e ROSETTA. Em (**??**) utilizaram a função de energia GROMOS em conjunto com uma versão modificada do AG utilizando *Crowding*. As proteínas utilizadas como teste nessa abordagem foram a 18ALA e 23ALA. A função de energia AMBER foi utilizada em (INOSTROZA-PONTA; FARFáN; DORN, 2015) utilizando um algoritmo memético para as proteínas 2EVQ (12 aminoácidos), 1K43 (14 aminoácidos), 1DEP (15 aminoácidos), 1E0Q (17 aminoácidos), 1RPV (19 aminoácidos) e 1L2Y. No trabalho de (CALVO; ORTEGA; ANGUITA, 2011) foi proposto o algoritmo PITAGORAS-PSP que também faz uso da função energia AM-

BER. Nesta abordagem, foi feito uso da técnica de otimização multi-objetivo, separando a função de energia entre ligantes e não-ligantes. O algoritmo proposto conseguiu resultados interessantes no CASP8. Uma terceira abordagem usando AMBER foi feita em (DORN; BURIOL; LAMB, 2011), sendo o algoritmo utilizado o AG com *path-relinking*. As proteínas utilizadas para a predição foram a 1ZDD e 2L56 (18 aminoácidos).

A função de energia ROSETTA foi utilizada em (BORGUESAN et al., 2015) em conjunto com uma base de dados de probabilidades de ângulos intitulada como *Angle Probability List* (APL). A APL e a função de energia ROSETTA foram aplicados com GA e PSO utilizando inércia. Nos testes realizados utilizando GA foi aplicada uma técnica de estruturação da população, dividindo ela em 3 categorias. A primeira categoria (classe A) estão os indivíduos elitistas que representem 10% da população. Os próximos 50% da população são os que possuem bons valores de *fitness* e por isso são classificados como classe B. Por fim está o restante da população, que possuem os piores valores de *fitness*, classificados como classe C. O controle de diversidade feito pelo GA consiste em manter os indivíduos da classe A, modificar os indivíduos da classe B e substituir os indivíduos da classe C por novos que vem da APL. Com essas abordagens, foi possível identificar que as versões que a utilização da APL ajudou na geração de indivíduos mais semelhantes a estrutura nativa. Em (INOSTROZA-PONTA; FARFáN; DORN, 2015) a função de energia ROSETTA é também utilizada com a APL. O método de busca aplicado pelos autores é um algoritmo memético que utiliza de diversos operadores para a busca global e como busca local é utilizado o método de *Simulated Annealing* (SA). Outros modelos de referência que utilizam o ROSETTA em combinação com uma base de dados de informação são: QUARK(XU; ZHANG, 2012), Zhang-Server (ZHANG et al., 2016) e BAKER-ROSETTA(KIM; CHIVIAN; BAKER, 2004). Todos esses modelos são reconhecidos como estado-da-arte devido aos bons resultados que obtiveram no CASP. Porém, eles não são comparados nesse trabalho devido a maior quantidade de dados biológicos que são agregados ao algoritmo, algo que não é feito nessa dissertação.

Apesar de existirem diversas funções de energia, nem sempre é possível comparar os valores de energia obtidos entre elas. Isso se dá devido a diferença das unidades de medidas utilizadas no ROSETTA, AMBER, CHARMM e GROMOS. Enquanto que AMBER, CHARMM e GROMOS mensuram um polipeptídio em kcal mol^{-1}, a função de energia ROSETTA utiliza de uma medida própria chamada de *Rosetta Energy Unit* (REU).

2.6 RESUMO DO CAPÍTULO

Nesse capítulo foram apresentados os principais conceitos biológicos e tecnológicos necessários para o entendimento desse trabalho de dissertação. Como pode ser visto, as

proteínas são macromoléculas complexas, formadas por aminoácidos e que, ao atingirem suas conformações nativas, elas exercem importantes funções biológicas. Dessa forma, o PPEP tem sido classificado de um dos problemas mais importantes e desafiadores da bioinformática, pois contribuições para o entendimento das estruturas das proteínas podem ajudar na produção de drogas para prevenção, tratamento ou até mesmo cura de doenças como Alzheimer, Parkinson e alguns tipos de cânceres.

Nesse capítulo também foram apresentados os principais conceitos que definem o PPEP, suas representações computacionais e as principais funções de energia utilizadas pela literatura. É importante ressaltar que cada função de energia calcula a energia de uma proteína de maneira diferente, podendo variar a unidade de medida como visto: kcal mol^{-1} para CHARMM e AMBER e REU para ROSETTA. Outra métrica importante é conhecida como o RMSD ao qual verificar o quão parecida está a proteína predita da que já possui sua conformação nativa conhecida.

Como algumas instâncias do PPEP são tratadas como problemas NP-completos, a utilização de metaheurísticas torna-se viável e atrativa já que métodos exatos não conseguem predizer uma estrutura terciária em um tempo aceitável. Muitos dos trabalhos encontrados da literatura utilizam de algoritmos bioinspirados para a solução do PPEP. Porém, esses algoritmos são dependentes do equilíbrio das capacidades de intensificação e diversificação durante o período de busca por uma solução. Apesar dos diferentes autores reconhecerem essa necessidade, em nenhum dos trabalhos revisados nesse capítulo utilizaram alguma métrica para mensuração da diversidade populacional. À vista disso, foi apresentado uma revisão de alguns mecanismos de diversidade conhecidos. Além disso, uma métrica de monitoração da diversidade genotípica proposta por Corriveau et al. (2013) é explorada, justificando a sua aplicação no presente trabalho.

3 MODELO *IN SILICO* PARA PREDIÇÃO DE ESTRUTURA DE PROTEÍNAS

Nesse capítulo é abordado o modelo *in silico* para a solução do PPEP descrito em (NARLOCH; PARPINELLI, 2017a). Segundo Dorn et al. (DORN et al., 2014) são necessárias três definições para um modelo de predição *Ab Initio* de proteínas: **(a)** a representação computacional de uma proteína; **(b)** uma função de energia potencial para guiar o processo de busca; **(c)** uma metodologia de busca de conformações.

3.1 REPRESENTAÇÃO DA PROTEÍNA

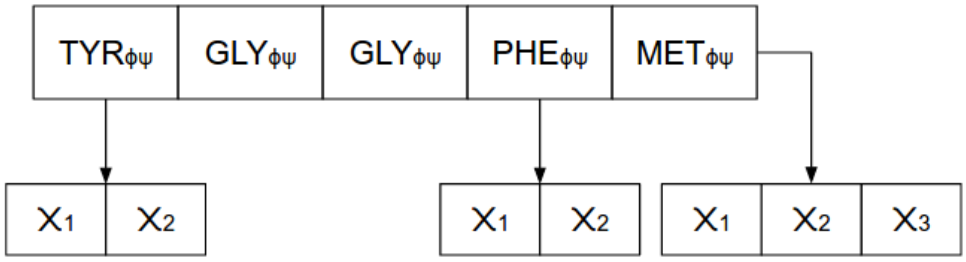
Existem diversas maneiras de representar uma proteína computacionalmente conforme visto na Subseção 2.2.1. Como o presente trabalho é aplicado para a predição terciária de proteínas, a modelagem realizada se dá a nível atômico. Dentre as representações atômicas mais utilizadas está a de ângulos e torções da cadeia principal e da cadeia lateral. A escolha desse tipo de representação é devido ao fato que uma cadeia polipeptídica pode ser representada por um conjunto de ângulos conhecidos entre as ligações da cadeia principal conforme explicado na Subseção 2.2.1. São três os ângulos que formam a cadeia principal da proteína: o ângulo ϕ presente entre a ligação do nitrogênio e carbono α; o ângulo ψ presente entre a ligação do carbono α e o carbono; e o ângulo ω presente na ligação entre carbono e nitrogênio conforme a Figura 3 da Subseção 2.1.1.

Além dos três ângulos da cadeia principal, existem os ângulos da cadeia secundária que também são levados em consideração. Esses ângulos são representados por χ, tendo uma variação da quantidade de χ em cada aminoácido de acordo com suas características. A quantidade de ângulos por resíduo foi apresentado na Subseção 2.1.1 pela Tabela 2.

A definição da representação computacional de uma proteína é importante pois irá servir como entrada de dados no algoritmo proposto de otimização. Como exemplo, a Figura 10 apresenta como a proteína *Met-Enkephalin* (1PLW) com 5 aminoácidos e 17 ângulos para serem otimizados é representada pelo algoritmo. Dessa forma, cada resíduo é transformado em um objeto que possui os ângulos ϕ e ψ da cadeia principal e uma quantidade variável de ângulos χ. O ângulo ω não é representado por ter seu valor fixo em 180°, não sendo necessário otimiza-lo.

Esses ângulos teoricamente podem variar entre 180° e −180°. Entretanto, com o objetivo de diminuir o espaço de busca, no presente trabalho utiliza-se da classificação conhecida como DSSP-8 (KABSCH; SANDER, 1983), ao qual determinados tipos de estruturas secundárias possuem uma variação limitada de possíveis ângulos a serem assumidos conforme a Tabela 4. Para saber as possíveis estruturas secundárias que a com-

Figura 10 – Representação da Proteína 1PLW.



binação de aminoácidos pode formar, o preditor de estrutura secundária conhecido como Scratch (POLLASTRI et al., 2002) é utilizado.

Tabela 4 – Classificação DSSP-8

Estrutura Secundária	Limites ϕ	Limites ψ
H (α-helix)	$[-67°, -47°]$	$[-57°, -37°]$
B (β-bridge)	$[-130°, -110°]$	$[110°, 130°]$
E (β-strand)	$[-130°, -110°]$	$[110°, 130°]$
G (3-10-helix)	$[-59°, -39°]$	$[-36°, 16°]$
I (pi-helix)	$[-67°, -47°]$	$[-80°, -60°]$
T (turn)	$[-180°, 180°]$	$[-180°, 180°]$
S (bend)	$[-180°, 180°]$	$[-180°, 180°]$
U (undefined)	$[-180°, 180°]$	$[-180°, 180°]$

3.2 FUNÇÃO DE ENERGIA

Na presente dissertação foram utilizadas três funções de energia potencial para a determinação das estruturas 3D de proteínas. Dessa forma, é possível comparar a eficiência entre as funções de energia utilizando o mesmo método de otimização e também comparar os resultados obtidos com trabalhos encontrados na literatura. Todas as energias utilizam de componentes físicos e químicos para o cálculo de energia. As funções de energia CHARMM e AMBER, apesar de serem diferentes, elas seguem um padrão conforme Equação 3.1.

$$E_{Total} = \sum_{ligações} + \sum_{ângulos} + \sum_{torções} + \sum_{não-ligados} \tag{3.1}$$

As funções de energia potencial utilizadas foram a CHARMM (BROOKS et al., 2009), a AMBER (CORNELL et al., 1995) e ROSETTA (ROHL et al., 2004). Apesar das funções de energia CHARMM e AMBER seguirem a forma apresentada pela Equação 3.1, elas possuem maneiras diferentes de calcular cada componente conforme visto na Subseção 2.2.4.

Apesar das duas funções de energia serem bastante utilizadas na literatura, não foram encontrados trabalhos que comparassem as funções de energia utilizando a mesma metodologia de busca. Dessa forma, esse estudo permite identificar se há alguma diferença em utilizar a energia dada pela função CHARMM, AMBER ou ROSETTA. Para utilizar as funções de energia CHARMM e AMBER o pacote de dinâmica molecular conhecido como TINKER[1] foi necessário. O TINKER provê ferramentas para a tradução de modelos de ângulos e torções para planos cartesianos que são utilizados para o cálculo da energia. Para a função de energia ROSETTA foi utilizada a ferramenta PyROSETTA[2]. Um fator importante a ser considerado é que o pacote de dinâmica molecular TINKER utiliza de muitos arquivos de texto para a criação e avaliação das proteínas, causando um grande acesso ao disco rígido do equipamento e consequentemente aumentando o tempo de processamento do algoritmo. Por outro lado, o PyROSETTA permite que todo o processamento seja feito em memória, evitando o acesso a disco e otimizando o tempo de criação e avaliação de uma proteína.

3.3 MÉTODO DE BUSCA

Como o modelo de PPEP presente nessa dissertação é tratado como um problema pertencente a classe NP-Completo, algoritmos exatos tornam-se inviáveis devido ao tempo de processamento necessário para encontrar a solução ótima. Devido às limitações desses algoritmos exatos, as metaheurísticas tornam-se interessantes por conseguirem encontrar soluções boas a um custo computacional aceitável. Muitas metaheurísticas podem ser utilizadas no PPEP como citados na Seção 2.5, porém a escolhida neste trabalho é a ED, por ser considerado um algoritmo competitivo para a solução de problemas de domínio contínuo como em (NARLOCH; PARPINELLI, 2017a). Além disso, foram apresentados bons resultados pelo algoritmo ADEMO/D (VENSKE et al., 2016) quando comparado com outras metaheurísticas.

Apesar de suas qualidades em problemas de domínio contínuo, algumas versões do DE são bastantes agressivas e tendem a ter um desequilíbrio entre suas rotinas de intensificação e diversificação. Com essa falta de equilíbrio o algoritmo a ser pego por um atrator local do espaço de busca, forçando a população a convergir de maneira prematura. Dessa forma, é necessário monitorar a diversidade da população durante o processo de otimização para identificar tais comportamentos.

Na Subseção 3.3.1 é apresentado o algoritmo DE com o mecanismo de mutação rand/1/bin e duas estratégias estratégias de diversificação: *Generation Gap* e mutação Gaussiana. Já na Subseção 3.3.2 é apresentada uma abordagem que, de maneira determi-

[1] https://dasher.wustl.edu/tinker/
[2] http://www.pyrosetta.org/

nista, utiliza mais de um mecanismo de mutação no processo de busca, sem a aplicação de estratégias de diversificação.

3.3.1 Evolução Diferencial com Estratégias de Diversificação

No presente trabalho são empregadas diferentes variações do algoritmo DE para o PPEP. Inicialmente são feitas simulações da versão $\mathrm{DE}_{rand/1/bin}$ que utiliza o três indivíduos aleatórios no processo de composição de uma nova solução. Dois mecanismos de diversificação são incorporados a versão canônica: *Generation Gap* (DE_{GG}) e a perturbação gaussiana (DE_{GP}). A função do *Generation Gap* é preservar parte da população durante a criação de novas gerações, modificando dessa forma o operador geracional do algoritmo e diminuindo a pressão da seleção. A segunda modificação é feita no operador de mutação, ao qual dois indivíduos sofrem uma perturbação em seus valores para diminuir a deriva genética. Essa perturbação ocorre entre os limites de uma distribuição Gaussiana. Devido a essa característica, a abordagem é conhecida como mutação Gaussiana (KOENIG, 2002). Ambas as abordagens são discutidas na Seção 2.4. Uma combinação dessas estratégias também é feita, compondo o DE_{GG-GP}. Essas combinações tem como objetivo verificar se a diversidade populacional é mantida por um maior número de gerações e se com a manutenção dessa diversidade o algoritmo tem capacidade de encontrar melhores resultados.

Dessa forma, o algoritmo DE implementado pode ser visualizado no Algoritmo 2. Inicialmente o algoritmo gera uma população de N indivíduos de maneira pseudoaleatória utilizando a biblioteca *Mersenne Twister* na linha 2 e na linha 3 faz a avaliação do *fitness* desses indivíduos utilizando CHARMM, AMBER ou ROSETTA. A cada ciclo do algoritmo, um percentual da população será armazenado no conjunto de indivíduos temporários chamados de Pop_{temp} conforme o valor do parâmetro G (linhas 5 a 8). Após preservar parte da população incia-se a etapa de geração de novos indivíduos através do mecanismo de *crossover* e mutação. É importante ressaltar que indivíduos que foram selecionados para Pop_{temp} não são substituídos, esse controle é feito nas linhas 13 a 15 do pseudocódigo. Na linha 18 é aplicada a fórmula da versão $\mathrm{DE}_{rand/1/bin}$ em conjunto com a mutação Gaussiana nos indivíduos $\vec{x}_{r1}$ e $\vec{x}_{r2}$ conforme o parâmetro GSD definido *a priori*. O indivíduo é avaliado utilizando a função de *fitness* (linha 23) e caso ele seja de melhor qualidade que seu predecessor ele é inserido na população (linha 25). Caso ele seja inferior ao seu predecessor, o novo indivíduo é descartado e o predecessor permanece para a próxima geração (linha 27).

Algorithm 2 Evolução Diferencial $rand/1/bin$ + *Generation Gap* e Mutação Gaussiana

Configura os parâmetros : N, F, CR, G, GSD, MAX-AV
Inicializa a população com soluções candidatas aleatórias $\vec{x_i}$
Avalia a população com a função objetivo $f(\vec{x_i})$
enquanto Critério de parada não satisfeito **faça** {Número máximo de avaliações}
  $Pop_{temp} = \varnothing$
  para $i = 1$ até $|N \cdot G - N|$ **faça**
    $Pop_{temp} = rand(Pop)$
  fim para
  para $i = 1$ **até** $N \cdot G$ **faça**
    Selecione aleatoriamente $r_1, r_2, r_3, r_3 \in N$ com $r_1 \neq r_2 \neq r_3 \neq i$
    Selecione uma dimensão aleatoriamente $p \in D$
    se $x_i \in Pop_{temp}$ **então**
      Próximo
    fim se
    para $j = 1$ **até** D **faça**
      se $(j == p \vee rand(0,1) \leq CR)$ **então**
        $y_j = x_{r_1,j} + F \cdot (gauss(x_{r2,j}) - gauss(x_{r3,j}))$
      senão
        $y_j = x_{ij}$
      fim se
    fim para
    Avalia a solução $\vec{y}$
    se $f(\vec{y})$ é melhor do que $f(\vec{x_i})$ **então**
      $Pop_{temp} = \vec{y}$
    senão
      $Pop_{temp} = \vec{x_i}$
    fim se
  fim para
  $Pop = Pop_{temp}$
  Memorize a melhor solução encontrada até o momento
fim enquanto
Reportar os resultados obtidos

Todas essas quatro versões ($DE_{rand/1/bin}$, DE_{GG}, DE_{GP} e DE_{GG-GP}) são avaliadas pelas funções de energia CHARMM, AMBER e ROSETTA e seus resultados comparados no próximo capítulo (NARLOCH; PARPINELLI, 2017a).

3.3.2 Evolução Diferencial em Cascata

Devido a função de energia ROSETTA, do módulo PyROSETTA, ter melhor tempo de processamento quando comparado ao TINKER, um novo método de busca é definido neste trabalho e aplicado exclusivamente para a função de energia ROSETTA: o $DE_{cascata}$ (NARLOCH; PARPINELLI, 2017b). Esse método de busca consiste em separar a quan-

tidade de gerações em quatro partes e aplicar a cada quarto um mecanismo de mutação diferente. Nesse caso não são utilizados os mecanismos de diversificação como nas abordagens DE_{GG}, DE_{GP} e DE_{GG-GP}.

A aplicação desse método tem como objetivo explorar os pontos positivos de diferentes abordagens de mutação existentes para o algoritmo DE e minimizar suas deficiências. Dentre os mecanismos de mutação utilizados nesse trabalho estão o $\mathrm{DE}_{rand/1/bin}$, $\mathrm{DE}_{best/1/bin}$, $\mathrm{DE}_{curr-to-rand}$ e $\mathrm{DE}_{curr-to-best}$. Espera-se que as versões que utilizem indivíduos aleatórios para a composição de um novo indivíduo ($\mathrm{DE}_{rand/1/bin}$ e $\mathrm{DE}_{curr-to-rand}$) tenham maior capacidade de diversificação quando comparadas às versões que forçam a população a usar valores do melhor indivíduo ($\mathrm{DE}_{best/1/bin}$ e $\mathrm{DE}_{curr-to-best}$). Dessa forma, inicializa-se o processo de otimização com as versões que possuam maior capacidade de diversificação e que ao fim do processo de otimização sejam utilizadas as rotinas que possuem maior capacidade de intensificação para que a população de soluções tenha sua convergência.

Como visto no Algoritmo 3, o $\mathrm{DE}_{cascata}$ mantém a mesma estrutura da versão canônica do DE, a diferença é que nesta versão existe uma alteração entre os mecanismos de mutação utilizados. A cada quarto do processo de otimização esse mecanismo é alterado. A linha 6 faz o controle de qual mecanismo deve ser utilizado. Na linha 10 é aplicado o mecanismo de mutação conforme a variável *estrategia_mutacao*.

Com o intuito de verificar o comportamento de todas as versões propostas e implementadas nesta dissertação, é utilizado o índice de diversidade populacional calculado pela Equação 2.12 abordada na Subseção 2.4. Com esse índice é possível relacionar a convergência do valor de *fitness* com a diversidade existente durante o processo de otimização. Além da correlação com o *fitness*, esse índice se é importante devido a capacidade de identificar a convergência populacional durante o processo de otimização, identificando a capacidade de diversificação e intensificação do algoritmo.

Ao fim do processo de otimização, o algoritmo retorna os ângulos que compõem o melhor indivíduo, permitindo que seja avaliada a qualidade da estrutura obtida. Além do valor de energia encontrado, a estrutura predita pode ser comparada com a estrutura da proteína alvo (caso ela exista na base do PDB) para mensurar o quão eficiente foi o algoritmo. Essa comparação é feita através da métrica RMSD dada pela Equação 2.1 na Seção 2.2.1.

Todas as etapas do processo de otimização podem ser observadas na Figura 11. Inicialmente é necessário definir qual proteína será predita pelo modelo. A estrutura primária é adquirida pelo PDB e é enviada ao *Scratch* para a predição de sua estrutura secundária para ser feita a classificação DSSP-8 (como exemplo visto na Tabela 4), objetivando redimensionar o espaço de busca utilizando de informações contidas na estrutura

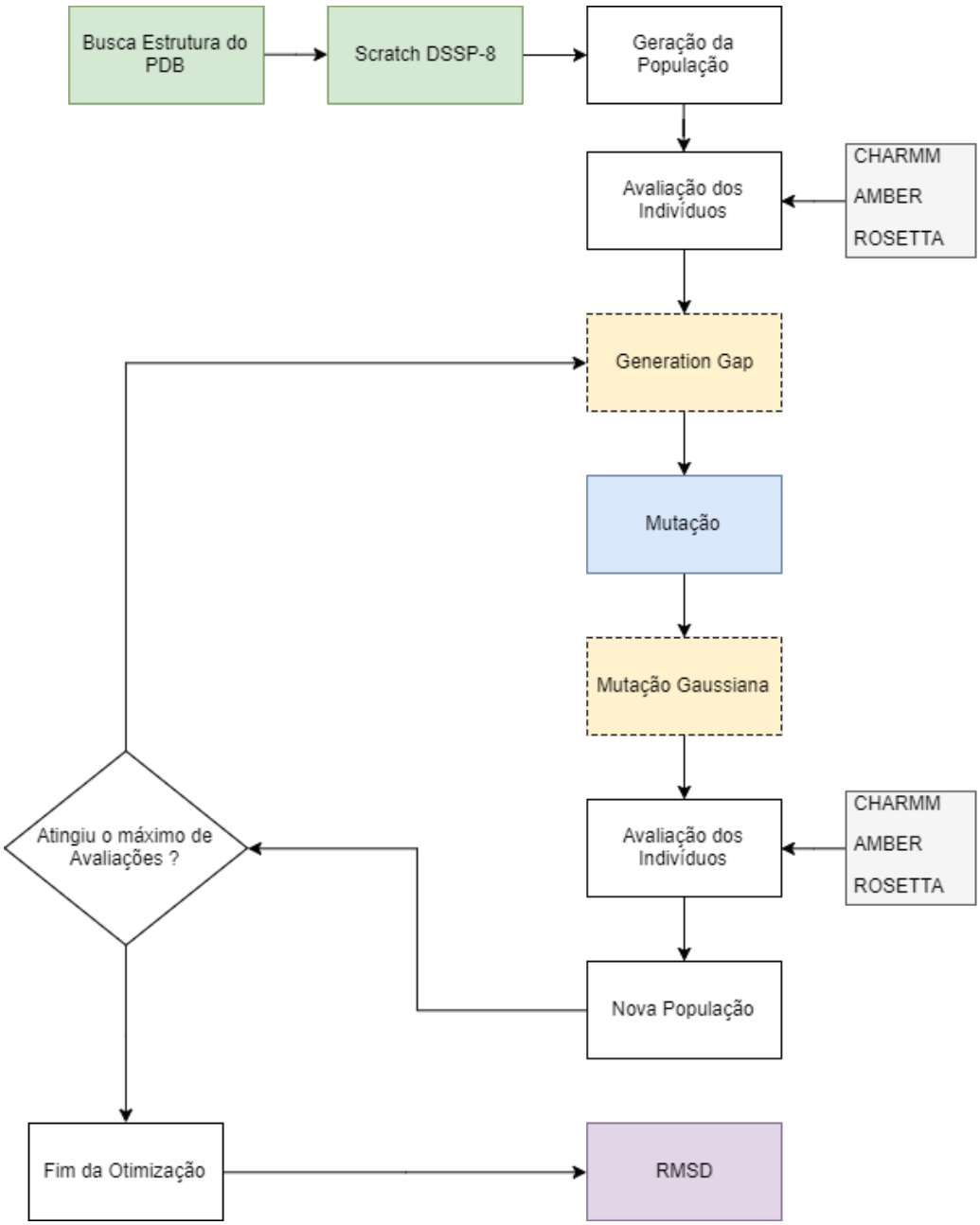
Algorithm 3 Evolução Diferencial em Cascata

1: Configura os parâmetros : N, F, CR, MAX-AV
2: Inicializa a população com soluções candidatas aleatórias $\vec{x_i}$
3: Avalia a população com a função objetivo $f(\vec{x_i})$
4: $estrategiaMutacao = 0$
5: **enquanto** Critério de parada não sastifeito **faça** {Número máximo de avaliações}
6: A cada quarto do número de avaliações mudar $estrategiaMutacao$
7: **para** $i = 1$ **até** N **faça**
8: **para** $j = 1$ **até** D **faça**
9: **se** $(j == p \vee rand(0, 1) \leq CR)$ **então**
10: Aplicar $estrategiaMutacao$ na dimensão y_j
11: **senão**
12: $y_j = x_{ij}$
13: **fim se**
14: **fim para**
15: Avalia a solução $\vec{y}$
16: **se** $f(\vec{y})$ é melhor do que $f(\vec{x_i})$ **então**
17: $Pop_{temp} = \vec{y}$
18: **senão**
19: $Pop_{temp} = \vec{x_i}$
20: **fim se**
21: **fim para**
22: $Pop = Pop_{temp}$
23: Memorize a melhor solução encontrada até o momento
24: **fim enquanto**
25: Reportar os resultados obtidos

primária conforme destacado com retângulo da cor vermelha. Após o retorno da classificação, a sequência primária e a classificação dos resíduos são informados como parâmetros para o algoritmo. A partir desse momento o algoritmo inicia o processo de otimização, sendo que cada um dos indivíduos gerados é avaliado conforme sua energia obtida pelas funções (CHARMM, AMBER ou ROSETTA). Os melhores indivíduos são selecionados para formar a próxima geração. Nos retângulos em tom amarelo e com bordas pontilhadas representam os mecanismos de diversificação utilizados no Algoritmo 2. O retângulo azul representa a etapa de mutação que no Algoritmo 2 é a versão rand/1/bin enquanto que para o Algoritmo 3 é variável conforme as abordagens escolhidas.

Ao fim, é retornado o indivíduo predito que possui a menor energia para comparação e representação gráfica. Esse indivíduo é comparado com a proteína alvo para verificar o quão próximo ele chegou de sua conformação nativa (RMSD), sendo quanto menor o valor em Å , maior a similaridade da proteína predita com a nativa. Essa fase final é sinalizada com o retângulo roxo do fluxograma.

Figura 11 – Fluxograma do Modelo Proposto.



Fonte: Autoria Própria

Com essas informações, os três requisitos para um modelo de predição de estrutura de proteínas são descritos. Inicialmente a definição da representação da proteína serve como entrada para o algoritmo como uma lista de ângulos. Esses ângulos são redimensionados conforme as recomendações da classificação DSSP-8 para a cadeia principal e a biblioteca de rotâmeros para a cadeia secundária. A segunda etapa é a definição da função de energia que serve como fórmula para mensurar o *fitness* dos indivíduos da população. Nesse caso são utilizadas as funções de energia potencial CHARMM, AMBER e ROSETTA. Por fim, é especificado o algoritmo e as alterações feitas para amenizar problemas que podem vir a ocorrer devido a pressão da seleção e deriva genética, seja através

de rotinas de diversificação como o *generation gap* e a mutação Gaussiana ou através da combinação de diferentes mecanismos de mutação ($\text{DE}_{cascata}$).

4 EXPERIMENTOS, RESULTADOS E ANÁLISES

Ao decorrer desse capítulo são especificados os algoritmos utilizados para comparação, quais testes estatísticos, proteínas alvo, parâmetros utilizados pelos algoritmos, os resultados e as análises desses resultados. Durante a fase de experimentos foram feitas 10 execuções por caso de testes. Essa quantidade de execuções é encontradas na maioria dos trabalhos da literatura. Dessa forma, obteve-se um total de 580 experimentos por ter utilizado cada configuração do algoritmo, para cada proteína alvo em cada função de energia potencial. Os experimentos foram realizados em computadores Intel Core i7 (3.4GHz) com 16Gb RAM e sistema operacional GNU/Linux na distribuição Ubuntu versão 14.04. As abordagens que fizeram uso das funções de energia CHARMM e AMBER foram desenvolvidos na linguagem de programação C++ em sua versão 11 em arquitetura paralela com 7 *threads* devido à limitação dos computadores utilizados. O pacote de dinâmica molecular que possui essas duas funções de energia é o TINKER. Para a função de energia ROSETTA o sistema foi desenvolvido em Python 3 devido a interface com Python fornecida pelo pacote de dinâmica molecular possuir a função de energia.

As proteínas alvo utilizadas para testar a eficiência das abordagens são a 1PLW (*Met-Enkephalin*), 1ZDD e 1CRN por serem proteínas em comum com trabalhos da literatura como em (VENSKE et al., 2016). Duas outras proteínas (1ENH e 1AIL) são preditas utilizando somente a função de energia ROSETTA, também selecionadas com base em trabalhos da literatura como em (BORGUESAN et al., 2015). O tamanho de cada proteína é apresentado na Tabela 5. A primeira coluna é referente à identificação das proteínas, a segunda referente à quantidade de aminoácidos e a terceira coluna se refere à quantidade de ângulos de torções da cadeia principal e lateral. É importante ressaltar que o tamanho dos indivíduos da população é equivalente à quantidade de ângulos.

Tabela 5 – Proteínas Alvo.

Proteína	Quantidade de Aminoácidos	Quantidade de Ângulos	Estrutura Secundária
1PLW	5	22	-
1ZDD	34	181	α
1CRN	46	191	$\alpha + \beta$
1ENH	54	289	α
1AIL	72	363	α

A fase de experimentação pode ser distribuída em duas etapas: utilizando mecanismos de diversificação e utilizando diferentes mecanismos de mutação de maneira determinista ($\text{DE}_{cascata}$).

Para análises comparativas entre as abordagens aplicadas nessa dissertação, são utilizados dois testes estatísticos não-paramétricos: o de Kruskall-Wallis para verificar se há diferença estatística entre as abordagens, e o teste de Dunn para identificação dos algo-

ritmos que possuem essa diferença. Os testes não paramétricos foram selecionados devido à refutação da hipótese nula que a amostragem dos dados pertencem a uma distribuição normal (SPRENT, 2007).

4.1 EVOLUÇÃO DIFERENCIAL COM MECANISMOS DE DIVERSIFICAÇÃO

Inicialmente são feitos os testes e análises utilizando o algoritmo DE em quatro configurações, a primeira configuração é o $DE_{rand/1/bin}$. A segunda versão utiliza a estratégia *Generation Gap* e é identificada por DE_{GG}. A terceira versão é o algoritmo canônico com a mutação Gaussiana (DE_{GP}). Por fim, a quarta versão utiliza *Generation Gap* e mutação Gaussiana em conjunto e é denominada de DE_{GG-GP}. As quatro configurações utilizam as três funções de energia (CHARMM, AMBER e ROSETTA).

Os parâmetros utilizados de configuração dos algoritmos são listados na Tabela 6. Os parâmetros de tamanho da população, fator de mutação, fator de *crossover* e número de avaliações foram recomendados pelo trabalho de Venske et al. (2016).

Tabela 6 – Configuração de Parâmetros.

Parâmetro	Valor	Descrição
NP	100	Tamanho da População
F	0,5	Fator de Mutação
CR	1,0	Fator de *Crossover*
MAX-AV	500.000	Número de Avaliações
G	0,8	Fator do *Generation Gap*
GSD	0,1	Desvio padrão da mutação Gaussiana

A parâmetrização G do *Generation Gap* foi definida em $0,8$. Esse valor faz com que o mecanismo geracional substitua 80% da população e preserve os outros 20% que são escolhidos aleatoriamente, sendo que o indivíduo com o melhor *fitness* sempre seja preservado durante as gerações. Não é recomendado valores menores do parâmetro G devido às características evolutivas do algoritmo. Caso o mecanismo geracional não substitua muitos indivíduos a evolução pode não ocorrer e o algoritmo não convergir durante o processo de otimização. O fator GSD utilizado pelo mecanismo de mutação Gaussiana pode tornar o mecanismo de mutação aleatório caso esse parâmetro possua valor próximo a 1. Dessa forma, o parâmetro foi definido em $0,1$. Como ambas as estratégias não foram exploradas nesse problema na literatura, os valores foram definidos com base em testes preliminares.

O tempo médio de execução por proteína e função de energia é apresentado na Tabela 7 em horas como unidade de medida. Para essa métrica, são somadas todas as

abordagens por proteína e função de energia e calculada a média aritmética.

Tabela 7 – Tempo médio de Execução.

	1PLW	1ZDD	1CRN
CHARMM	15h36m	15h51m	15h54m
AMBER	15h42m	15h45m	15h51m
ROSETTA	0h14m	0h47m	0h51m

Com os valores apresentados pela Tabela 7 pode-se perceber que a função de energia ROSETTA possui um custo computacional muito menor do que as funções CHARMM e AMBER do pacote de dinâmica molecular TINKER, tendo diferenças de 15 horas no tempo de execução. Esse fato pode ser relacionado ao excesso de transações de escrita e leitura do TINKER ao disco rígido durante o cálculo do valor de energia.

A Tabela 8 apresenta os resultados obtidos pelas três funções de energias utilizadas neste trabalho. A coluna 4 traz o valor de energia mínima encontrada pela abordagem dentre as 10 execuções enquanto que o RMSD_α desse valor de energia está na coluna 5. A unidade de medida para as energias CHARMM e AMBER são calculadas em kcal mol^{-1} enquanto que a ROSETTA é calculada em *Rosetta Energy Unit* (REU). Dessa forma, a comparação entre as funções de energia não podem ser feitas levando em consideração o valor de energia, somente o RMSD_α. As células em negrito são referentes ao melhor valor absoluto encontrado dentre todas as execuções, mesmo que em alguns casos as abordagens sejam provadas estatisticamente equivalentes. Essa distinção é feita pois as formas 3D preditas apresentadas são referentes a esses valores (menor energia encontrada dentre todas).

Para a proteína 1PLW, os menores valores de RMSD_α encontrados foram de $1,63$Å com a abordagem DE_{GP} para a CHARMM e $1,52$Å com a abordagem $\text{DE}_{rand/1/bin}$ para a AMBER. Dessa forma, observa-se que para a proteína 1PLW, a função de energia AMBER conseguiu encontrar uma conformação melhor em comparação a CHARMM, seja em valores de energia ou pelos valores de RMSD_α. Dentre as funções de energia, a que obteve menor RMSD_α e, consequentemente, a conformação mais próxima foi a ROSETTA com a abordagem DE_{GG} ($0,77$Å), sendo essa a abordagem que obteve a menor energia em REU com $-1,93$.

A Tabela 9 apresenta a comparação estatística de múltiplas alternativas para a proteína 1PLW. Nela são expostos os valores *p-value*, para um resultado menor que $0,05$ representa diferença estatisticamente significativa com uma taxa de 95% de confiança.

Quando analisados os valores retornados pelo teste de Dunn descritos na Tabela 9, é possível observar quais comparações possuem resultados estatisticamente diferentes. Para a função de energia CHARMM, a abordagem DE_{GG-GP} é estatisticamente equivalente às abordagens $\text{DE}_{rand/1/bin}$ e DE_{GP}, tendo desempenho superior somente em relação à

Tabela 8 – Resultados Obtidos CHARMM, AMBER e ROSETTA.

Proteína	Função de Energia	Versão	Energia Mínima	$RMSD_\alpha$	Energia Média
1PLW	CHARMM	$DE_{rand/1/bin}$	$-34,69$	$1,90$Å	$-28,58 \pm 3,00$
		DE_{GG}	$-33,95$	$1,99$Å	$-26,35 \pm 2,77$
		DE_{GP}	$-32,10$	**1,63Å**	$-27,96 \pm 1,91$
		DE_{GG-GP}	**-35,82**	$1,98$Å	**-30,47 ± 4,44**
	AMBER	$DE_{rand/1/bin}$	$-95,97$	**1,52Å**	$-87,83 \pm 4,70$
		DE_{GG}	$-100,33$	$1,68$Å	$-85,78 \pm 14,38$
		DE_{GP}	$-107,81$	$1,62$Å	**-107,36 ± 1,12**
		DE_{GG-GP}	**-109,89**	$1,77$Å	$-106,22 \pm 1,90$
	ROSETTA	$DE_{rand/1/bin}$	$-1,30$	$1,72$Å	$-0,37 \pm 0,78$
		DE_{GG}	**-1,93**	**0,77Å**	$-0,48 \pm 1,29$
		DE_{GP}	$-1,24$	$1,76$Å	$-0,49 \pm 0,62$
		DE_{GG-GP}	$-1,41$	$1,73$Å	**-0,51 ± 1,02**
1ZDD	CHARMM	$DE_{rand/1/bin}$	$-955,68$	$2,65$Å	$-508,52 \pm 262,99$
		DE_{GG}	$-796,68$	$4,25$Å	$95,03 \pm 1.503,92$
		DE_{GP}	**-1.216,40**	**2,36Å**	**-1.086,99 ± 105,74**
		DE_{GG-GP}	$-1.156,95$	$5,69$Å	$-983,97 \pm 119,80$
	AMBER	$DE_{rand/1/bin}$	$343,49$	$6,28$Å	$24.440,87 \pm 17.839,12$
		DE_{GG}	$1.901,21$	$8,05$Å	$21.081,19 \pm 15.881,01$
		DE_{GP}	**-586,26**	**2,86Å**	**-343,08 ± 208,39**
		DE_{GG-GP}	$-387,43$	$9,33$Å	$12,68 \pm 369,86$
	ROSETTA	$DE_{rand/1/bin}$	$114,51$	$5,85$Å	$137,70 \pm 10,43$
		DE_{GG}	$126,93$	**3,84Å**	$154,86 \pm 19,93$
		DE_{GP}	**66,93**	$8,55$Å	**89,61 ± 12,91**
		DE_{GG-GP}	$76,72$	$7,92$Å	$118,05 \pm 34,44$
1CRN	CHARMM	$DE_{rand/1/bin}$	$818,04$	**7,89Å**	$2.483,87 \pm 3.424,94$
		DE_{GG}	$594,07$	$13,12$Å	$1.608,72 \pm 1.152,38$
		DE_{GP}	**166,83**	$10,71$Å	**288,52 ± 86,67**
		DE_{GG-GP}	$260,12$	$8,60$Å	$464,60 \pm 133,59$
	AMBER	$DE_{rand/1/bin}$	$716,87$	$9,97$Å	$6.956,85 \pm 7.176,66$
		DE_{GG}	$6.310,39$	$14,06$Å	$12.584,24 \pm 5.640,21$
		DE_{GP}	**212,62**	**7,61Å**	**457,39 ± 175,71**
		DE_{GG-GP}	$244,03$	$9,08$Å	$607,14 \pm 207,07$
	ROSETTA	$DE_{rand/1/bin}$	$131,73$	$15,21$Å	$181,29 \pm 40,63$
		DE_{GG}	$157,70$	$20,15$Å	$185,73 \pm 16,62$
		DE_{GP}	$86,72$	**13,80Å**	**94,68 ± 7,50**
		DE_{GG-GP}	**86,25**	$20,60$Å	$106,70 \pm 15,96$

abordagem DE_{GG}.

Para a função de energia AMBER, nota-se que são estatisticamente equivalentes as abordagens que utilizam da perturbação Gaussiana como mecanismo de diversificação (DE_{GP} e DE_{GG-GP}), sendo essas as abordagens que conseguiram melhores valores de energia para a proteína 1PLW. Dessa forma, pode-se notar que a utilização do mecanismo de perturbação Gaussiana teve influência na obtenção de melhores valores de energia, demonstrando significância estatística em comparação às abordagens $DE_{rand/1/bin}$ e DE_{GG} para a função de energia AMBER. Por fim, para a função de energia ROSETTA, o teste de Dunn demonstrou equivalência estatística entre todas as quatro abordagens ($DE_{rand/1/bin}$, DE_{GG}, DE_{GP} e DE_{GG-GP}).

Para a proteína 1ZDD, é possível observar que a abordagem DE_{GG-GP} também consegue valores de energia mínima melhores do que as outras duas abordagens que não fazem uso da perturbação Gaussiana como mecanismo de diversificação. Diferentemente

Tabela 9 – Teste de Dunn para 1PLW

		$DE_{rand/1/bin}$	DE_{GG}	DE_{GP}
CHARMM	DE_{GG}	**0,0084**	–	–
	DE_{GP}	0,4467	**0,0058**	–
	DE_{GG-GP}	0,3369	**0,0025**	0,3871
AMBER	DE_{GG}	0,3798	–	–
	DE_{GP}	**0,0000**	**0,0000**	–
	DE_{GG-GP}	**0,0002**	**0,0005**	0,2221
ROSETTA	DE_{GG}	0,1141	0,2455	–
	DE_{GP}	0,4619	0,2766	–
	DE_{GG-GP}	0,2054	0,4467	0,2337

do ocorrido com a proteína 1PLW, para a 1ZDD a função de energia CHARMM foi a que conseguiu uma conformação mais próxima a real e com valor de energia menor. A função de energia ROSETTA retornou maiores valores de $RMSD_\alpha$ em comparação às outras duas funções.

Na Tabela 10 são mostrados os resultados do teste de Dunn para a proteína 1ZDD. Pode-se observar que é confirmada a superioridade das abordagens DE_{GP} e DE_{GG-GP} em comparação com as abordagen $DE_{rand/1/bin}$ e DE_{GG} pois ambas possuem diferenças estatisticamente significativas em relação às duas versões que não fazem uso da perturbação Gaussiana. Para as funções de energia CHARMM e AMBER as abordagens DE_{GP} e DE_{GG-GP} podem ser consideradas estatisticamente equivalentes devido a valores superiores a $0,05$. Já para a função de energia ROSETTA a abordagem DE_{GP} é estatisticamente diferente de todas as outras abordagens, demonstrando ser a melhor opção nesse caso.

Tabela 10 – Teste de Dunn para 1ZDD.

		$DE_{rand/1/bin}$	DE_{GG}	DE_{GP}
CHARMM	DE_{GG}	0,3725	–	–
	DE_{GP}	**0,0000**	**0,0000**	–
	DE_{GG-GP}	**0,0012**	**0,0004**	0,1647
AMBER	DE_{GG}	0,3231	–	–
	DE_{GP}	**0,0000**	**0,0000**	–
	DE_{GG-GP}	**0,0041**	**0,0019**	0,0903
ROSETTA	DE_{GG}	0,1464	–	–
	DE_{GP}	**0,0002**	**0,0000**	–
	DE_{GG-GP}	**0,0466**	**0,0031**	**0,0346**

Por fim, o mesmo comportamento foi identificado para a proteína 1CRN, ao qual ambas as abordagens que fazem uso da perturbação Gaussiana conseguiram melhores valores de energia mínima em todas as funções de energia avaliadas. Para a proteína 1CRN, observou-se que a função de energia AMBER foi a que conseguiu a melhor conformação 3D devido ao menor valor de $RMSD_\alpha$ obtido.

Quando aplicado o teste de Dunn, o valores apresentados na Tabela 11 monstram que as abordagens DE_{GP} e DE_{GG-GP} são novamente estatisticamente equivalentes. Essas abordagens apresentaram diferenças significativas em comparação às abordagens

$DE_{rand/1/bin}$ e DE_{GG} que não obtiveram resultados competitivos.

Tabela 11 – Teste de Dunn para 1CRN.

		$DE_{rand/1/bin}$	DE_{GG}	DE_{GP}
CHARMM	DE_{GG}	0,4695	–	–
	DE_{GP}	**0,0000**	**0,0000**	–
	DE_{GG-GP}	**0,0008**	**0,0010**	0,0903
AMBER	DE_{GG}	0,1694	–	–
	DE_{GP}	**0,0001**	**0,0000**	–
	DE_{GG-GP}	**0,0018**	**0,0001**	0,2106
ROSETTA	DE_{GG}	0,3440	–	–
	DE_{GP}	**0,0000**	**0,0000**	–
	DE_{GG-GP}	**0,0010**	**0,0002**	0,1554

Com essas informações, pode-se observar que em todos os casos a utilização da perturbação Gaussiana contribuiu de maneira positiva para encontrar melhores resultados em comparação às abordagens que não fazem uso dessa estratégia de diversidade, sendo as abordagens DE_{GP} e DE_{GG-GP} provadas equivalentes quando utilizado o teste de Dunn para verificação. Além disso, em todos os casos para as três proteínas, as abordagens DE_{GP} e DE_{GG-GP} obtiveram um valor de energia média melhor em comparação às outras abordagens, sendo que em sete dos nove casos o DE_{GP} foi o que mostrou melhores resultados médios e com menores valores de desvio padrão.

Perante as análises comparativas entre as três funções de energia utilizando os mesmos algoritmos de busca, cada função de energia obteve a melhor conformação para uma proteína. Para a proteína 1PLW a energia ROSETTA obteve melhor valor de $RMSD_{\alpha}$. A proteína 1ZDD com menor diferença da proteína alvo foi a CHARMM. Por fim, a energia AMBER foi a que obteve melhor conformação para a proteína 1CRN.

Apesar da função de energia ROSETTA não ter sido superior às outras duas funções de energias nos casos de teste, o seu tempo de processamento é um fator que deve ser levado em consideração. A função de energia ROSETTA apresenta uma redução aproximada de 15 horas em comparação às outras duas funções de energia.

De maneira geral, pode-se observar que a utilização da perturbação Gaussiana contribuiu para o processo de otimização na grande maioria dos casos. A única exceção foi para a proteína 1PLW com a função de energia ROSETTA na qual a energia mínima foi encontrada pela abordagem DE_{GG}.

4.1.1 Análise de Convergência e Diversidade

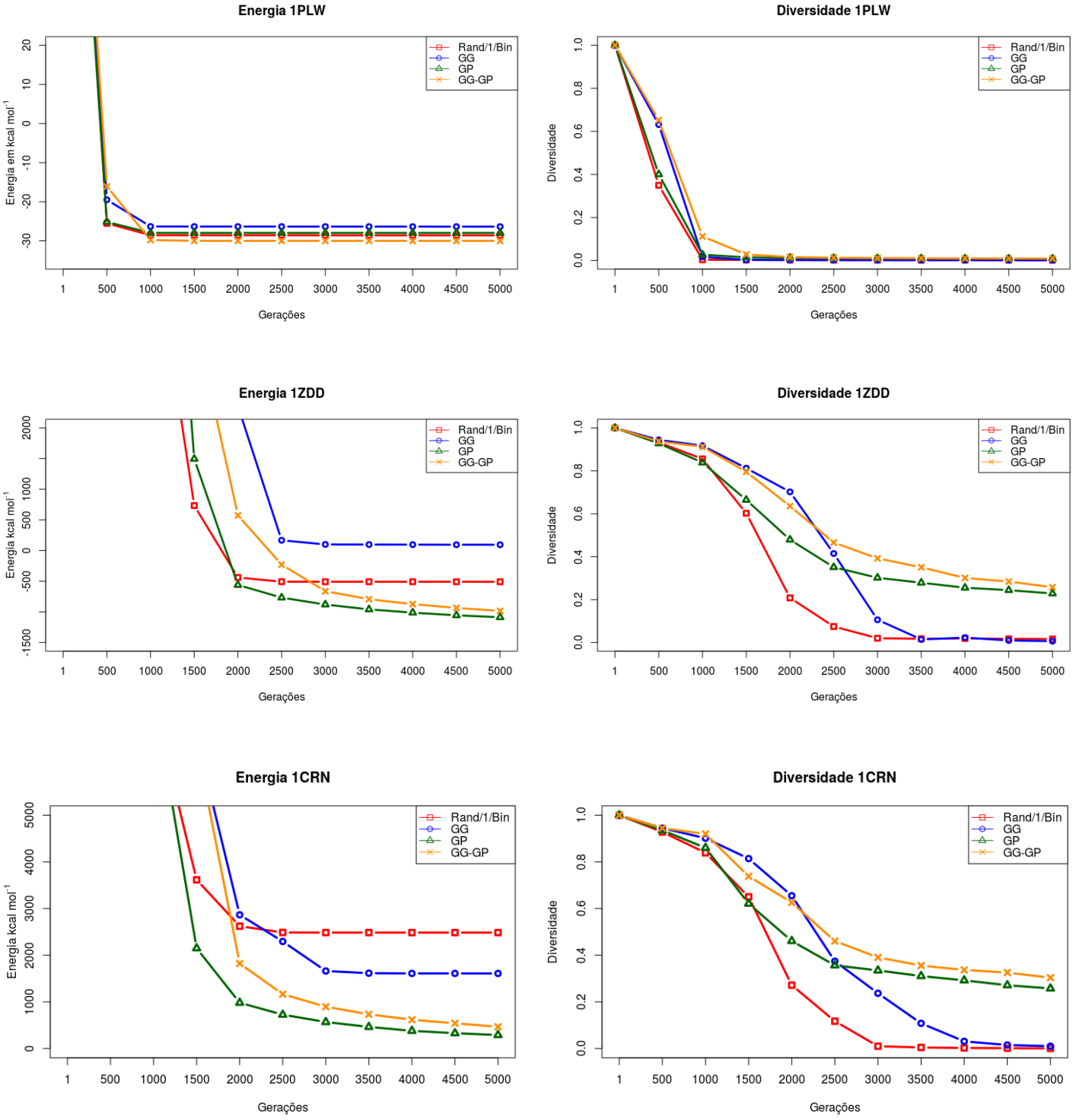
O monitoramento da convergência da energia e da diversidade das abordagens $DE_{rand/1/bin}$, DE_{GG}, DE_{GP} e DE_{GG-GP} é apresentado pela Figura 12 para a função de energia CHARMM, pela Figura 13 para a função de energia AMBER e pela Figura 14 para a função de energia ROSETTA. Na coluna da esquerda dessas figuras estão dispostas

as convergências de energia em cada uma das proteínas e na coluna da direita a diversidade populacional. Para os gráficos de energia o eixo y representa os valores de energia em kcal mol^{-1} para as funções de energia CHARMM e AMBER e para a função ROSETTA os dados estão na unidade de medida conhecida como *Rosetta Energy Unit* (REU). Para os gráficos de diversidade o eixo y representa o valor de diversidade que possui intervalo de $[0, 1]$. Em ambos os gráficos, o eixo x apresenta as gerações, sendo que na geração 5.000 é alcançado o máximo de avaliações configuradas no algoritmo.

Para a proteína 1PLW, onde os resultados dos algoritmos foram similares, pode-se observar que perto da geração 1.000 o valor da função de energia se estabiliza para todas as quatro abordagens. Porém, quando analisada a convergência genotípica da população, é possível verificar que ainda há diversidade populacional após a geração 1.000 para a versão DE_{GG-GP}, monstrando que os mecanismos de diversificação estão conseguindo manter a diversidade por mais gerações. Como o espaço de busca da proteína 1PLW é pequeno em comparação as outras, todas as quatro abordagens obtiveram resultados semelhantes de energia conforme visto previamente e também uma convergência genotípica similar. Porém, a abordagem que obteve melhor resultado foi a abordagem que manteve a diversidade populacional por mais gerações durante o processo de otimização.

Quando analisada a convergência para a proteína 1ZDD, é possível notar maiores diferenças entre os valores de energia e também a diversidade populacional obtida para cada uma das abordagens. A versões $\text{DE}_{rand/1/bin}$ e DE_{GG} estabilizaram seus valores de energia perto da geração 2.500 e 3.000 respectivamente. Entretanto, as versões DE_{GP} e DE_{GG-GP} apresentam um caimento na geração 5.000 em seus valores de energia, que é o limite de gerações que atinge o máximo de avaliações parametrizadas. Esse comportamento, que caimento do valor de energia, pode ser associado com a manutenção da diversidade populacional para essas duas abordagens, mostrando que a manutenção da diversidade é um fator importante.

Figura 12 – Energia (esquerda) e diversidade populacional (direita) para cada proteína - CHARMM.



Nota-se que para as duas versões que obtiveram resultados ruins de energia, $\mathrm{DE}_{rand/1/bin}$ e DE_{GG}, a diversidade populacional acabou rapidamente. Para as abordagens DE_{GP} e DE_{GG-GP}, que atingiram valores melhores de energia, observa-se que a diversidade foi mantida durante todo o processo de otimização, tendo aproximadamente

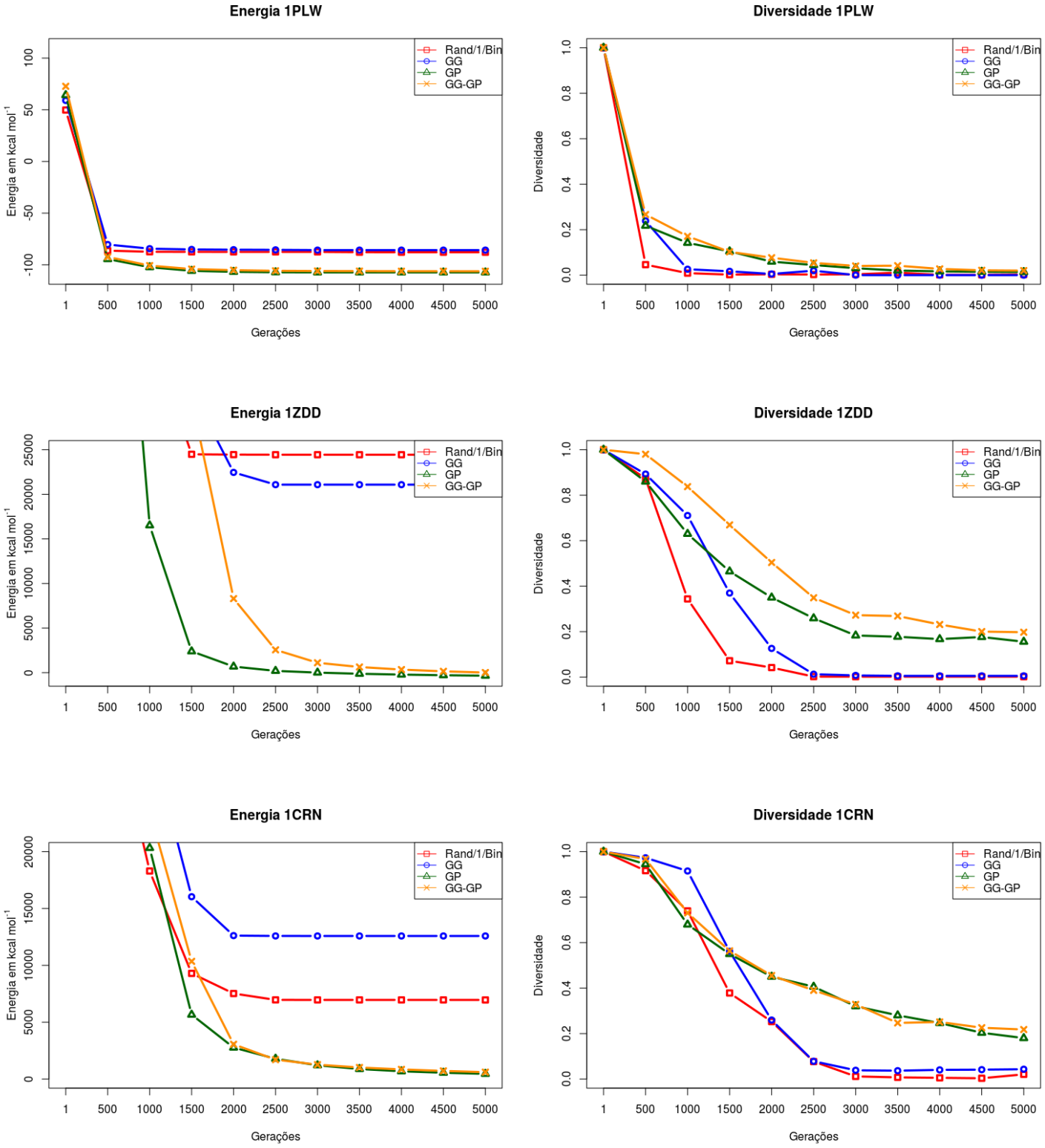
30% de diversidade ao fim da execução do algoritmo, demonstrando que ainda o número de avaliações parametrizadas (500 mil) pode ser extendida ou rotinas de intensificação podem ser aplicadas para tentar obter conformações ainda melhores. A mesma análise de convergência feita para a proteína 1ZDD pode ser observada para a proteína 1CRN, demonstrando que as rotinas que tiveram melhor manutenção da diversidade durante o processo de otimização obtiveram melhores resultados em comparação com as duas piores abordagens ($\mathrm{DE}_{rand/1/bin}$ e DE_{GG}).

Assim, como feito com a energia CHARMM, os gráficos de convergência de energia e de diversidade para a função de energia AMBER estão na Figura 13. O comportamento geral observado utilizando CHARMM também ocorreu com a energia AMBER. As versões que obtiveram melhores resultados de energia (DE_{GP} e DE_{GG-GP}) mantiveram por mais gerações a diversidade populacional da população, evitando a convergência prematura em pontos locais que poderiam prejudicar o resultado. Já os que não conseguiram manter a diversidade por um número maior de iterações ($\mathrm{DE}_{rand/1/bin}$ e DE_{GG}) acabaram obtendo resultados ruins de energia.

A Figura 14 apresenta os gráficos de convergência dos valores de energia (ao lado esquerdo) e diversidade (ao lado direito) para a função de energia ROSETTA. Diferentemente do que ocorreu para as funções de energia CHARMM e AMBER, nenhuma das abordagens conseguiu convergir para a proteína 1PLW. Nota-se que em todos as quatro versões do DE a medida de diversidade genotípica ficou em índices aproximados de 60%. Esse comportamento pode ser relacionado com os limites de angulação preditos pela classificação DSSP-8, que nesse caso ficam entre -180 e 180, além da mudança do espaço de busca proporcionado pela função de energia ROSETTA.

Para as proteínas 1ZDD e 1CRN, pode ser observado que os mecanismos de diversidade conseguiram manter por um número maior de gerações a diversidade populacional quando comparado a abordagem $\mathrm{DE}_{rand/1/bin}$ em todas as três proteínas. Tendo um comportamento diferente do ocorrido com a 1PLW. Entretanto, nota-se que as abordagens DE_{GP} e DE_{GG-GP} mantiveram índices aproximados a 40% de diversidade para as proteínas 1ZDD e 1CRN, não convergindo totalmente apesar dos melhores valores de energia mínima, assim como ocorreu com as funções de energia CHARMM e AMBER.

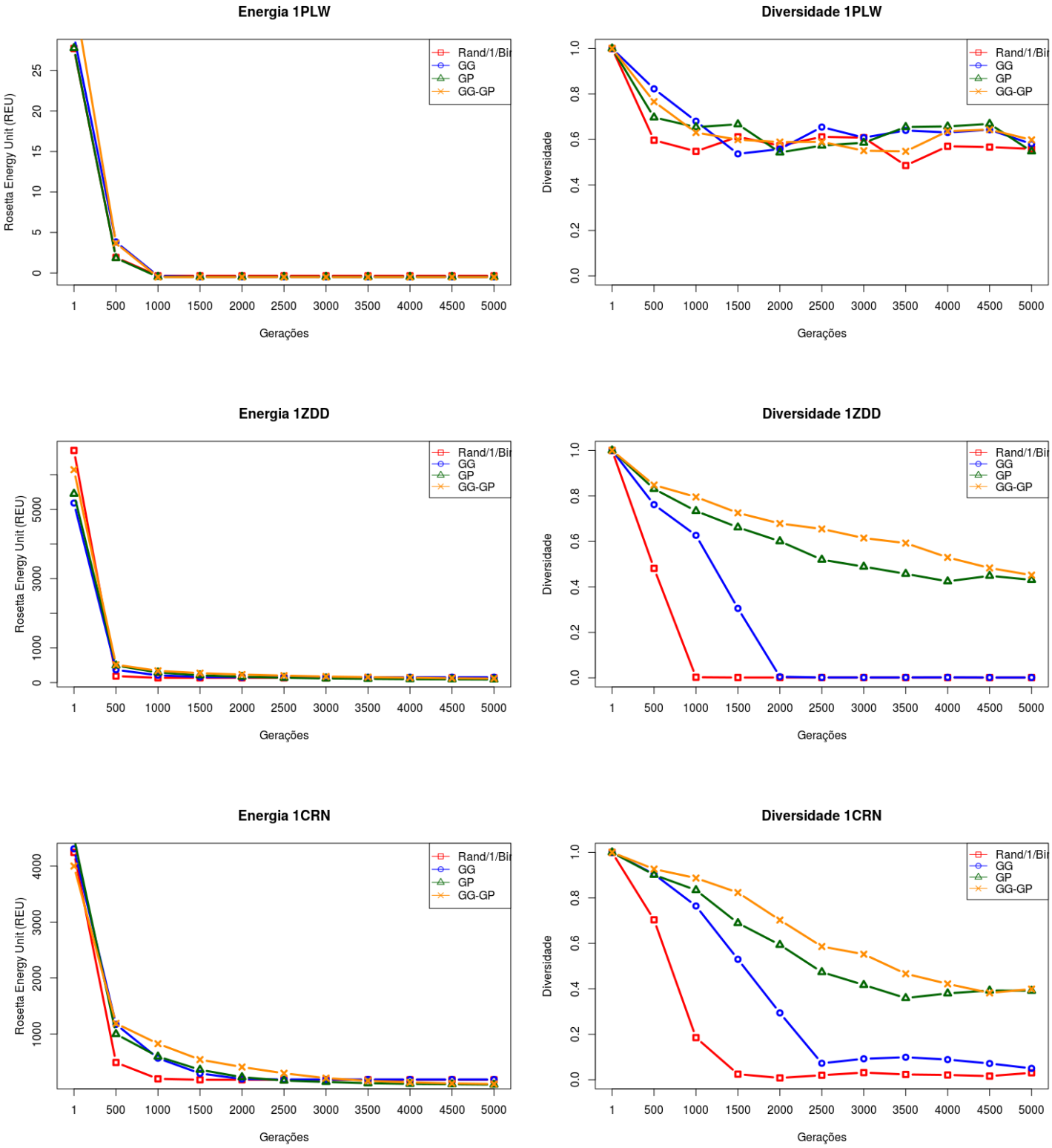
Figura 13 – Energia (esquerda) e diversidade populacional (direita) para cada proteína - AMBER.



Dessa forma, apesar das abordagens que utilizaram da estratégia de mutação Gaussiana terem obtidos melhores resultados na maioria dos casos, elas ainda podem ser melhor

exploradas para que haja a convergência da população, possibilitando a obtenção resultados melhores dos atuais.

Figura 14 – Energia (esquerda) e diversidade populacional (direita) para cada proteína - ROSETTA.



Outra análise relevante observada, independente da função de energia para as proteínas 1ZDD e 1CRN, é que o mecanismo de diversificação *Generation Gap* não conseguiu obter muita diversidade durante o processo de otimização quando comparado com o mecanismo de mutação Gaussiana. Porém, quando as duas técnicas são combinadas a diversidade populacional é mantida em um percentual maior do que a utilização de uma ou outra de maneira isolada. Com os gráficos de convergência de energia e de diversidade populacional para as funções de energia, pode-se verificar a eficiência dos mecanismos de diversificação e também a relação entre os valores de energia obtidos com os níveis de diversidade durante o processo de otimização. Neste quesito, para as funções de energia, os comportamentos observados foram similares.

4.1.2 Comparação com Outras Abordagens

Os resultados obtidos pelas abordagens implementadas no presente trabalho são comparados com trabalhos da literatura que utilizaram a representação da proteína em ângulos de torções da cadeia principal e da cadeia secundária e que utilizaram as mesmas proteínas (1PLW, 1ZDD e 1CRN). Como não foram encontrados na literatura trabalhos que utilizam a função de energia AMBER para essas proteínas, somente as funções de energia CHARMM e ROSETTA são comparadas com a literatura. A Tabela 12 apresenta os valores obtidos pelas abordagens $\text{DE}_{rand/1/bin}$, DE_{GG}, DE_{GP} e DE_{GG-GP} com trabalhos da literatura que foram revisados na Seção 2.5 para a função de energia CHARMM. O símbolo '−' é utilizado quando não foram encontrados os dados no trabalho de referência. Os melhores resultados absolutos estão destacados em negrito.

Para a proteína *Met-Enkephalin* (1PLW) pode-se observar que a versão DE_{GG-GP} obteve o resultado com menor energia mínima em $-35,82$ kcal mol^{-1} e com RMSD_α de $1,98$Å em comparação com os demais trabalhos da literatura. Os valores obtidos por todas as abordagens aplicadas nessa dissertação são competitivos com os encontrados na literatura e com o algoritmo tido como estado-da-arte conhecido como ADEMO/D (VENSKE et al., 2016). O menor RMSD_α encontrado foi pelo algoritmo NSGA-II no trabalho (ROMERO, 2010) com $1,26$Å. Os resultados da proteína 1ZDD demonstram que o valor mais baixo de energia foi obtida pela versão DE_{GP} com $-1.216,40$ kcal mol^{-1} e RMSD_α de $2,36$Å. Em comparação com a literatura, a versão DE_{GP} obteve valores inferiores mas competitivos com o algoritmo ADEMO/D (VENSKE et al., 2016). Nessa proteína o ADEMO/D obteve a menor energia com $-1.301,38$ kcal mol^{-1} e RMSD_α de 2,14Å.

Para a proteína 1CRN a abordagem que obteve melhor resultado foi a DE_{GP} com valor de energia mínimo de $166,83$ kcal mol^{-1} e RMSD_α de $10,71$Å. Apesar do valor de energia obtido pela versão DE_{GP} ter sido melhor que os valores da literatura, o RMSD_α

Tabela 12 – Resultados obtidos CHARMM.

Proteína	Versão	Energia Mínima	$RMSD_\alpha$	Energia Média
1PLW	$DE_{rand/1/bin}$	$-34,69$	$1,90$Å	$-28,58 \pm 3,00$
	DE_{GG}	$-33,95$	$1,99$Å	$-26,35 \pm 2,77$
	DE_{GP}	$-32,10$	$1,63$Å	$-27,96 \pm 1,91$
	DE_{GG-GP}	**-35,82**	$1,98$Å	$-30,47 \pm 4,44$
	ADEMO/D	$-30,43$	$1,77$Å	–
	BFOA	$-19,10$	$3,60$Å	–
	I-PAES	$-20,56$	$2,83$Å	–
	NSGA-II	$-22,73$	**1,26Å**	–
	SSORIGA	$42,82$	–	$46,23 \pm 1,64$
1ZDD	$DE_{rand/1/bin}$	$-955,68$	$2,65$Å	$-508,52 \pm 262,99$
	DE_{GG}	$-796,68$	$4,25$Å	$95,03 \pm 1.503,92$
	DE_{GP}	$-1.216,40$	$2,36$Å	$-1.086,99 \pm 105,74$
	DE_{GG-GP}	$-1.156,95$	$5,69$Å	$-983,97 \pm 119,80$
	ADEMO/D	**-1.301,38**	**2,14Å**	–
	I-PAES	$-1.052,09$	$2,27$Å	–
	NSGA-II	$-1.218,57$	$3,81$Å	–
1CRN	$DE_{rand/1/bin}$	$818,04$	$7,89$Å	$2.483,87 \pm 3.424,94$
	DE_{GG}	$594,07$	$13,12$Å	$1.608,72 \pm 1.152,38$
	DE_{GP}	**166,83**	$10,71$Å	$288,52 \pm 86,67$
	DE_{GG-GP}	$260,12$	$8,60$Å	$464,60 \pm 133,59$
	ADEMO/D	$253,25$	$6,06$Å	–
	I-PAES	$509,09$	**4,43 Å**	–
	NSGA-II	$262,68$	$7,32$Å	–
	SSORIGA	$503,56$	–	$535,09 \pm 20,98$

ficou bastante alto. O melhor $RMSD_\alpha$ para essa proteína pertence ao algoritmo I-PAES (CUTELLO; NARZISI; NICOSIA, 2008) com $4,43$Å. Com essa comparação foi possível perceber que apesar das abordagens utilizadas nesse trabalho serem mais simples por não serem multi-objetivo como a maioria dos algoritmos, com exceção do BFOA (PAL, 2014), foram obtidos resultados competitivos com os da literatura, conseguindo superar até o algoritmo considerado estado-da-arte ADEMO/D em duas das três proteínas.

A Tabela 13 apresenta os resultados obtidos com a função de energia ROSETTA em comparação com dois algoritmos encontrados na literatura: GA e PSO. A primeira coluna lista as proteínas alvo enquanto que a segunda coluna identifica os algoritmos. Na coluna 3 estão os melhores valores de energia encontrados e na quarta coluna o $RMSD_\alpha$ condizente com a energia. A última coluna representa a energia média das 10 execuções e também o desvio padrão. O símbolo '–' é utilizado quando não foram encontrados os dados no trabalho de referência. Os melhores resultados absolutos estão destacados em negrito. Para a proteína 1PLW não foram encontrados resultados na literatura, mantendo dessa forma a abordagem DE_{GG} como melhor resultado até o momento.

Para a proteína 1ZDD observa-se que o GA teve o pior desempenho em relação a energia e ganha somente do PSO quando comparados os valores de $RMSD_\alpha$. Dentre os seis algoritmos, a versão DE_{GP} foi a que obteve melhor valor de energia com $66,93$ REU. Em questões de $RMSD_\alpha$ a abordagem DE_{GG} teve o melhor resultado. Para a proteína 1CRN os valores de referência da literatura são piores que a versão canônica do $DE_{rand/1/bin}$. Dentre todas as abordagens, o melhor resultado de energia mínima foi obtido pela versão

Tabela 13 – Resultados obtidos ROSETTA.

Proteína	Versão	Energia Mínima	$RMSD_\alpha$	Energia Média
1PLW	$DE_{rand/1/bin}$	$-1,30$	$1,72$Å	$-0,37 \pm 0,78$
	DE_{GG}	**-1,93**	**0,77Å**	$-0,48 \pm 1,29$
	DE_{GP}	$-1,24$	$1,76$Å	$-0,49 \pm 0,62$
	DE_{GG-GP}	$-1,41$	$1,73$Å	$-0,51 \pm 1,02$
	GA	–	–	–
	PSO	–	–	–
1ZDD	$DE_{rand/1/bin}$	$114,51$	$5,85$Å	$137,70 \pm 10,43$
	DE_{GG}	$126,93$	**3,84Å**	$154,86 \pm 19,93$
	DE_{GP}	**66,93**	$8,55$Å	$89,61 \pm 12,91$
	DE_{GG-GP}	$76,72$	$7,92$Å	$118,05 \pm 34,44$
	GA	230.60	9.50Å	274.50 ± 33.60
	PSO	$95,80$	$11,00$Å	$239,30 \pm 40,00$
1CRN	$DE_{rand/1/bin}$	$131,73$	$15,21$Å	$181,29 \pm 40,63$
	DE_{GG}	$157,70$	$20,15$Å	$185,73 \pm 16,62$
	DE_{GP}	$86,72$	**13,80Å**	$94,68 \pm 7,50$
	$DE_{GG\text{-}GP}$	**86,25**	$20,60$Å	$106,70 \pm 15,96$
	GA	$467,70$	$19,30$Å	$676,0 \pm 121,30$
	PSO	$230,09$	$15,90$Å	$478,30 \pm 102,40$

DE_{GG-GP} com $86,25$ REU.

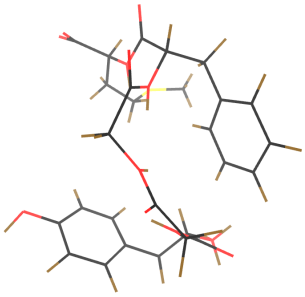
Por fim, todas as abordagens do DE obtiveram melhores valores de energia em comparação com a literatura, tendo destaque para a abordagem DE_{GG-GP} que conseguiu o menor valor de energia e para a abordagem DE_{GP} com o menor valor de $RMSD_\alpha$.

A partir dessas comparações, pode-se observar que em todas as ocasiões o algoritmo DE, mesmo em sua forma canônica, conseguiu valores de energia superiores aos encontrados na literatura que utilizaram a função de energia ROSETTA. Dessa forma, pode-se concluir que para as funções de energia CHARMM e ROSETTA, as abordagens utilizadas nessa dissertação conseguiram resultados competitivos com a literatura com abordagens simples mas que conseguem manter a diversidade populacional durante o período de otimização.

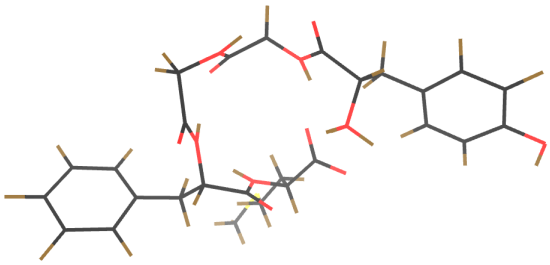
4.1.3 Representação Gráfica das Conformações

As representações gráficas das proteínas são criadas pelo *software Visual Molecular Dynamics* (HUMPHREY; DALKE; SCHULTEN, 1996). As diferenças entre as conformações nativas e preditas das proteínas utilizando a função de energia CHARMM estão na Figura 15, sendo as do lado esquerdo as nativas e do lado direto as preditas. As representações foram escolhidas conforme o menor valor de energia encontrado dentre as abordagens por proteína.

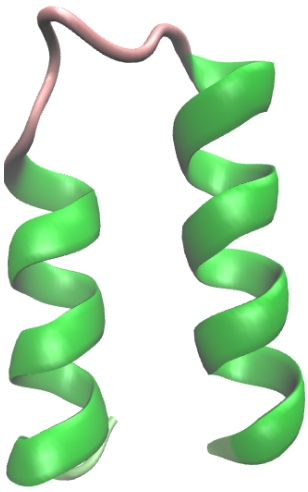
Figura 15 – Representações gráficas das proteínas - CHARMM.



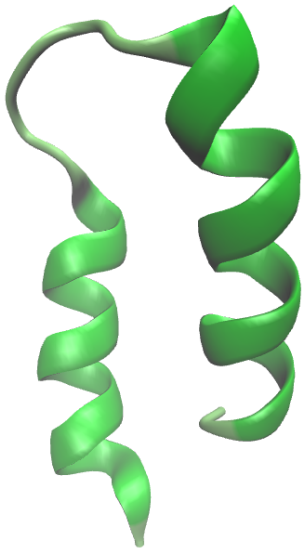
(a) – 1PLW - Nativa.



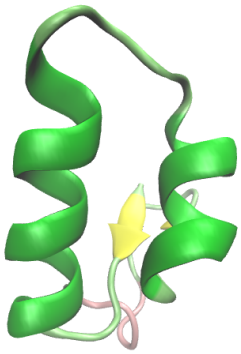
(b) – 1PLW - Predita (1.98Å).



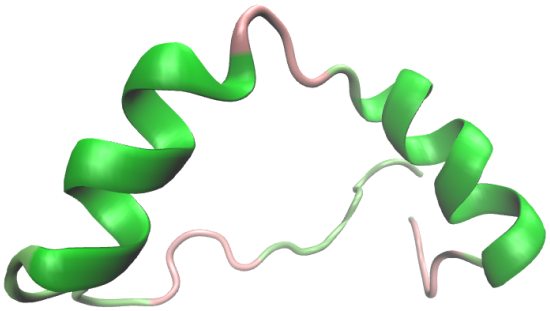
(c) – 1ZDD - Nativa.



(d) – 1ZDD - Predita (2, 36Å).



(e) – 1CRN - Nativa.

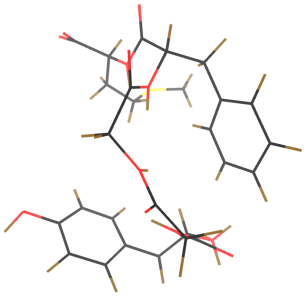


(f) – 1CRN - Predita (10, 71Å).

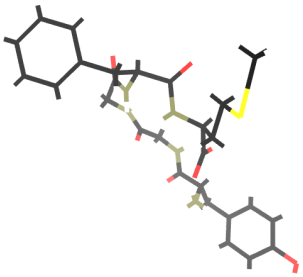
Nas proteínas 1PLW e 1ZDD é possível verificar que as principais estruturas foram preditas, tendo pequenos erros de alinhamento das estruturas. Para a proteína 1CRN é possível verificar que as duas hélices α são preditas mas possuem pequenos erros de alinhamento também. Entretanto, a conformação nativa dessa proteína possui duas folhas β que não foram preditas utilizando a função de energia CHARMM. A Figura 16 traz

a comparação das conformações nativas (a esquerda) com as conformações preditas pela função de energia AMBER (a direita). As representações foram escolhidas conforme o menor valor de energia encontrado dentre as abordagens por proteína.

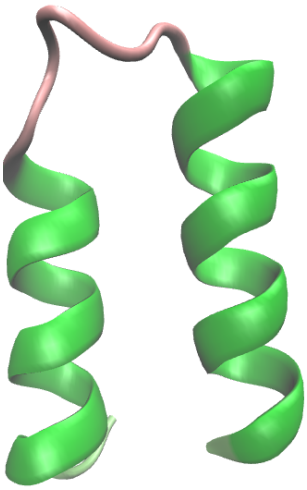
Figura 16 – Representações gráficas das proteínas - AMBER.



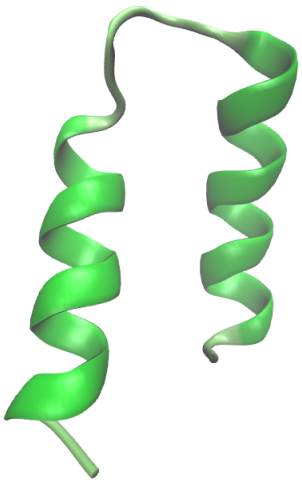
(a) – 1PLW - Nativa.



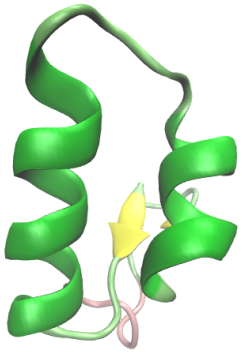
(b) – 1PLW - Predita (1, 77Å).



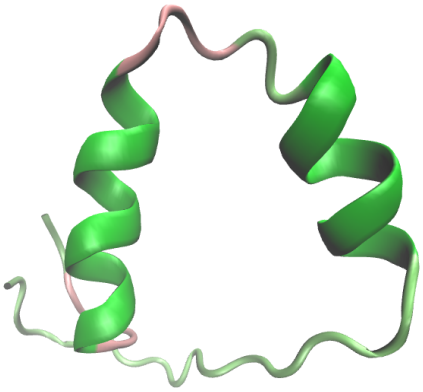
(c) – 1ZDD - Nativa.



(d) – 1ZDD - Predita (2, 86Å).



(e) – 1CRN - Nativa.

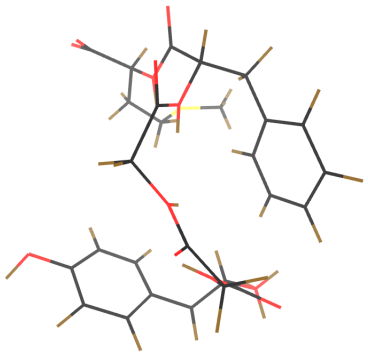


(f) – 1CRN - Predita (7, 61Å).

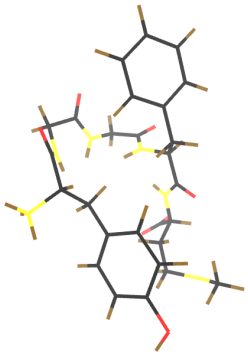
Assim como ocorreu com a CHARMM, a utilização da função de energia AMBER permitiu que o algoritmo obtivesse conformações parecidas com as nativas. Porém, uma diferença entre a função de energia CHARMM e AMBER foi no alinhamento da proteína,

demonstrando que a AMBER conseguiu deixar as suas conformações preditas mais similares das nativas do que as predições que utilizaram CHARMM. Para a proteína 1CRN novamente as folhas β não foram formadas. A Figura 17 apresenta as comparação gráfica entre as conformações nativas (esquerda) e preditas (direita) utilizando a função de energia ROSETTA. Essas conformações também são referentes aos menores valores de energia absolutos encontrados em cada proteína.

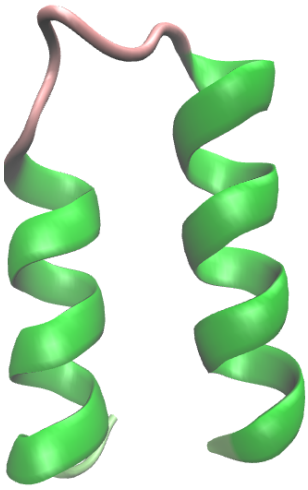
Figura 17 – Representações gráficas das proteínas - ROSETTA.



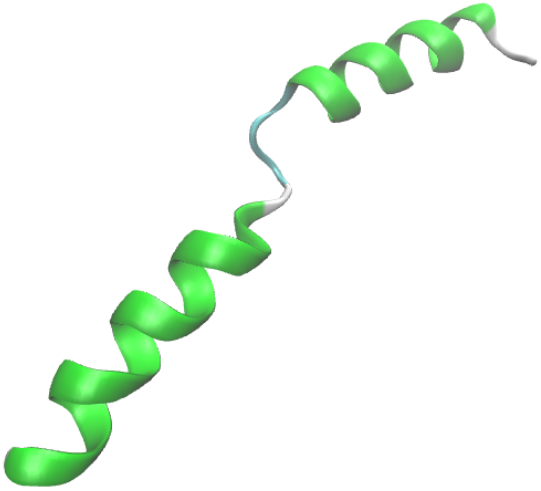
(a) – 1PLW - Nativa.



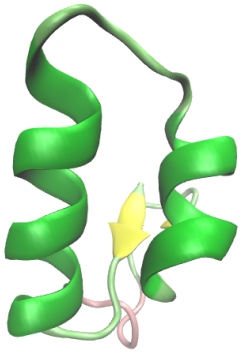
(b) – 1PLW - Predita (1, 77Å).



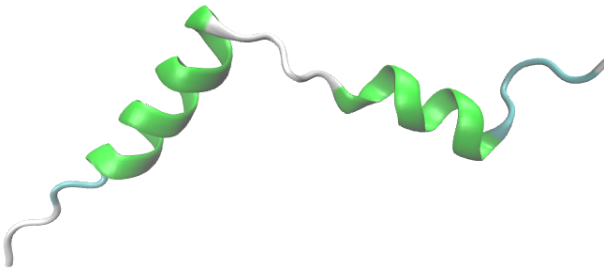
(c) – 1ZDD - Nativa.



(d) – 1ZDD - Predita (8, 55Å).



(e) – 1CRN - Nativa.



(f) – 1CRN - Predita (15, 96Å).

Como as abordagens que utilizaram o ROSETTA não conseguiram bons resultados de RMSD, é esperado que as estruturas preditas possuam certas deformações. Nesse caso, pode-se observar que para a proteína 1ZDD as estruturas de hélice-α são preditas mas com erro de alinhamento entre elas. Para a proteína 1CRN ocorre algo semelhante, tendo as duas estruturas em hélice com erro de alinhamento e com a folha-β não predita.

4.2 EVOLUÇÃO DIFERENCIAL EM CASCATA

Nessa seção a abordagem $\text{DE}_{cascata}$ é aplicada somente para a função de energia ROSETTA devido ao custo computacional relacionado ao tempo de execução do algoritmo ser menor que as funções de energia CHARMM e AMBER. Duas ordenações de cascateamento foram avaliadas, a versão DE_{C1} inicia o primeiro quarto (1.250 gerações) com a estratégia de mutação $\text{DE}_{rand/1/bin}$. Em seguida é aplicada a estratégia $\text{DE}_{curr-to-rand}$ da geração 1.250 até a 2.500. No terceiro quarto aplica-se a versão $\text{DE}_{curr-to-best}$ até a geração 3.750. Por fim o $\text{DE}_{best/1/bin}$ é utilizado como mecanismo de mutação até o fim do processo de otimização. Para a segunda versão (DE_{C2}) são utilizadas somente duas estratégias de mutação, iniciando com o $\text{DE}_{rand/1/bin}$ para o primeiro quarto, no segundo e terceiro quarto é utilizada a versão $\text{DE}_{curr-to-rand}$ e para o último quarto aplica-se novamente o $\text{DE}_{rand-1-bin}$. As ordenações foram definidas com base no comportamento de convergência das abordagens que utilizam somente um mecanismo de mutação, inicializando com abordagens que conseguem manter a diversidade por mais gerações com o intuito do algoritmo convergir a população somente ao fim do processo de otimização.

Para essa etapa de testes, as proteínas utilizadas foram retiradas do trabalho de (BORGUESAN et al., 2015) para fins comparativos. São elas: 1ZDD, 1CRN, 1ENH e 1AIL. Elas foram escolhidas conforme trabalhos da literatura que fazem uso da função de energia ROSETTA e o mesmo modelo de representação (ângulos de torsões da cadeia princial e lateral). A Tabela 14 apresenta o tempo de execução médio entre as duas abordagens em cascata (DE_{C1} e DE_{C2}) para cada uma das proteínas utilizadas nessa etapa de testes.

Tabela 14 – Tempo médio de execução.

Proteína	Tempo
1ZDD	0h48m
1CRN	0h53m
1ENH	2h09m
1AIL	2h52m

É importante ressaltar que essas proteínas foram preditas em (BORGUESAN et al., 2015) com um tempo de execução de 12 horas para cada uma delas.

A Tabela 15 apresenta os resultados obtidos para todas as proteínas utilizadas nesse trabalho com a função de energia ROSETTA. A primeira coluna classifica as proteínas, a segunda o algoritmo e a terceira a energia mínima obtida em 10 execuções. A quarta coluna é o RMSD_α da energia mínima e a quinta e última coluna estão as informações de energia média e desvio padrão. Os valores são comparados com as estratégias de mutação apresentadas no Capítulo 2.3, as abordagens em cascata propostas nesse trabalho (DE_{C1} e DE_{C2}) e por dois algoritmos encontrados na literatura: um GA e um PSO. Para as proteínas 1ENH e 1AIL os autores não fizeram testes com o PSO devido ao melhor resultado obtido pelo AG nas outras proteínas. Ambos os resultados dessas abordagens encontram-se em (BORGUESAN et al., 2015). Os valores em negrito são referentes aos melhores resultados absolutos encontrados em dez execuções.

Tabela 15 – Resultados obtidos ROSETTA.

Proteína	Versão	Energia Mínima	RMSD_α	Energia Média
1ZDD	$\text{DE}_{best/1/Bin}$	334,46	9,10Å	$642,71 \pm 210,77$
	$\text{DE}_{rand/1/Bin}$	114,51	**5,85Å**	$137,70 \pm 10,43$
	$\text{DE}_{curr-to-rand}$	224,55	7,12Å	$356,47 \pm 55,07$
	$\text{DE}_{curr-to-best}$	267,20	8,95Å	$488,70 \pm 200,18$
	DE_{C1}	**54,27**	7,67Å	$82,98 \pm 15,46$
	DE_{C2}	65,77	9,42Å	$\mathbf{82,76 \pm 9,22}$
	GA (BORGUESAN et al., 2015)	230,60	9,50Å	$274,50 \pm 33,60$
	PSO (BORGUESAN et al., 2015)	95.80	11.00Å	$239,30 \pm 40,00$
1CRN	$\text{DE}_{best/1/Bin}$	639,12	13,83Å	$1.104,09 \pm 339,84$
	$\text{DE}_{rand/1/Bin}$	131,73	**15,21Å**	$181,30 \pm 40,63$
	$\text{DE}_{curr-to-rand}$	356,35	19,23Å	$612,14 \pm 136,78$
	$\text{DE}_{curr-to-best}$	361,44	17,26Å	$965,14 \pm 288,51$
	DE_{C1}	82,86	21,56Å	$126,95 \pm 25,98$
	DE_{C2}	**72,48**	15,44Å	$\mathbf{109,01 \pm 22,96}$
	GA (BORGUESAN et al., 2015)	467,70	19,30Å	$676,70 \pm 121,30$
	PSO (BORGUESAN et al., 2015)	230,90	19,30Å	$478,30 \pm 102,40$
1ENH	$\text{DE}_{best/1/Bin}$	1.567,07	18,15Å	$2.340,60 \pm 822,78$
	$\text{DE}_{rand/1/Bin}$	353,81	17,76Å	$451,10 \pm 64,62$
	$\text{DE}_{curr-to-rand}$	918,06	**10,13Å**	$1.390,75 \pm 296,18$
	$\text{DE}_{curr-to-best}$	984,38	21,17Å	$1.725,72 \pm 677,56$
	DE_{C1}	294,25	14,72Å	$372,11 \pm 52,05$
	DE_{C2}	**255,54**	19,28Å	$\mathbf{320,38 \pm 41,06}$
	GA (BORGUESAN et al., 2015)	433,12	20,23Å	$521,05 \pm 46,68$
1AIL	$\text{DE}_{best/1/Bin}$	1.288,18	29,60Å	$2.529,53 \pm 908,70$
	$\text{DE}_{rand/1/Bin}$	441,27	26,09Å	$546,69 \pm 71,09$
	$\text{DE}_{curr-to-rand}$	764,54	21,13Å	$1.467,19 \pm 357,72$
	$\text{DE}_{curr-to-best}$	1.338,29	23,75Å	$1.819,81 \pm 401,25$
	DE_{C1}	357,84	25,00Å	$440,63 \pm 58,11$
	DE_{C2}	**332,54**	**16,88Å**	$\mathbf{411,81 \pm 56,84}$
	GA (BORGUESAN et al., 2015)	460,27	24,65Å	$553,72 \pm 65,68$

Com os resultados da proteína 1ZDD, é possível notar que, tanto em comparação com as versões que utilizaram somente um mecanismo de mutação durante a busca, quanto ao GA e PSO encontrados na literatura, ambas as abordagens em cascata (DE_{C1} e DE_{C2}) propostas nesse trabalho obtiveram melhores resultados de energia mínima, energia média e desvio padrão. Apesar da diferença entre a energia mínima dessas duas abordagens, a média ficou em valores equivalentes. Porém, o algoritmo DE_{C2} demonstrou ter uma menor

variação em seus resultados. É possível observar também que entre as abordagens que utilizam de uma única estratégia de mutação, as versões que não fazem uso do melhor indivíduo da população ($DE_{rand/1/bin}$ e $DE_{curr-to-rand}$) conseguem melhores valores em relação aos que forçam a população a se aproximarem do melhor indivíduo.

Para as demais proteínas (1CRN, 1ENH e 1AIL) ocorre o mesmo comportamento dos resultados, tendo destaque para as abordagens que fazem uso do método em cascata do mecanismo de mutação que sempre conseguem atingir melhores valores em comparação com as abordagens que fazem uso de somente um mecanismo de mutação. Mesmo em comparação com a literatura, os valores de energia mínima e média demonstram ser competitivos com os AG e PSO (BORGUESAN et al., 2015). Os resultados de $RMSD_\alpha$ variam bastante entre as abordagens, nem sempre tendo o melhor $RMSD_\alpha$ para a melhor energia mínima. Esse acontecimento pode ser relacionado às deficiências das funções de energia.

Na Tabela 16 é possível correlacionar quais as abordagens possuem significância estatística com 95% de intervalo de confiança para todas as quatro proteínas (1ZDD, 1CRN, 1ENH e 1AIL). Em cada célula da tabela estão os *p-values* retornados pelo teste de Dunn. Caso esses valores sejam inferiores à $0,05$ pode-se assumir que há diferenças entre as abordagens. Não foi possível aplicar o teste estatístico em abordagens da literatura devido a não publicação de todos os resultados da amostragem.

Tabela 16 – Teste de Dunn para o $DE_{cascata}$

		$DE_{best/1/bin}$	$DE_{rand/1/bin}$	$DE_{curr-to-rand}$	$DE_{curr-to-best}$	DE_{C1}
1ZDD	$DE_{rand/1/bin}$	**0,0004**	–	–	–	–
	$DE_{curr-to-rand}$	0,0506	**0,0408**	–	–	–
	$DE_{curr-to-best}$	0,2063	**0,0052**	0,2063	–	–
	DE_{C1}	**0,0000**	**0,0290**	**0,0001**	**0,0000**	–
	DE_{C2}	**0,0000**	**0,02580**	**0,0001**	**0,0000**	0,4796
1CRN	$DE_{rand/1/bin}$	**0,0003**	–	–	–	–
	$DE_{curr-to-rand}$	**0,0345**	**0,00562**	–	–	–
	$DE_{curr-to-best}$	0,3552	**0,0012**	0,0740	–	–
	DE_{C1}	**0,0000**	0,0758	**0,0013**	**0,0000**	–
	DE_{C2}	**0,0000**	**0,0196**	**0,0001**	**0,0000**	0,2652
1ENH	$DE_{rand/1/bin}$	**0,0001**	–	–	–	–
	$DE_{curr-to-rand}$	**0,0493**	**0,0229**	–	–	–
	$DE_{curr-to-best}$	0,1685	**0,0036**	0,2447	–	–
	DE_{C1}	**0,0000**	0,1220	**0,0008**	**0,0001**	–
	DE_{C2}	**0,0000**	**0,0215**	**0,0000**	**0,0000**	0,1955
1AIL	$DE_{rand/1/bin}$	**0,0002**	–	–	–	–
	$DE_{curr-to-rand}$	**0,0467**	**0,0266**	–	–	–
	$DE_{curr-to-best}$	0,2174	**0,0023**	0,1851	–	–
	DE_{C1}	**0,0000**	**0,0853**	**0,0005**	**0,0000**	–
	DE_{C2}	**0,0000**	**0,0376**	**0,0001**	**0,0000**	0,3410

Dessa forma, pode-se verificar que a abordagem DE_{C1} possui diferenças estatisticamente relevantes quando comparadas às abordagens $DE_{best/1/bin}$, $DE_{curr-to-rand}$ e $DE_{curr-to-best}$. Apesar do DE_{C1} ter obtido resultados de energia mínima melhores que o $DE_{rand/1/bin}$, o teste de Dunn aponta que com 95% de confiança, que as abordagens

DE_{C1} e $DE_{rand/1/bin}$ são equivalentes para as proteínas 1CRN, 1ENH e 1AIL. Analisando a abordagem DE_{C2} é possível observar que ela possui diferenças com todas as abordagens que utilizaram de apenas um mecanismo de mutação para todas as quatro proteínas. A abordagem DE_{C2} foi considerada somente equivalente à DE_{C1}.

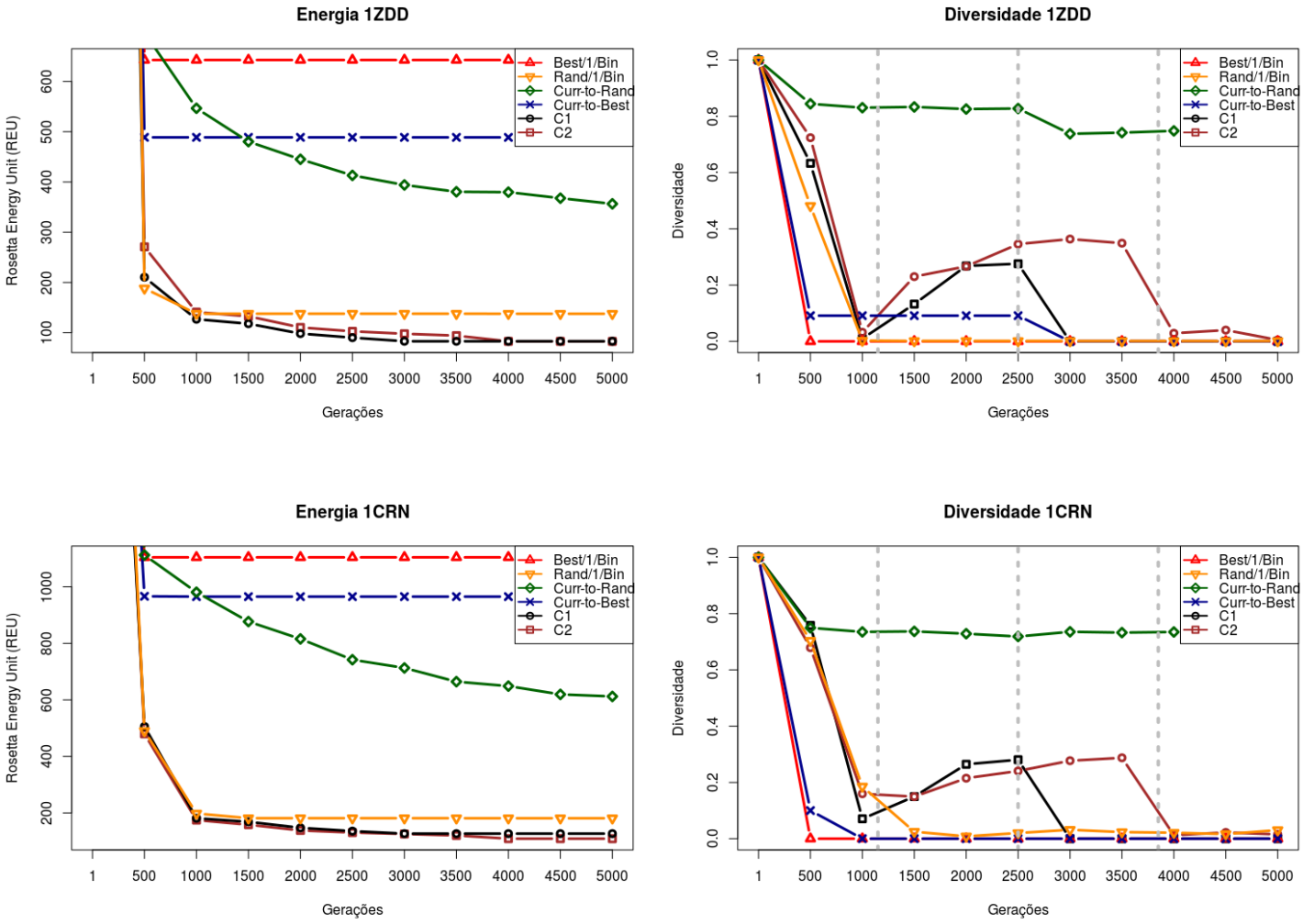
Quando comparados os resultados obtidos com os valores retornados pelo teste de Dunn expostos na Tabela 16, é possível observar que a utilização da abordagem em cascata para os mecanismos de mutação contribuiu para a obtenção de melhores resultados em comparação aos algoritmos que utilizaram somente um mecanismo de mutação durante todo o período de otimização. Sendo DE_{C2} estatisticamente relevante em comparação a todas as abordagens que utilizaram somente um mecanismo de mutação, obtendo também, os melhores valores médios de energia em todas as proteínas.

4.2.1 Análise de Convergência e Diversidade

O monitoramento da convergência da energia e da diversidade das abordagens $\mathrm{DE}_{best/1/bin}$, $\mathrm{DE}_{rand/1/bin}$, $\mathrm{DE}_{curr-to-rand}$, $\mathrm{DE}_{curr-to-best}$, DE_{C1} e DE_{C2} é apresentado pela Figura 18 para as proteínas 1ZDD e 1CRN e a Figura 19 para as proteínas 1ENH e 1AIL utilizando a função de energia ROSETTA. Na coluna da esquerda dessas figuras estão dispostas as convergências de energia em cada uma das proteínas e na coluna da direita a diversidade populacional. Para os gráficos de energia o eixo y representa os valores de energia em REU. Para os gráficos de diversidade o eixo y representa o valor de diversidade que possui intervalo de $[0, 1]$. Em ambos os gráficos, o eixo x apresenta as gerações, sendo que na geração 5.000 é obtido o máximo de avaliações configuradas no algoritmo. Para os gráficos de diversidade populacional existem linhas verticais em cinza, separando o número de gerações em quatro partes para a diferenciação do comportamento de cada mecanismo de mutação aplicado em cada quarto no caso das abordagens DE_{C1} e DE_{C2}.

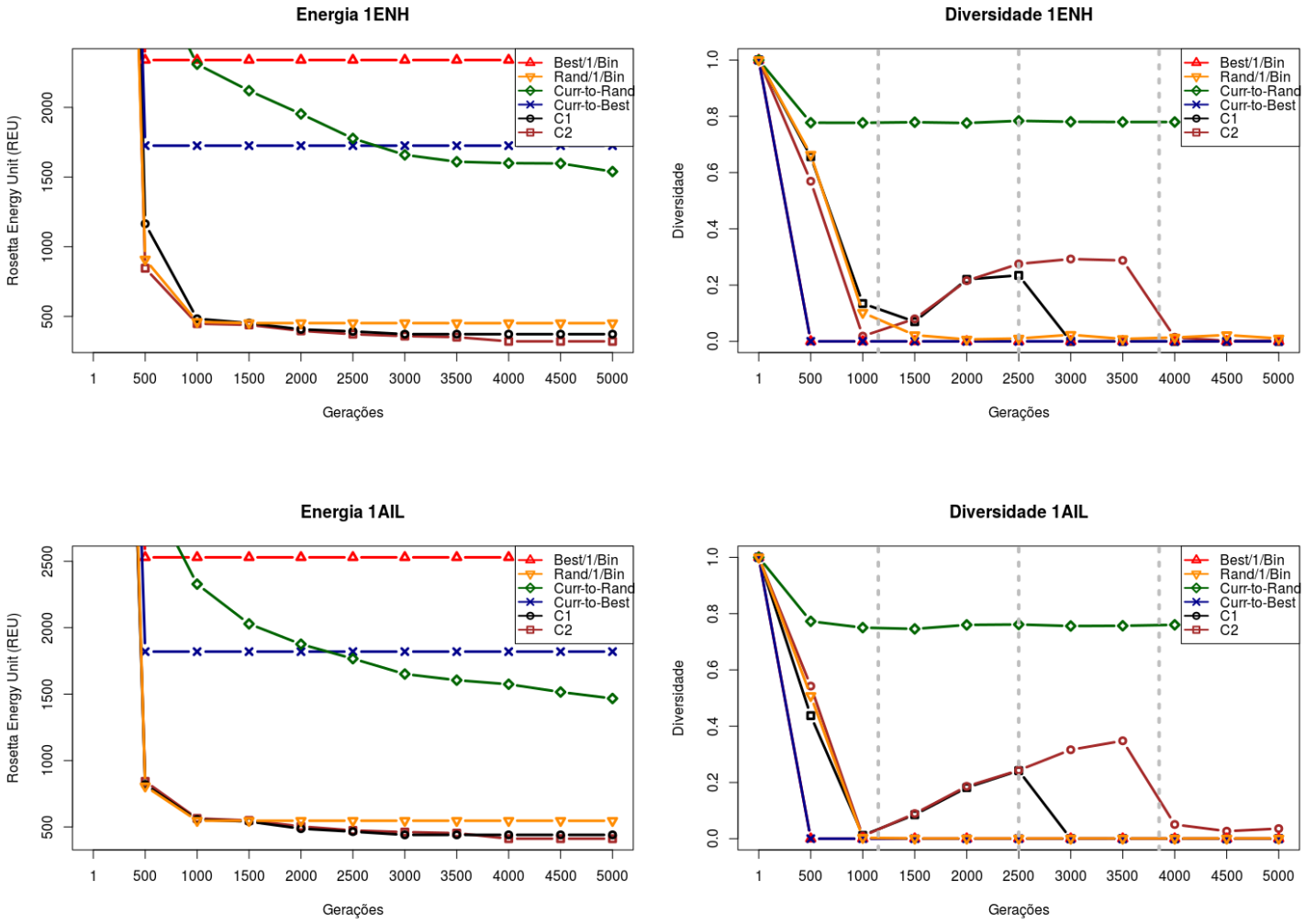
Inicialmente pode-se observar nos gráficos de todas as proteínas que as abordagens que fazem uso dos melhores indivíduos como mecanismo de mutação único ($\mathrm{DE}_{best/1/bin}$ e $\mathrm{DE}_{curr-to-best}$), possuem comportamentos de convergência muito abruptos, formando rapidamente um cotovelo, tanto nos gráficos de energia quanto nos gráficos de diversidade populacional. Esse tipo de comportamento é gerado devido a agressividade do algoritmo em forçar toda a população a se aproximar do melhor indivíduo, aumentando as chances do algoritmo ficar preso em um ponto local e permanecer nele até o fim do processo de otimização, caracterizando a convergência prematura do algoritmo.

Figura 18 – Energia (esquerda) e diversidade populacional (direita) para 1ZDD e 1CRN.



Para as abordagens que possuem características de maior aleatoriedade, $\text{DE}_{rand/1/bin}$ e $\text{DE}_{curr-to-rand}$, nota-se dois comportamentos diferentes. Na abordagem $\text{DE}_{rand/1/bin}$ a convergência e estagnação da energia ocorre mais rapidamente, em torno da geração 1.000. Esse fato está relacionado à perda de diversidade populacional que ocorre mais ou menos no mesmo número de gerações, levando o algoritmo a estagnar. Por sua vez, a abordagem $\text{DE}_{curr-to-rand}$ apresentou uma convergência de diversidade populacional bastante lenta, ficando acima de 75% para todas as proteínas, o que significa que o algoritmo não convergiu totalmente. Em consequência, apesar do valor de energia manter um caimento lento, os valores obtidos foram ruins em comparação à abordagem $\text{DE}_{rand/1/bin}$ devida a falta de intensificação do algoritmo.

Figura 19 – Energia (esquerda) e diversidade populacional (direita) para 1ENH e 1AIL.



Com essas informações, percebe-se que as versões do algoritmo DE que utilizam somente de um mecanismo de mutação possuem dificuldades em manter o balanço entre a diversificação e intensificação durante o processo de busca. Quando observados os comportamentos dos algoritmos DE_{C1} e DE_{C2} é possível observar que a combinação desses mecanismos de mutação durante o período de otimização mantiveram um melhor balanço entre a diversificação e intensificação, justificando a melhora nos resultados obtidos por essas abordagens.

É interessante ressaltar que com a utilização do $\text{DE}_{curr-to-rand}$ foi possível reverter a perda de diversidade que ocorreu ao fim do primeiro quarto, ao qual o mecanismo do algoritmo $\text{DE}_{rand/1/bin}$ estava sendo utilizado. Para a versão DE_{C1} nota-se que, após a recuperação de aproximadamente 30% da diversidade populacional no segundo quarto, a aplicação do $\text{DE}_{curr-to-best}$ rapidamente fez a população convergir e estagnar no terceiro quarto do processo de otimização. Para a abordagem DE_{C2}, que manteve o $\text{DE}_{curr-to-rand}$ em dois quartos do processo de otimização, a diversidade chegou a níveis aproximados de

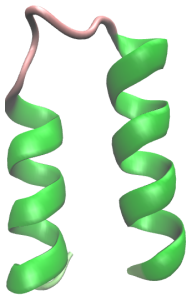
40% e foi explorada somente no quarto final pela abordagem $\text{DE}_{rand/1/bin}$.

Assim como ocorreram com as abordagens testadas na seção anterior, a manutenção da diversidade populacional tem impactos positivos no algoritmo de busca que consegue encontrar melhores soluções. Dessa forma, entende-se que apesar de não ser uma tarefa trivial, o monitoramento e controle da diversidade é um fator a ser considerado na elaboração de metaheurísticas para problemas de otimização.

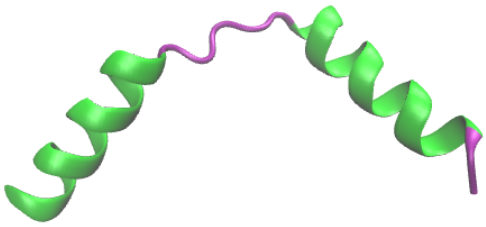
4.2.2 Representação Gráfica das Conformações

As representações gráficas das proteínas são criadas pelo *software Visual Molecular Dynamics* (HUMPHREY; DALKE; SCHULTEN, 1996). As diferenças entre as conformações nativas e preditas das proteínas utilizando a função de energia ROSETTA estão na Figura 20 para as proteínas 1ZDD e 1CRN e Figura 21 para as proteínas 1ENH e 1AIL. Ao lado esquerdo encontram-se as conformações nativas das proteínas e ao lado direito as conformações encontradas pelos melhores valores de energia absolutos encontrados (DE_{C1} para a proteína 1ZDD e DE_{C2} para as proteínas 1CRN, 1ENH e 1AIL). Em parênteses estão identificados os valores de RMSD_{α}.

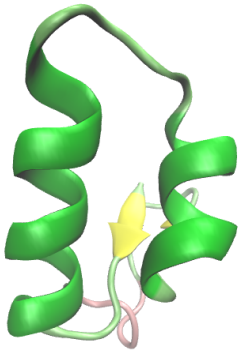
Figura 20 – Representações gráficas das proteínas.



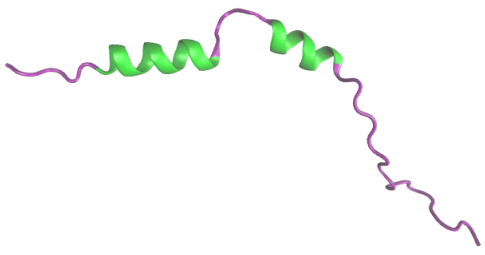
(a) – 1ZDD - Nativa.



(b) – 1ZDD - Predita (7.67Å).

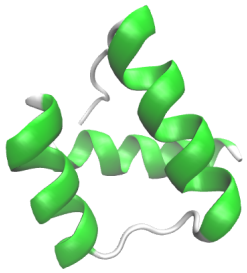


(c) – 1CRN - Nativa.

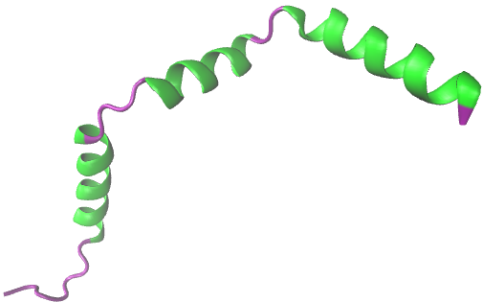


(d) – 1CRN - Predita (15, 44Å).

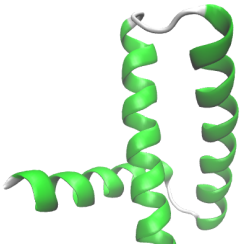
Figura 21 – Representações gráficas das proteínas.



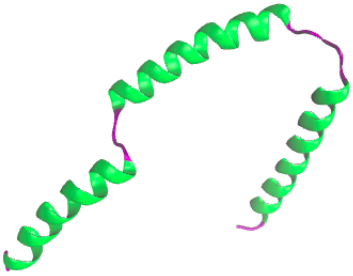
(a) – 1ENH - Nativa.



(b) – 1ENH - Predita (19, 28Å).



(c) – 1AIL - Nativa.



(d) – 1AIL - Predita (16, 88Å).

Apesar dos valores de RMSD serem altos para todas as conformações preditas, é possível identificar que o algoritmo encontrou a maioria das estruturas secundárias para as proteínas 1ZDD, 1CRN, 1ENH e 1AIL, assim como ocorreu com as funções CHARMM e AMBER. O $RMSD_\alpha$ foi penalizado devido ao arranjamento espacial dessas estruturas, algo que não ocorreu com as outras funções de energia. Novamente é possível observar que o algoritmo não foi capaz de identificar a estrutura secundária folha-β da proteína 1CRN.

5 CONSIDERAÇÕES FINAIS

Nessa dissertação foram aplicados duas metologias de busca para a solução do PPEP. Em ambas as metodologias a diversidade populacional foi calculada, validando a hipótese de que o equilíbrio entre as capacidades de intensificação e diversificação das metaheurísticas é um fator importante. A primeira metodologia de busca aplica o algoritmo DE com quatro versões: $\text{DE}_{rand/1/bin}$; DE_{GG} que utiliza a estratégia conhecida como *Generation Gap*; DE_{GP} que utiliza a estratégia conhecida como Perturbação Gaussiana e DE_{GG-GP} que une todas as quatro abordagens.

Todas as quatro versões foram aplicadas em 10 testes para três proteínas alvo: 1PLW com 5 aminoácidos e 22 ângulos; 1ZDD com 34 aminoácidos e 181 ângulos; 1CRN com 46 aminoácidos e 191 ângulos. Os resultados obtidos monstraram que os mecanismos de diversificação conseguiram manter a diversidade por mais gerações durante o processo de otimização com todas as três funções de energia (CHARMM, AMBER e ROSETTA). Porém, a versão DE_{GG} (que utiliza somente do *Generation Gap*) teve resultados com desvios padrões muito altos em todas as três proteínas, demonstrando que quando o algoritmo encontra um ponto local ele não consegue sair e acaba tendo uma convergência prematura. Já quando utilizada a mutação Gaussiana (DE_{GP}), percebe-se que o mecanismo ajuda o algoritmo a manter a diversidade e a não convergir prematuramente, tendo resultados superiores nas proteínas 1ZDD e 1CRN. Além disso, quando combinados os algoritmos na versão DE_{GG-GP} os índices de diversidade se mantém ainda mais altos nas três proteínas, mas tendo melhor resultado somente na proteína 1PLW. Quando comparadas as abordagens DE_{GG-GP} e DE_{GP} com trabalhos da literatura, os resultados demonstram que as versões utilizadas nesse trabalho são competitivas com os da literatura. Um diferencial do presente trabalho está na análise de diversidade populacional que justifica o uso de mecanismos de diversificação, algo que não foi encontrado em nenhum dos trabalhos relacionados. Além disso, as abordagens utilizadas no presente trabalho são mais simples que dos trabalhos da literatura encontrados.

Através das observações feitas pela a aplicação do $\text{DE}_{rand/1/bin}$ e pelos resultados obtidos com a manutenção da diversidade populacional, uma abordagem diferente foi proposta com base no comportamento de diversificação e intensificação de quatro diferentes mecanismos de mutação do DE: $\text{DE}_{rand/1/bin}$, $\text{DE}_{best/1/bin}$, $\text{DE}_{curr-to-rand}$ e $\text{DE}_{curr-to-best}$. Essa abordagem proposta foi chamada de $\text{DE}_{cascata}$ devida a característica de separar o período de otimização em quatro partes e aplicar os mecanismos de mutação. Com base nos testes realizados e análises utilizando o teste de Dunn, pode-se verificar que a abordagem DE_{C2} obteve os melhores resultados de energia em comparação com os algoritmos que utilizaram somente um mecanismo de mutação. Em questão de valores de energia absolutos, as duas versões do $\text{DE}_{cascata}$ obtiveram resultados competitivos quando com-

parados e com dois algoritmos da literatura (GA e PSO). Pode-se relacionar os resultados competitivos devido ao controle da diversidade populacional durante todo o período de otimização, mesmo que de maneira determinista, validando novamente que a manutenção da diversidade durante o período de otimização é um fator importante e que deve ser levado em consideração. Esse método foi aplicado em quatro proteínas, com diferentes estruturas e tamanhos: 1ZDD, 1CRN, 1ENH e 1AIL.

Uma importante contribuição desse trabalho foi a análise comparativa entre as funções de energia CHARMM, AMBER e ROSETTA para a predição de estrutura de proteínas *Ab Initio*. Para essa análise foram feitas simulações com todas as funções de energia e elas demonstraram a capacidade de identificar as principais estruturas das proteínas alvo com exceção das folhas β da proteína 1CRN. Quando comparados os resultados obtidos para as funções de energia, cada uma delas obteve resultados mais próximos da conformação de cada uma das três proteínas. A ROSETTA teve melhor valor para a 1PLW, a CHARMM para a proteína 1ZDD e AMBER para a proteína 1CRN.

Apesar da hipótese inicial de que: a quantidade de componentes na formula das funções de energia está relacionada ao melhor valor de RMSDα, não ter sido validado, uma vantagem observada em se utilizar a função de energia ROSETTA está relacionado com o esforço computacional empreendido, tendo diferenças de 15 horas em comparação com a CHARMM e AMBER. Essa diferença está relacionada ao excesso de acesso em disco que o pacote de dinâmica molecular responsável por elas necessita. Dessa forma, recomenda-se a utilização da função de energia ROSETTA.

Em todas as funções de energia foi observado a dificuldade em predizer as folhas β para a proteína 1CRN. Esse comportamento está relacionado a estrutura primária da proteína pois, os aminoácidos que compõem a folha β estão no início e no fim da estrutura primária. Com isso, os algoritmos não conseguiram fazer o alinhamento necessário para que haja interação entre os componentes para que a estrutura se forme.

5.1 TRABALHOS FUTUROS

Existem diversos trabalhos futuros que podem ser explorados futuramente para a solução do PPEP. Dentre eles, pode-se citar:

- Aplicar estratégias de intensificação nas versões que mantiveram valores de diversidade ao fim do processo de otimização, refinando a busca e aproveitando a diversidade restante.

- Agregar mais informações ao processo de predição de estrutura de proteínas é uma abordagem interessante para melhorar a qualidade da predição.

- Outra abordagem interessante é a utilização do índice de diversidade obtido durante o processo de otimização como informação de *feedback* para o algoritmo e transformá-lo em uma abordagem adaptativa, diminuindo a quantidade de parâmetros que devem ser configurados *a priori*. Essa informação pode ser utilizada para controle dos parâmetros de *crossover*, mutação e quantidade de avaliações. Tornar a abordagem $\text{DE}_{cascata}$ adaptativa conforme o índice de diversidade é interessante também.

- Testar novas configurações para o $\text{DE}_{cascata}$, seja alterando a quantidade de partições do processo de otimização ou até mesmo outros mecanismos de mutação.

- Em nenhum trabalho encontrado na literatura foram utilizadas as conformações já encontradas em modelos simplificados como informação de entrada para a resolução de modelos atômicos. A possibilidade de correlacionar os modelos é algo que deve ser investigado.

5.2 TRABALHOS PUBLICADOS

Alguns trabalhos foram publicados no decorrer do desenvolvimento dessa Dissertação. São eles:

- P. H. Narloch and R. S. Parpinelli. Análise da Diversidade Genotípica no Algoritmo Evolução Diferencial Aplicado ao Problema de Predição de Estrutura de Proteínas. In Workshop em Bioinformática, UTFPR, Cornélio Procópio, 2016.

- P. H. Narloch and R. S. Parpinelli. Diversification Strategies in Differential Evolution Algorithm to Solve the Protein Structure Prediction Problem. In A. M. Madureira, A. Abraham, D. Gamboa, and P. Novais, editors, Intelligent Systems Design and Applications, volume 557, pages 125–134. Springer International Publishing, Cham, 2017.

- P. H. Narloch and R. S. Parpinelli. The Protein Structure Prediction Problem Approached by a Cascade Differential Evolution Algorithm Using Rosetta. In 7th Brazilian Conference on Intelligent Systems (BRACIS), 2017.

- R. S. Parpinelli, G. F. Plichoski, R. S. da Silva and P. H. Narloch. A Review of Technique for On-line Control of Parameters in Swarm Intelligence and Evolutionary Computation Algorithms. In Engineering Applications of Artificial Intelligence. **Submetido para Avaliação**

Referências

ANFINSEN, C. B.; HABER, E.; SELA, M.; WHITE, F. H. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proceedings of the National Academy of Sciences*, v. 47, n. 9, p. 1309–1314, 1961.

BÄCK, T.; SCHWEFEL, H.-P. An overview of evolutionary algorithms for parameter optimization. *Evol. Comput.*, MIT Press, Cambridge, MA, USA, v. 1, n. 1, p. 1–23, mar. 1993.

BERMAN, H. M. The Protein Data Bank. *Nucleic Acids Research*, v. 28, n. 1, p. 235–242, jan. 2000.

BONNEAU, R.; BAKER, D. Ab Initio Protein Structure Prediction: Progress and Prospects. *Annual Review of Biophysics and Biomolecular Structure*, v. 30, n. 1, p. 173–189, jun. 2001.

BORGUESAN, B.; SILVA, M. B. e; GRISCI, B.; INOSTROZA-PONTA, M.; DORN, M. Apl: An angle probability list to improve knowledge-based metaheuristics for the three-dimensional protein structure prediction. *Computational Biology and Chemistry*, v. 59, n. Part A, p. 142 – 157, 2015.

BOUSSAïD, I.; LEPAGNOT, J.; SIARRY, P. A survey on optimization metaheuristics. *Information Sciences*, v. 237, n. Supplement C, p. 82 – 117, 2013. Prediction, Control and Diagnosis using Advanced Neural Computations.

BOUTET, E.; LIEBERHERR, D.; TOGNOLLI, M.; SCHNEIDER, M.; BANSAL, P.; BRIDGE, A. J.; POUX, S.; BOUGUELERET, L.; XENARIOS, I. UniProtKB/Swiss-Prot, the Manually Annotated Section of the UniProt KnowledgeBase: How to Use the Entry View. Springer New York, New York, NY, v. 1374, p. 23–54, 2016.

BROOKS, B. R.; BROOKS, C. L.; MACKERELL, A. D.; NILSSON, L.; PETRELLA, R. J.; ROUX, B.; WON, Y.; ARCHONTIS, G.; BARTELS, C.; BORESCH, S.; CAFLISCH, A.; CAVES, L.; CUI, Q.; DINNER, A. R.; FEIG, M.; FISCHER, S.; GAO, J.; HODOSCEK, M.; IM, W.; KUCZERA, K.; LAZARIDIS, T.; MA, J.; OVCHINNIKOV, V.; PACI, E.; PASTOR, R. W.; POST, C. B.; PU, J. Z.; SCHAEFER, M.; TIDOR, B.; VENABLE, R. M.; WOODCOCK, H. L.; WU, X.; YANG, W.; YORK, D. M.; KARPLUS, M. CHARMM: The biomolecular simulation program. *Journal of Computational Chemistry*, v. 30, n. 10, p. 1545–1614, jul. 2009.

CALVO, J.; ORTEGA, J.; ANGUITA, M. Pitagoras-psp: Including domain knowledge in a multi-objective approach for protein structure prediction. *Neurocomputing*, v. 74, n. 16, p. 2675 – 2682, 2011. Advances in Extreme Learning Machine: Theory and Applications Biological Inspired Systems. Computational and Ambient Intelligence.

CASTRO, L. N. D.; TIMMIS, J. *Artificial immune systems: a new computational intelligence approach*. London ; New York: Springer, 2002.

CORNELL, W. D.; CIEPLAK, P.; BAYLY, C. I.; GOULD, I. R.; MERZ, K. M.; FERGUSON, D. M.; SPELLMEYER, D. C.; FOX, T.; CALDWELL, J. W.; KOLLMAN, P. A. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *Journal of the American Chemical Society*, v. 117, n. 19, p. 5179–5197, maio 1995.

CORRIVEAU, G.; GUILBAULT, R.; TAHAN, A.; SABOURIN, R. Review and Study of Genotypic Diversity Measures for Real-Coded Representations. *"IEEE Transactions on Evolutionary Computation"*, IEEE, v. 16, n. 5, p. 695–710, out. 2012.

CORRIVEAU, G.; GUILBAULT, R.; TAHAN, A.; SABOURIN, R. Review of phenotypic diversity formulations for diagnostic tool. *Applied Soft Computing*, v. 13, n. 1, p. 9 – 26, 2013.

CUTELLO, V.; NARZISI, G.; NICOSIA, G. A multi-objective evolutionary approach to the protein structure prediction problem. *Journal of The Royal Society Interface*, v. 3, n. 6, p. 139–151, fev. 2006.

CUTELLO, V.; NARZISI, G.; NICOSIA, G. Computational studies of peptide and protein structure prediction problems via multiobjective evolutionary algorithms. Springer Berlin Heidelberg, Berlin, Heidelberg, p. 93–114, 2008.

DILL, K. A.; BROMBERG, S.; YUE, K.; CHAN, H. S.; FTEBIG, K. M.; YEE, D. P.; THOMAS, P. D. Principles of protein folding - a perspective from simple exact models. *Protein Science*, Cold Spring Harbor Laboratory Press, v. 4, n. 4, p. 561–602, 1995.

DORIGO, M.; CARO, G. D. New ideas in optimization. In: CORNE, D.; DORIGO, M.; GLOVER, F.; DASGUPTA, D.; MOSCATO, P.; POLI, R.; PRICE, K. V. (Ed.). Maidenhead, UK, England: McGraw-Hill Ltd., UK, 1999. cap. The Ant Colony Optimization Meta-heuristic, p. 11–32.

DORN, M.; BURIOL, L. S.; LAMB, L. C. A hybrid genetic algorithm for the 3-D protein structure prediction problem using a path-relinking strategy. *2011 IEEE Congress of Evolutionary Computation (CEC)*, IEEE, Los Angeles, United States of America, p. 2709–2716, 2011.

DORN, M.; SILVA, M. B. e; BURIOL, L. S.; LAMB, L. C. Three-dimensional protein structure prediction: Methods and computational strategies. *Computational Biology and Chemistry*, v. 53, n. Part B, p. 251 – 276, 2014. ISSN 1476-9271.

FILHO, C. J. A. B.; NETO, F. B. de L.; LINS, A. J. C. C.; NASCIMENTO, A. I. S.; LIMA, M. P. *Fish School Search*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009. 261–277 p.

FRAENKEL, A. S. Complexity of protein folding. *Bulletin of Mathematical Biology*, v. 55, n. 6, p. 1199–1210, 1993.

GUNSTEREN, W. F. van; DAURA, X.; MARK, A. E. GROMOS Force Field. John Wiley & Sons, Ltd, Chichester, UK, abr. 2002.

GUPTA, D.; GHAFIR, S. An overview of methods maintaining diversity in genetic algorithms. *International journal of emerging technology and advanced engineering*, v. 2, n. 5, p. 56–60, 2012.

GUYEUX, C.; CÔTÉ, N. M.-L.; BAHI, J. M.; BIENIA, W. Is protein folding problem really a np-complete one? first investigations. *Journal of Bioinformatics and Computational Biology*, v. 12, n. 01, p. 1350017, 2014.

HALGREN, T. A. Potential energy functions. *Current Opinion in Structural Biology*, v. 5, n. 2, p. 205 – 210, 1995.

HERRERA, F.; LOZANO, M. Adaptation of genetic algorithm parameters based on fuzzy logic controllers. *Genetic Algorithms and Soft Computing*, v. 8, p. 95–125, 1996.

HOLLAND, J. H. *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence*. Cambridge, MA, USA: MIT Press, 1992.

HUMPHREY, W.; DALKE, A.; SCHULTEN, K. VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics*, v. 14, p. 33–38, 1996.

INOSTROZA-PONTA, M.; FARFáN, C.; DORN, M. A memetic algorithm for protein structure prediction based on conformational preferences of aminoacid residues. In: *Proceedings of the Companion Publication of the 2015 Annual Conference on Genetic and Evolutionary Computation*. New York, NY, USA: ACM, 2015. (GECCO Companion '15), p. 1403–1404.

JONG, K. A. D. *An Analysis of the Behavior of a Class of Genetic Adaptive Systems*. Tese (Doutorado), Ann Arbor, MI, USA, 1975.

JONG, K. A. D.; DE, K. A.; SARMA, J. Generation gaps revisited. Morgan Kaufmann, 1992.

JR., M. M. G.; ARAJO, A. F. R. Diversity-Based Adaptive Evolutionary Algorithms. InTech, fev. 2010.

KABSCH, W.; SANDER, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, Wiley Subscription Services, Inc., A Wiley Company, v. 22, n. 12, p. 2577–2637, 1983.

KAR, A. K. Bio inspired computing – a review of algorithms and scope of applications. *Expert Systems with Applications*, v. 59, n. Supplement C, p. 20 – 32, 2016.

KARABOGA, D.; GORKEMLI, B.; OZTURK, C.; KARABOGA, N. A comprehensive survey: artificial bee colony (abc) algorithm and applications. *Artificial Intelligence Review*, v. 42, n. 1, p. 21–57, Jun 2014.

KARI, L.; ROZENBERG, G. The many facets of natural computing. *Communications of the ACM*, v. 51, n. 10, p. 72, out. 2008.

KIM, D. E.; CHIVIAN, D.; BAKER, D. Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Research*, v. 32, n. Web Server, p. W526–W531, jul. 2004.

KOENIG, A. C. A study of mutation methods for evolutionary algorithms. *Advanced Topics in Artificial Intelligence*, University of Missouri-Rolla, 2002.

KOLINSKI, A.; SKOLNICK, J. Reduced models of proteins and their applications. *Polymer*, v. 45, n. 2, p. 511 – 524, 2004. Conformational Protein Conformations.

LAU, K. F.; DILL, K. A. A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules*, v. 22, n. 10, p. 3986–3997, 1989.

LEACH, A. R. *Molecular modelling : principles and applications*. 2nd. ed. Harlow, UK: Prentice Hall, 2001.

LEE, J.; WU, S.; ZHANG, Y. *Ab Initio Protein Structure Prediction*. Dordrecht: Springer Netherlands, 2009. 3–25 p.

MARTI-RENOM, M. A.; MADHUSUDHAN, M.; SALI, A. Alignment of protein sequences by their profiles. *Protein Science*, Cold Spring Harbor Laboratory Press, v. 13, n. 4, p. 1071–1087, 2004.

MáRQUEZ-CHAMORRO, A. E.; ASENCIO-CORTéS, G.; SANTIESTEBAN-TOCA, C. E.; AGUILAR-RUIZ, J. S. Soft computing methods for the prediction of protein tertiary structures: A survey. *Applied Soft Computing*, v. 35, n. Supplement C, p. 398 – 410, 2015.

NARLOCH, P. H.; PARPINELLI, R. S. Diversification Strategies in Differential Evolution Algorithm to Solve the Protein Structure Prediction Problem. *Intelligent Systems Design and Applications*, Springer International Publishing, Cham, v. 557, p. 125–134, 2017.

NARLOCH, P. H.; PARPINELLI, R. S. The protein structure prediction problem approached by a cascade differential evolution algorithm using rosetta. *7th Brazilian Conference on Intelligent Systems (BRACIS)*, 2017. To Appear.

NELSON, D. L.; COX, M. M.; LEHNINGER, A. L. *Lehninger principles of biochemistry*. 6., internat. ed. ed. New York: Freeman, 2013.

O, V. T. D.; TINOS, R. A self-organizing genetic algorithm for protein structure prediction. *Learning & Nonlinear Models*, v. 8, n. 3, p. 135–147, 2010.

OLIVEIRA, M. V.; FREITAS, A. R.; GUIMARÃES, F. G. Uma Estratégia De Ranking Baseada Em Diversidade Em Algoritmos Genéticos. *Simpósio Brasileiro de Pesquisa Operacional*, Rio de Janeiro, Brasil, 2012.

PAL, A. *Ab-Initio Protein Structure Prediction using Bacterial Foraging Optimization Algorithm*. Tese (Doutorado) — Jadavpur University KOLKATA, 2014.

PARPINELLI, R. S.; BENITIEZ, C. M.; CORDEIRO, J.; LOPES, H. S. Performance Analysis of Swarm Intelligence Algorithms for the 3d-AB off-lattice Protein Folding Problem. *Multiple-Valued Logic and Soft Computing*, v. 22, n. 3, p. 267–286, 2014.

PARPINELLI, R. S.; LOPES, H. S. An eco-inspired evolutionary algorithm applied to numerical optimization. *Nature and Biologically Inspired Computing*, IEEE, p. 466–471, out. 2011.

PARPINELLI, R. S.; LOPES, H. S. New inspirations in swarm intelligence: a survey. *International Journal of Bio-Inspired Computation*, Inderscience Publishers, Inderscience Publishers, Geneva, SWITZERLAND, v. 3, n. 1, p. 1–16, fev. 2011.

PARPINELLI, R. S.; LOPES, H. S. An Ecology-Based Evolutionary Algorithm Applied to the 2d-AB Off-Lattice Protein Structure Prediction Problem. In: *Proceedings of the 2013 Brazillian Conference on Intelligent Systems*. Fortaleza, Brazil: IEEE, 2013. p. 64–69.

POLI, R.; KENNEDY, J.; BLACKWELL, T. Particle swarm optimization. *Swarm Intelligence*, v. 1, n. 1, p. 33–57, Jun 2007.

POLLASTRI, G.; PRZYBYLSKI, D.; ROST, B.; BALDI, P. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins: Structure, Function, and Genetics*, v. 47, n. 2, p. 228–235, maio 2002.

PRICE, K.; STORN, R. M.; LAMPINEN, J. A. *Differential Evolution: A Practical Approach to Global Optimization (Natural Computing Series)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2005.

ROHL, C. A.; STRAUSS, C. E.; MISURA, K. M.; BAKER, D. Protein Structure Prediction Using Rosetta. *Methods in Enzymology*, Elsevier, Washington, United States of America, v. 383, p. 66–93, 2004.

ROMERO, D. C. B. *A Multi-objective Ab-initio Model for Protein Folding Prediction at an Atomic Conformation Level*. Tese (Doutorado) — Universidad Nacional de Colombia. Facultad de Ingeniera. Departamento de Ingeniera de Sistemas y Computacin, 2010.

RUSSELL, S. J.; NORVIG, P. *Artificial intelligence: a modern approach*. 3. ed. ed. Upper Saddle River, NJ: Prentice-Hall, 2010. (Prentice-Hall series in artificial intelligence).

SARENI, B.; KRAHENBUHL, L. Fitness sharing and niching methods revisited. *Trans. Evol. Comp*, IEEE Press, Piscataway, NJ, USA, v. 2, n. 3, p. 97–106, set. 1998.

SCALABRIN, M. H.; PARPINELLI, R. S.; BENíTEZ, C. M.; LOPES, H. S. Population-based harmony search using GPU applied to protein structure prediction. *International Journal of Computational Science and Engineering*, v. 9, n. 1/2, p. 106–118, 2014.

SCHNEIDER, M.; FU, X.; KEATING, A. E. X-ray vs. nmr structures as templates for computational protein design. *Proteins: Structure, Function, and Bioinformatics*, Wiley Subscription Services, Inc., A Wiley Company, v. 77, n. 1, p. 97–110, 2009.

SHAKHNOVICH, E. I.; GUTIN, A. M. Engineering of stable and fast-folding sequences of model proteins. *Proceedings of the National Academy of Sciences*, v. 90, n. 15, p. 7195–7199, 1993.

SHIR, O. M. Niching in evolutionary algorithms. *Handbook of Natural Computing*, Springer Berlin Heidelberg, Berlin, Heidelberg, p. 1035–1069, 2012.

SPRENT, N. C. S. P. *Applied Nonparametric Statistical Methods, Fourth Edition*. 4. ed. London, England: Chapman and Hall/CRC, 2007.

STILLINGER, F. H.; HEAD-GORDON, T. Collective aspects of protein folding illustrated by a toy model. *Phys. Rev. E*, American Physical Society, v. 52, p. 2872–2877, 1995.

STORN, R.; PRICE, K. Differential Evolution – A Simple and Efficient Heuristic for global Optimization over Continuous Spaces. *Journal of Global Optimization*, v. 11, n. 4, p. 341–359, 1997.

SáNCHEZ, R.; ŠALI, A. Advances in comparative protein-structure modelling. *Current Opinion in Structural Biology*, v. 7, n. 2, p. 206 – 214, 1997.

TANTAR, A.-A.; MELAB, N.; TALBI, E.-G.; PARENT, B.; HORVATH, D. A parallel hybrid genetic algorithm for protein structure prediction on the computational grid. *Future Generation Computer Systems*, v. 23, n. 3, p. 398 – 409, 2007.

THOMAS, P. D.; DILL, K. A. Local and nonlocal interactions in globular proteins and mechanisms of alcohol denaturation. *Protein Science*, v. 2, n. 12, p. 2050–2065, dez. 1993.

UNGER, R.; MOULT, J. Finding the lowest free energy conformation of a protein is an np-hard problem: Proof and implications. *Bulletin of Mathematical Biology*, v. 55, n. 6, p. 1183–1198, 1993.

VENSKE, S. M.; GONçALVES, R. A.; BENELLI, E. M.; DELGADO, M. R. Ademo/d: An adaptive differential evolution for protein structure prediction problem. *Expert Systems with Applications*, v. 56, n. Supplement C, p. 209 – 226, 2016.

WALSH, G. *Proteins: biochemistry and biotechnology*. 2nd ed. ed. Chichester: Wiley Blackwell, 2014.

WOOLEY, J. C.; YE, Y. *A Historical Perspective and Overview of Protein Structure Prediction*. New York, NY: Springer New York, 2007. 1–43 p.

XU, D.; ZHANG, Y. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins: Structure, Function, and Bioinformatics*, Wiley Subscription Services, Inc., A Wiley Company, v. 80, n. 7, p. 1715–1735, 2012.

ZHANG, W.; YANG, J.; HE, B.; WALKER, S. E.; ZHANG, H.; GOVINDARAJOO, B.; VIRTANEN, J.; XUE, Z.; SHEN, H.-B.; ZHANG, Y. Integration of QUARK and I-TASSER for Ab Initio Protein Structure Prediction in CASP11: Ab Initio Structure Prediction in CASP11. *Proteins: Structure, Function, and Bioinformatics*, v. 84, p. 76–86, set. 2016.

Apêndices

A Utilização do PyRosetta

O PyRosetta é uma interface interativa do pacote de dinâmica molecular ROSETTA utilizado para a manipulação de proteínas. Nesse trabalho essa interface é utilizada em conjunto com metaheurísticas para a predição de estrutura de proteínas. Nesse apêndice está um passo a passo simples para a utilização do PyRosetta. É necessário nesse caso que o pacote PyRosetta seja instalado conforme as configurações da máquina. Todas as informações de instalação do PyRosetta encontram-se no site da ferramenta[1].

Inicialmente é necessário que seja selecionada a proteína alvo que será predita. As proteínas utilizadas nessa dissertação possuem suas sequências e estruturas armazenadas no PDB. Nesse exemplo a proteína 1PLW, que possui 5 aminoácidos, é utilizada. É necessário que seja coletado a sequência de aminoácidos no modelo FASTA de representação, pois é essa notação que será utilizada como informação de entrada para as funções do PyRosetta. Para essa proteína, a sequência FASTA é "YGGFM". Outra ferramenta utilizada é o *SCRATCH Protein Predictor*[2], que com as informações da estrutura primária da proteína ele tenta encontrar suas estruturas secundárias para limitação de ângulos conforme a classificação DSSP-8.

Para a utilização do PyRosetta, é necessário que sua biblioteca seja importada para o ambiente de desenvolvimento com o comando `import pyrosetta` e suas variáveis inicializadas com `pyrosetta.init()`. Essas são duas exigências para a utilização de qualquer objeto para manipulação de macromoléculas do PyRosetta. A classe responsável pela função de energia é a `Score Function`. A partir da instanciação dos objetos dessa classe é possível ter acesso a função de energia *full atom* padrão do ROSETTA ou até mesmo modificar ela. Nesse caso a função de energia utilizada é a padrão, podendo ser obtida com a linha de comando `pyrosetta.get_fa_scorefxn()`.

Assim como existe a classe para as funções de energia, também há a classe para a representação das proteínas: `Pose()`. Para criar uma proteína utilizando os aminoácidos obtidos pelo PDB é necessário utilizar o comando `pyrosetta.pose_from_sequence(aminoacids, "fa_standard")`, sendo o primeiro argumento a sequência FASTA e o segundo a utilização de resíduos *full atom*. A partir do momento que o objeto `Pose` esteja instanciado é possível utilizar a função `scorefxn(pose)` para obter o valor de energia dessa proteína.

Para modificar os ângulos do objeto `Pose` é possível acessar os atributos do objeto através dos métodos: `set_phi(aminoacid, value)` para o ângulo ϕ; `set_psi(aminoacid, value)` para o ângulo ψ; `set_omega(aminoacid, value)` para o ângulo ω; `set_chi(chi_i,aminoacid, value)`. Sendo o parâmetro `aminoacid` referente

[1] http://www.pyrosetta.org/
[2] http://scratch.proteomics.ics.uci.edu/

a posição do aminoácido na estrutura primária, *value* o valor do ângulo e, no caso dos ângulos χ, é necessário informar primeiro qual a posição do χ.

Com base nessas informações, o Algoritmo 4 apresenta a utilização do pacote PyRosetta para a mensuração de energia de um objeto `Pose` que represente a proteína 1PLW.

Algorithm 4 Utilização do PyRosetta

```
1: import pyrosetta
2: pyrosetta.init()
3: aminosequence = 'YGGFM'
4: self.scorefxn = pyrosetta.get_fa_scorefxn()
5: self.generalPose = pyrosetta.pose_from_sequence(aminosequence, "fa_standard")
6: pose = self.generalPose
7: para i = 1 até Quantidade de Aminoácidos faça
8:     pose.set_phi(i, 180)
9:     pose.set_psi(i, 180)
10:    pose.set_omega(i, 180)
11:    para j = 1 até Quantidade χ faça
12:        pose.set_chi(j, i, 180)
13:    fim para
14: fim para
15: retorne score = self.scorefxn(pose)
```

A sequência de comandos apresentados no Algoritmo 4 apresenta a utilização do PyRosetta nessa dissertação, utilizando a representação de ângulos de torsões da cadeia principal e secundária e a função de energia *full atom* padrão do ROSETTA com a linguagem Python em sua versão 3. Outros comandos que podem servir para a manipulação de informações do PyRosetta estão presentes em: <http://graylab.jhu.edu/pyrosetta/downloads/documentation/PyRosetta_Workshops_Appendix_A.pdf>.