UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

BRUNO FERREIRA DE FARIA ALIXANDRE

Explorando a Diversidade em Algoritmos Genéticos Distribuídos Para o Problema de Predição de Estrutura Tridimensional de Proteínas

Dissertação apresentada como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação

Orientador: Prof. Dr. Márcio Dorn

Porto Alegre
2018

CIP — CATALOGAÇÃO NA PUBLICAÇÃO

Alixandre, Bruno Ferreira de Faria

Explorando a Diversidade em Algoritmos Genéticos Distribuídos Para o Problema de Predição de Estrutura Tridimensional de Proteínas / Bruno Ferreira de Faria Alixandre. – Porto Alegre: PPGC da UFRGS, 2018.

117 f.: il.

Dissertação (mestrado) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR–RS, 2018. Orientador: Márcio Dorn.

1. Bioinformática Estrutural. 2. Predição de Estrutura Tridimensional de Proteínas. 3. Otimização. 4. Meta-heurísticas. 5. Algoritmo Genético Distribuído. 6. Controle de Diversidade. I. Dorn, Márcio. II. Título.

“Quem não vive para servir, não serve pra viver”

— MAHATMA GANDHI

AGRADECIMENTOS

A Deus e todos os meu familiares, em especial meus pais Edson e Maria Aparecida, por proverem todos os meios aos quais aqui me trouxeram, além de todo o amor e compreensão despendidos nessa jornada. A meus avós e padrinhos Geraldo (*in memoriam*) e Iracema, por todo o carinho, atenção e ensinamentos ao longo de minha vida. Agradeço também ao meu orientador Professor Doutor Márcio Dorn, pelo apoio incondicional e conduta exemplar, os quais foram imprescindíveis para meu êxito nessa etapa de minha carreira. Agraço também a todos os meus amigos de toda a vida, que sempre se fizeram presentes nos momentos mais difíceis. Também agradeço a todos os meus companheiros de laboratório, em especial aos Mestres Bruno Borguesan e Leonardo de Lima Corrêa, por toda a ajuda e dedicação, apoio este determinante no sucesso desta pesquisa. Todos vocês foram peças fundamentais na construção deste sonho, o meu muito obrigado a todos. Por fim, agradeço à Universidade Federal do Rio Grande do Sul, por meio de seu corpo docente e funcionários, por proverem todo o suporte necessário ao longo destes dois anos de pesquisa, além do CNPq pelo financiamento da mesma.

RESUMO

Conhecer a estrutura tridimensional de uma proteína é um passo fundamental para se determinar seu papel biológico. Ainda que o problema de predição da estrutura de proteínas permaneça sem solução definitiva, diversos métodos foram propostos nos últimos anos no intuito de solucioná-lo. Como principal alternativa, meta-heurísticas como o Algoritmo Genético e versões similares vem sendo aplicadas ao problema. Apesar de eficientes em diversas tarefas, estes algoritmos encontram alguns obstáculos ao tratarem o problema, principalmente devido à convergência prematura. Isso ocorre quando a população do algoritmo perde diversidade, convergindo então em valores não satisfatórios. Muitas maneiras de lidar com este problema foram desenvolvidas. O Algoritmo Genético Distribuído se divide em múltiplas populações, de modo a desacelerar a convergência global. Nesta dissertação, um Algoritmo Genético Distribuído com controle de diversidade é apresentado. Para tal, uma nova política de migração foi proposta, a qual busca manter e explorar de maneira eficiente a diversidade populacional. Antes de aplicá-la ao problema de predição da estrutura de proteínas, uma série de testes foram realizados, com o objetivo de se obter a versão mais adequada ao problema. Em uma primeira etapa, esta nova política foi testada em conjunto com outras cinco políticas distintas, todas aplicadas no Algoritmo Genético canônico e no *Biased Random-Key Genetic Algorithm* (BRKGA), uma versão também baseada em diversidade. Os resultados apontam para um melhor desempenho da política proposta em relação às demais, principalmente combinada ao BRKGA. Na etapa seguinte, uma série de operadores genéticos foram propostos, no intuito de agregar informações específicas do problema à tomada de decisão. Além disso, utilizou-se também uma base de conhecimento experimental visando reduzir o espaço de busca conformacional. Como resultado, um novo método foi gerado (política proposta + BRKGA), incorporando novos operadores de inicialização baseado em conhecimento experimental, além de operadores de recombinação e mutação que utilizam informação da estrutura secundária. Por fim, este método foi comparado a outros dois que também fazem uso de informação da estrutura secundária e de bases de conhecimento. Os resultados apontam para um desempenho competitivo, se mostrando superior em diversos casos.

Palavras-chave: Bioinformática Estrutural. Predição de Estrutura Tridimensional de Proteínas. Otimização. Meta-heurísticas. Algoritmo Genético Distribuído. Controle de Diversidade.

ABSTRACT

The knowledge of the three-dimensional structure of a protein is a key step in determining its biological role. Although the protein structure prediction problem remains without a definitive solution, several methods have been proposed in the last years aiming to solve it. As the main alternative, meta-heuristics such as the Genetic Algorithm and similar versions have been applied to the problem. Despite efficient in several tasks, these algorithms face some obstacles when dealing with this problem, mainly due to the premature convergence. This happens when the population of the algorithm loses diversity, converging to unsatisfactory values. Many ways of dealing with this problem have been developed. The Distributed Genetic Algorithm seeks to divide the population into multiple groups to slow global convergence. In this dissertation, a Distributed Genetic Algorithm with diversity control is presented. To this end, a new migration policy has been developed, in which it seeks not only to maintain but also to efficiently explore population diversity. Before applying it to the protein strucure prediction problem, several tests were performed, in order to obtain the most appropriate version of the method to the problem. In a first stage, this new policy was tested along with five other distinct policies, all applied in the Canonical Genetic Algorithm and the Biased Random-Key Genetic Algorithm (BRKGA), a version also based on diversity. The results point to a better performance of the proposed policy in relation to the others, mainly combined with BRKGA. In the next stage, a series of genetic operators were proposed, to aggregate specific information of the problem to the decision making. Also, an experimental knowledge base was used to reduce the conformational search space. As a result, a novel method (proposed policy + BRKGA) was generated, incorporating new initialization operators based on experimental knowledge, as well as recombination and mutation operators that use information from the secondary structure of proteins. Finally, this method was compared to other two methods that also make use of secondary structure information and knowledge bases. The results point to a competitive performance, proving superior in several cases.

Keywords: Structural Bioinformatics, Three-dimensional Protein Strucute Prediction, Optimization, Metaheuristics, Distributed Genetic Algorithm, Diversity Control.

LISTA DE ABREVIATURAS E SIGLAS

AG	Algoritmo Genético
AGD	Algoritmo Genético Distribuído
AM	Algoritmo Memético
APL	*Angle Probability List*
BRKGA	*Biased Random-Key Genetic Algorithm*
CEC	*Congress on Evolutionary Computation*
GDT	*Global Distance Test*
MR	*Migration rate*
MS	*Migration size*
NMR	*Nuclear Magnetic Resonance*
PDB	*Protein Data Bank*
PSP	*Protein Structure Prediction*
RefSeq	*Reference Sequence Database*
RES	Reforço de Estrutura Secundária
RMSD	*Root Mean Square Deviation*

LISTA DE SÍMBOLOS

Å	*Ângström*
ϕ	Ângulo *phi*
ψ	Ângulo *psi*
ω	Ângulo *omega*
χ	Ângulo *chi*

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Em Bioinformática Estrutural existem vários problemas que até o presente momento não apresentam um método computacional que, mesmo com os avanços das tecnologias correntes e do poder computacional, possa garantir solução ótima. Em particular, a Bioinformática Estrutural trata de problemas onde, as regras que governam os processos bioquímicos e suas relações são parcialmente conhecidos o que torna difícil o desenvolvimento de estratégias computacionais eficientes (LESK, 2013; VERLI, 2014). Dentre estes problemas, dois deles são os mais desafiadores: a predição da estrutura tridimensional de macromoléculas (PSP - *Protein Structure Prediction*) e o problema de atracamento molecular (*Molecular Docking*) presente na área de desenvolvimento de fármacos. Predizer a estrutura de um polipeptídeo/proteína, apenas partindo de sua sequência linear de resíduos de aminoácidos representa um problema desafiador no campo da otimização matemática. O desafio ocorre devido à explosão de possíveis conformações que uma longa cadeia de resíduos de aminoácidos pode assumir (SCHEEF; FINK, 2005).

Considerado um dos principais problemas da Bioinformática Estrutural, o problema PSP pode ser definido como a utilização de meios computacionais para, a partir da estrutura primária (sequência de aminoácidos), predizer a estrutura (conformação) tridimensional de um polipeptídeo (DORN et al., 2014a). Predizer a estrutura correta de proteínas é uma tarefa intrigante e árdua (NGO; MARKS; KARPLUS, 1997). O problema de predição do enovelamento de uma proteína é classificado em complexidade computacional como um problema NP completo, isto é, ele está entre os mais difíceis problemas em termos de requisitos computacionais (CRESCENZI et al., 1998). Esta complexidade deve-se ao fato de o processo de enovelamento de uma proteína ser extremamente seletivo. Uma longa cadeia de resíduos de aminoácidos acaba assumindo um imenso número de conformações (LEVINTHAL, 1968). Determinar experimentalmente a estrutura de uma proteína é um processo caro (devido aos custos associados com a cristalografia ou ressonância magnética nuclear), e demorado (TRAMONTANO; LESK, 2006). Consequentemente, principalmente em decorrência do sequenciamento do genoma humano, muitas sequências de aminoácidos não possuem sua estrutura tridimensional conhecida. Para ser mais preciso, o *RefSeq* [1] (*Reference Sequence Database*), um dos principais banco de sequências de aminoácidos, possui aproximadamente 95 milhões de sequências catalogadas, ao passo que o principal repositório de estruturas tridimensionais de proteínas, o PDB

[1] <https://www.ncbi.nlm.nih.gov/refseq/>

(*Protein Data Bank*) [2] (BERMAN et al., 2006), possui em torno de apenas 125 mil estruturas conhecidas. Essa discrepância fez com que diversas abordagens computacionais fossem propostas ao logo dos últimos anos para se lidar com o PSP.

Estes métodos podem ser divididos em quatro classes (DORN et al., 2014a): (*i*) métodos de primeiros princípios que não utilizam informações estruturais de proteínas de bases de dados, também chamados de métodos *ab initio* (OSGUTHORPE, 2000); (*ii*) métodos de primeiros princípios que utilizam informações estruturais de bases de dados (ROHL et al., 2004); (*iii*) métodos de alinhavamento de estruturas (BOWIE; LUTHY; EISENBERG, 1991); e (*iv*) métodos por modelagem comparativa (MARTÍ-RENOM et al., 2000). O primeiro grupo busca realizar a predição sem nenhuma informação prévia de outras estruturas conhecidas, guiando-se apenas em simulações computacionais de propriedades físico-químicas do processo de enovelamento de proteínas na natureza, tendo como função objetivo um modelo de campo de força que determina a energia livre da molécula, partindo do princípio que a estrutura nativa corresponde ao mínimo global dessa função (ANFINSEN, 1973). Apesar de não dependerem de nenhum conhecimento além da sequência de aminoácidos, estes métodos encontram dificuldades principalmente pela alta dimensionalidade e complexidade do espaço de busca conformacional. Os grupos *ii*, *iii* e *iv* são classificados como métodos baseados em conhecimento, fazendo uso de informações relacionadas a fragmentos estruturais ou estruturas completas de proteínas determinadas experimentalmente. Estes métodos só podem ser aplicados quando houver informações estruturais disponíveis, ficando limitados a uma base de dados de proteínas. Especificamente, o grupo *ii* representa uma classe híbrida de métodos, no qual utilizam informações de fragmentos de aminoácidos combinadas a uma abordagem puramente *ab initio* (SRINIVASAN; ROSE, 1995). Dessa forma, para predizer a estrutura tridimensional de proteínas através de abordagens baseadas em primeiros princípios, diversas meta-heurísticas estão sendo aplicadas, uma vez que a estrutura nativa tende a representar valores mínimos de energia livre, caracterizando assim um processo de otimização (ANFINSEN, 1973; DORN et al., 2014a).

Meta-heurísticas são uma das mais comuns e poderosas técnicas utilizadas em situações onde o conhecimento sobre o problema é restrito e soluções exatas não são atualmente computáveis (CRAINIC; TOULOUSE, 2003; TANTAR; MELAB; TALBI, 2008; ABUAL-RUB et al., 2012). Meta-heurísticas não garantem a solução ótima, mas elas proporcionam uma boa aproximação com um esforço computacional limitado. Apesar

[2] <http://www.rcsb.org/>

disto, muitas técnicas apresentam dificuldades em escapar de ótimos locais e, consequentemente, convergem prematuramente em soluções não satisfatórias (LEUNG; WANG, 2001; CRAINIC; TOULOUSE, 2003; GENDREAU; POTVIN, 2005). Uma maneira de se escapar de ótimos locais é através da manutenção da diversidade populacional, uma vez que soluções diversas permitem explorar diferentes regiões do espaço de busca e consequentemente aumentar as chances de convergir para soluções de qualidade (HUANG; CHEN, 2001; URSEM, 2002). Outra maneira de se incrementar a qualidade das soluções obtidas consiste na incorporação de conhecimento experimental no processo de otimização (LESK, 2010). Bases de dados como os fragmentos do *Rosetta* (KIM; CHIVIAN; BAKER, 2004) são utilizadas em diversos trabalhos com o intuito de se reduzir as possibilidades do espaço de busca, guiando-se pelo conhecimento do problema.

1.1 Motivação

Apesar do grande volume de dados (sequências de resíduos de aminoácidos) obtidos pelo Projeto Genoma [3], existe uma grande discrepância em relação ao número de estruturas tridimensionais de proteínas conhecidas (DORN et al., 2014a). Considerando também o alto custo de se obter tais estruturas por meios experimentais, acredita-se que, mesmo ainda não atingindo resultados ótimos, os meios computacionais possam ser uma alternativa adequada para solucionar-se tal problema. Tais desafios motivam o estudo e desenvolvimento de meta-heurísticas com capacidade de lidar com grandes dimensionalidades e complexos espaços de busca. Apesar do sucesso dessas técnicas em diversos casos, problemas como o PSP podem ser extremamente complexos e apresentarem certas limitações (e.g., ineficiência das funções de energia em descrever precisamente o estado da estrutura analisada). Abordagens como o Algoritmo Genético (AG) e suas variações vem sendo utilizadas na tentativa de solucionar tal problema. Versões aprimoradas, como o Algoritmo Genético Distribuído (AGD), buscam lidar com problemas durante o processo de otimização tais como convergência prematura e perda de diversidade. Dessa forma, esta dissertação apresenta um estudo e desenvolvimento de um AGD com mecanismos de manutenção de diversidade e que explore de maneira eficiente bases de conhecimento experimental, tornando-o apto a lidar com tais desafios do problema PSP.

[3] DOE *Genomic Science*. <http://genomics.energy.gov>

1.2 Objetivos e Metas

O objetivo geral deste trabalho consiste em desenvolver uma meta-heurística capaz de lidar com problemas com espaço de busca altamente complexos, explorando conceitos de manutenção de diversidade e conhecimento experimental, aplicada ao problema PSP. Para tal, pretende-se inicialmente avaliar a abordagem desenvolvida em um ambiente controlado e genérico, livre de interferências indesejadas (e.g., viés das funções de energia) e, posteriormente ajustar tal abordagem ao problema PSP. Os objetivos específicos são: FFF

1. Desenvolver estratégias que possibilitem evitar a convergência em ótimos locais por meio de mecanismos de manutenção de diversidade;
2. Identificar maneiras eficientes de se utilizar conhecimento experimental no intuito de reduzir e espaço de busca e melhorar a qualidade das soluções encontradas;
3. Analisar as principais características dos problemas de predição de estruturas tridimensional de proteínas, incorporando tais conhecimentos à abordagem proposta, visando aprimorar as soluções obtidas;
4. Implementar e validar a performance do método proposto em um ambiente não enviesado, buscando uma maior precisão na validação do mesmo;
5. Adaptar o método proposto, implementando de maneira construtiva componentes que agreguem informações acerca do problema bem como a utilização de bases de conhecimento;
6. Avaliar a performance da versão completa do método no problema PSP, analisando tanto a precisão da otimização quanto a qualidade estrutural das soluções;

1.3 Organização do Trabalho

Esta dissertação está organizada da seguinte forma: o Capítulo 2 fornece os conceitos necessários para o entendimento do problema, abordando tanto os aspectos biológicos, quanto computacional; o Capítulo 3 apresenta de maneira detalhada os trabalhos relacionados ao método desenvolvido; o Capítulo 4 apresenta a metodologia utilizada e a concepção da abordagens proposta; o Capítulo 5 traz os resultados dos experimentos e as discussões à respeito; e por fim, o Capítulo 6 conclui o trabalho além de indicar trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo serão abordados conceitos e bases teóricas necessários para a compreensão do problema PSP, que envolve desde noções de caráter biológico de proteínas, até maneiras de solucionar o problema com métodos computacionais.

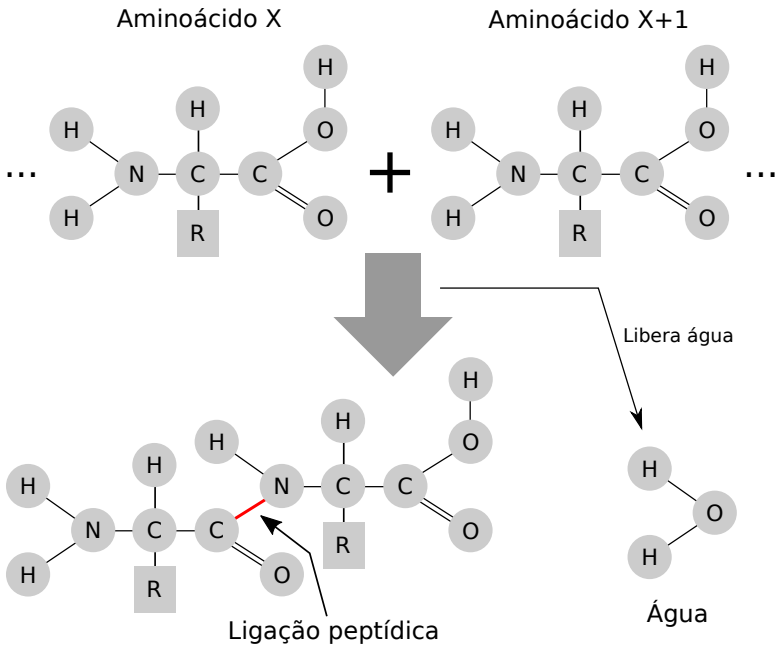
2.1 Proteínas

Podemos definir, sob uma visão estrutural, que proteínas (ou polipeptídeos) são moléculas compostas por uma sequência ordenada de resíduos de aminoácidos. Cada proteína possui uma sequência única e, sob condições fisiológicas naturais, se enovela em uma conformação denominada estado nativo (ANFINSEN, 1973; FASMAN, 1989). A estrutura dos resíduos de aminoácidos podem ser divididas em duas partes: (a) cadeia principal e; (b) cadeia lateral. Igualmente presente em todos os tipos de aminoácidos, a cadeia principal é composta por um grupo amino (H_2N^+), um grupo carboxílico ($COOH^-$) e um hidrogênio ligado à um carbono alfa (C_α) que une estes dois grupos. A cadeia lateral, também denominada de grupo R, é responsável por atribuir ao resíduo suas propriedades físico-químicas, de modo a diferenciá-los uns dos outros. O grupo R pode se distinguir em relação ao seu tamanho, à sua polaridade e à sua carga elétrica, sendo que dependendo da polaridade, o resíduo pode assumir caráter hidrofílico ou hidrofóbico. Em decorrência disto, 20 resíduos de aminoácidos distintos são encontrados na natureza, sendo tais diferenças determinantes no processo de enovelamento de uma proteína (RICHARDSON, 1981; LODISH et al., 1990).

Durante seu processo de sintetização, os aminoácidos de uma proteína são unidos por ligações peptídicas onde o grupo carboxílico de um resíduo reage com o grupo amino de outro, liberando uma molécula de água (H_2O) (BRANDEN; TOOZE, 1999; LESK, 2013). A Figura 2.1 ilustra este processo. Após a ligação, os dois resíduos formam um peptídeo, sendo que peptídeos maiores são referidos como polipeptídeos (CREIGHTON, 1990; LESK, 2010).

Na ligação peptídica (C-N) encontra-se o ângulo de rotação Omega (ω), todavia, devido ao seu caráter de ligação parcialmente dupla (devido à ressonância), este ângulo tende a ser planar, $\omega = 180°$ (*cis*, mais comum) ou $\omega = 0°$ (*trans*, mais raro), resultando em pouca mobilidade da molécula em torno desta ligação (BRANDEN; TOOZE, 1999). Rotações ocorrem com maior liberdade nas ligações N-C_α e C_α-C, ângulos Phi (ϕ) e Psi (ψ)

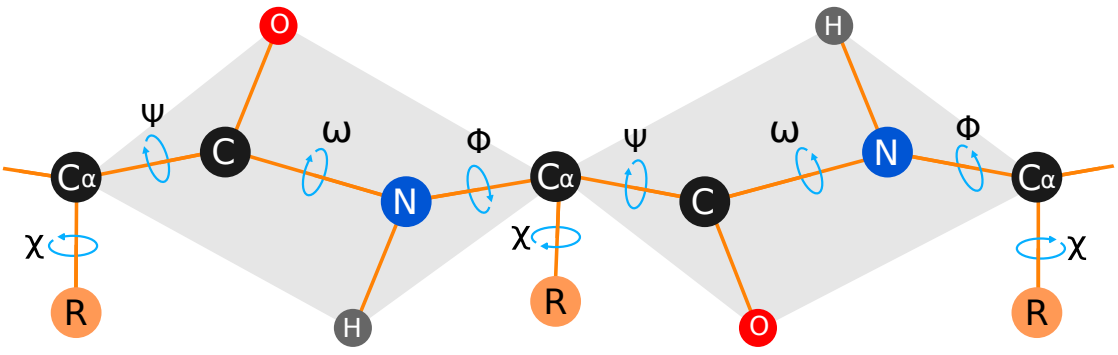
Figura 2.1: Ligação Peptídica. Grupos amino e carboxílico se ligam, liberando uma molécula de água.



Fonte: Do Autor.

respectivamente. Apesar de certas limitações de rotação devido a fatores de interferência estérica entre as cadeias principal e lateral, os ângulos ϕ e ψ são os principais responsáveis pela conformação adotada pela proteína (RICHARDSON, 1981; SCHEEF; FINK, 2005). De modo semelhante, os ângulos `Chi` (χ) são responsáveis pelo posicionamento da cadeia lateral e influenciam na estabilidade da conformação e no empacotamento da proteína. Também com rotações livres (de -180° a 180°), a quantidade de ângulos χ varia de 0 a 4 de acordo com o número de átomos da cadeia lateral. A Figura 2.2 ilustra esses ângulos.

Figura 2.2: Ângulos de rotação. ϕ, ψ e ω presentes na cadeia principal. χ's presentes na cadeia lateral.



Fonte: Adaptado de Borguesan et al. (2015a).

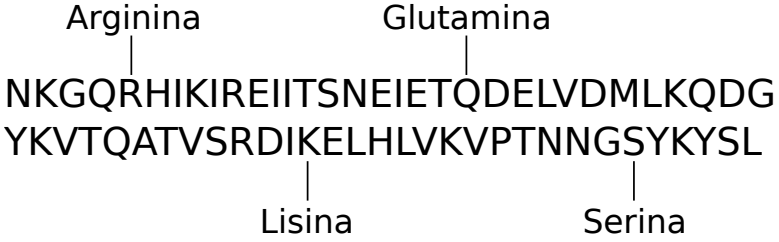
2.1.1 Níveis de Representação Estrutural

Proteínas podem ser representadas em quatro níveis estruturais: (i) estrutura primária, que se define pela sequência de aminoácidos que compõem a proteína; (ii) estrutura secundária, que são padrões de organização espacial recorrentes em trechos ao longo da sequência de aminoácidos ocorridos em função de suas propriedades físico-químicas; (iii) estrutura terciária, que consiste no agrupamento e interações dos padrões da estrutura secundária; e (iv) estrutura quartenária, onde tem-se o agregado de macromoléculas (NELSON; COX; LEHNINGER, 2005; LILJAS et al., 2009; LESK, 2013; VERLI, 2014). Cada um destes níveis estruturais são apresentados nas próximas seções.

2.1.1.1 Estrutura Primária

A estrutura primária descreve, de modo linear, a sequência de resíduos de aminácidos presentes em uma proteína. Conforme apresentado na seção anterior, os resíduos de aminoácidos se conectam por meio de uma ligação peptídica, onde o começo e o fim da estrutura correspondem à regiões `N`-terminal e `C`-terminal respectivamente (LODISH et al., 1990; BRANDEN; TOOZE, 1999). A Tabela 2.1 apresenta os 20 resíduos de aminoácidos presentes na natureza. A Figura 2.3 apresenta a sequência de aminoácidos que compõem a estrutura primária da proteína de ID 2P5K no PDB.

Figura 2.3: Estrutura primária. Proteína PDB ID: 2P5K.



Fonte: Próprio Autor.

2.1.1.2 Estrutura Secundária

A estrutura secundária caracteriza-se pela presença de ligações de hidrogênio. Esta ligação ocorre entre os átomos de hidrogênio do grupo amino e os átomos de oxigênio ou nitrogênio do grupo carboxílico. Devido a isso, arranjos estáveis são formados, resultando em padrões estruturais chamados de estruturas secundárias.

Tabela 2.1: Lista de aminoácidos.

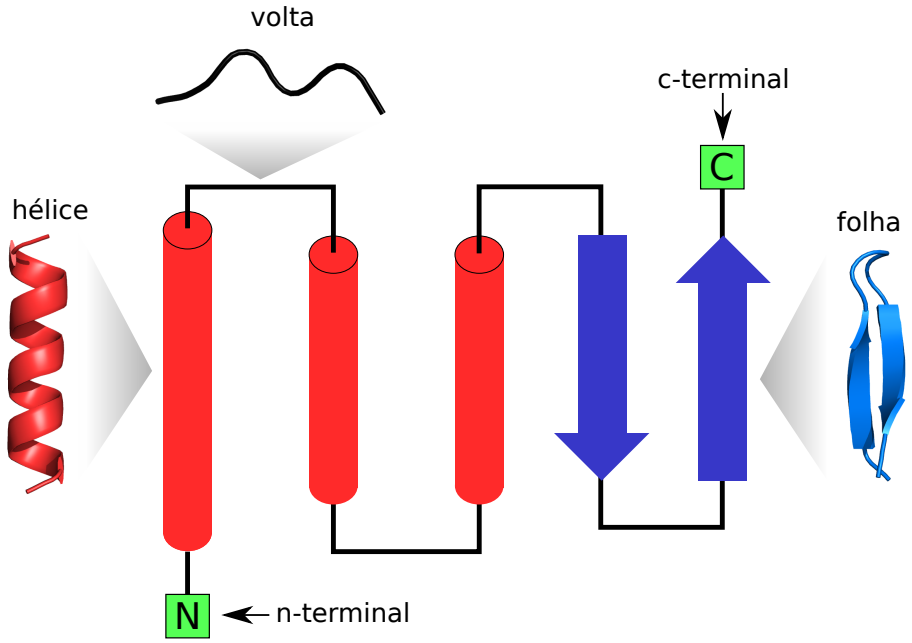
Aminoácido	Abreviação		Ângulos Chi	Polaridade
	3 Letras	1 Letra		
Alanina	ALA	A	0	Não-Polar
Arginina	ARG	R	4	Carregado Positivamente
Asparagina	ASN	N	2	Carregado Negativamente
Aspartato	ASP	D	2	Carregado Positivamente
Cisteína	CYS	C	1	Carregado Negativamente
Fenilalanina	PHE	F	2	Não-Polar
Glicina	GLY	G	0	Não-Polar
Glutamato	GLU	E	3	Carregado Positivamente
Glutamina	GLN	Q	3	Carregado Negativamente
Histidina	HIS	H	2	Carregado Positivamente
Isoleucina	ILE	I	2	Não-Polar
Leucina	LEU	L	2	Não-Polar
Lisina	LYS	K	4	Carregado Positivamente
Metionina	MET	M	3	Não-Polar
Prolina	PRO	P	0	Não-Polar
Serina	SER	S	1	Carregado Negativamente
Tirosina	TYR	Y	2	Carregado Negativamente
Treonina	THR	T	1	Carregado Negativamente
Triptofano	TRP	W	2	Carregado Negativamente
Valina	VAL	V	1	Não-Polar

Adaptado de Lehninger, Nelson e Cox (2005).

Segundo Lesk (2013), existem dois tipos de arranjos que são considerados os principais elementos estruturais contidos na conformação de proteínas: hélices Alfa (α-*helices*) (PAULING; COREY; BRANSON, 1951) e folhas Beta (β-*sheets*) (PAULING; COREY, 1951). As hélices, assim chamadas por se enovelarem em formato helicoidal, possuem ligações de hidrogênio voltadas para a parte interna do cilindro (RICHARDSON, 1981), enquanto que as folhas ocorrem quando as cadeias polipeptídicas vizinhas se posicionam paralelamente (fitas), permitindo o estabelecimento de ligações de hidrogênio entre resíduos adjacentes em um mesmo plano (PAULING; COREY, 1951). Ambas estruturas (hélices e folhas) possuem ligações mais estáveis, por isso chamadas de estruturas regulares. Além delas, existem dois tipos de arranjos chamados alças (*coils*) e voltas (*turns*). Devido ao seu caráter flexível, estas estruturas são consideras irregulares. Elas ocorrem entre hélices e folhas, tendo um importante papel na conformação/empacotamento global da proteína (BRANDEN; TOOZE, 1999; LESK, 2013). A Figura 2.4 ilustra a estrutura secundária da proteína de ID 2P5K no PDB, onde as hélices são representadas em vermelho, folhas em azul, e voltas/alças em preto.

Métodos de predição e atribuição de estrutura secundária costumam representar

Figura 2.4: Estrutura secundária. Proteína PDB ID: 2P5K.



Fonte: Próprio Autor. Preparadas com *PYMOL* (www.pymol.org).

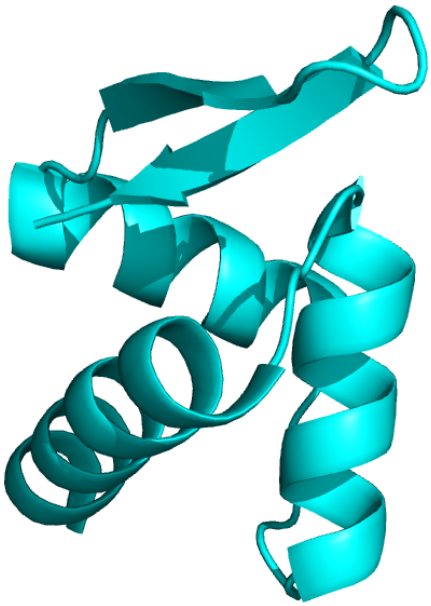
folhas, hélices e regiões irregulares como letras, permitindo assim que a estrutura secundária de uma proteína seja descrita por uma sequência de caracteres. O padrão utilizado varia de acordo com o método. O método DSSP (KABSCH; SANDER, 1983), que realiza a atribuição de estrutura secundária, representa folhas pelas letras E e B, dependendo de seu tipo; também é capaz de identificar três tipos de hélices, atribuindo as letras H, G e I; voltas e alças são representadas pelas letras T e S respectivamente; além da letra C, que é atribuída à regiões desordenadas em que se desconhece o padrão. O método STRIDE (HEINIG; FRISHMAN, 2004), que também é um método de atribuição, utiliza padrão quase idêntico ao DSSP, variando apenas a representação de voltas e alças, onde se utiliza apenas a letra T para ambas. Já o Psipred (JONES, 1999), que é um método de predição, possui um padrão menos detalhado, tratando todos os tipos de folhas pela letra E, as hélices pela letra H, e demais regiões irregulares pela letra C.

2.1.1.3 Estrutura Terciária

Formada pelo arranjo de estruturas secundárias no espaço tridimensional, a estrutura terciária também é chamada de estrutura nativa ou funcional (SCHEEF; FINK, 2005). A conformação descrita neste nível estrutural é obtida por meio de variações de fatores termodinâmicos (e.g., interações covalentes, ligações de hidrogênio, interações hidrofóbicas e eletrostáticas, forças de *van der Waals*) (GIBAS; JAMBECK, 2001; RICHARDSON,

1981; NELSON; COX; LEHNINGER, 2005). A Figura 2.5 ilustra a estrutura terciária da proteína de ID 2P5K no PDB, onde podemos notar padrões de estruturas secundárias, bem como suas respectivas disposições no espaço 3D.

Figura 2.5: Estrutura terciária. Proteína PDB ID: 2P5K.



Fonte: Próprio Autor. Preparadas com *PYMOL* (www.pymol.org).

2.1.1.4 Estrutura Quartenária

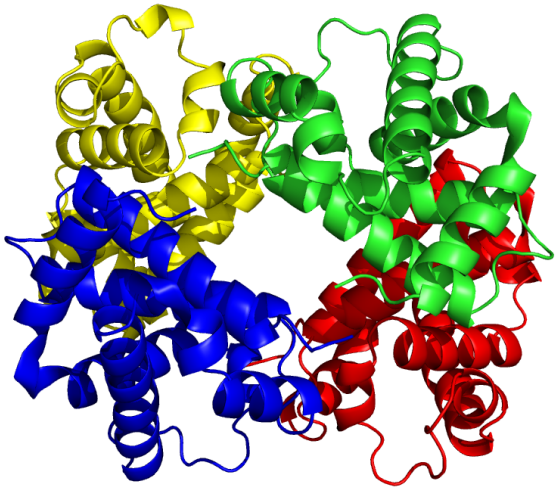
A estrutura quartenária de proteínas descrevem o arranjo tridimensional de diferentes cadeias polipeptídicas, também chamadas de subunidades (LESK, 2013). Essas subunidades podem ser distintas ou conter a mesma estrutura primária. A Figura 2.6 apresenta a estrutura quartenária da proteína de ID 1A3N no PDB, a qual é composta por quatro subunidades distintas.

2.2 Problema de Predição da Estrutura Tridimensional de Proteínas

Conhecer a estrutura tridimensional de uma proteína é um passo fundamental para compreensão de seu papel biológico (SCHEEFF; FINK, 2005; REID, 1999; DORN et al., 2013a), viabilizando inferir sua função na célula de um organismo, localizar seu sítio ativo e cavidades para atracamento molecular, entre outras informações (LEHNINGER; NELSON; COX, 2005; SCHEEF; FINK, 2005; LESK, 2010).

Os principais meios experimentais de se determinar a estrutura tridimensional de

Figura 2.6: Estrutura quartenária. Proteína PDB ID: 1A3N.



Fonte: Próprio Autor. Preparadas com *PYMOL* (www.pymol.org).

uma proteína são cristalografia de raios-x (MCREE, 1999) e Ressonância Magnética Nuclear (NMR - *Nuclear Magnetic Resonance*) (CAVANAGH et al., 1995). Apesar de sua importância, a obtenção dessas estruturas esbarram em obstáculos tais como o alto custo e o tempo despendido no processo (DORN et al., 2014b), além de algumas limitações em relação ao tamanho da proteína (WOOLEY; YE, 2007). Dessa forma, a obtenção da estrutura por meio da predição, apesar de não ter ainda a acurácia desejada, tem se mostrado muito promissora (RANGWALA; KARYPIS, 2010; VALENCIA; PAZOS, 2002). No decorrer dos últimos anos, inúmeras meta-heurísticas estão sendo propostas na tentativa de obter soluções aproximadas para o problema (DORN et al., 2014a), sendo que muitos trabalhos incorporam algum tipo de conhecimento prévio contido em bases de dados, de modo a melhorar a qualidade destas soluções (BORGUESAN et al., 2015b; CORRÊA et al., 2016).

2.2.1 Bases de Dados

Para que as abordagens atinjam resultados de qualidade utilizando conhecimento prévio, tais bases de dado devem ser exploradas de maneiras eficientes (SCHEEF; FINK, 2005; BORGUESAN et al., 2015b). Devido às dificuldades de se determinar experimentalmente a estrutura tridimensional de proteínas, alguns casos acabam por ter sua qualidade/grau de certeza inferior às outras, demandando análises e filtragens dos dados quanto à sua precisão, e até mesmo a utilização de técnicas mais avançadas de descoberta de conhecimento (HOVMOLLER; OHLSON, 2002; BORGUESAN et al., 2015b).

Uma base de conhecimento bastante utilizada é o conjunto de fragmentos do *BAKER-ROSETTASERVER* [1] (KIM; CHIVIAN; BAKER, 2004), utilizada pelo método *Rosetta* (ROHL et al., 2004). Esta base possui fragmentos do tamanho de 3 e de 9 resíduos de aminoácidos. Estes fragmentos representam todas as possibilidades de regiões subsequentes na proteína (no caso em questão, regiões de 3 e 9 aminoácidos). Para cada posição de fragmento da proteína, são disponibilizadas 200 ocorrências distintas. Essas ocorrências são informações de proteínas com a estrutura já conhecida e que possuem aspectos semelhantes à proteína a ser predita (GRONT et al., 2011). Para cada proteína-alvo a ser predita, uma biblioteca de dados é gerada, com base na semelhança entre a sequência de resíduos de aminoácidos e na estrutura secundária predita (utiliza para predição o Psipred conjunto com as técnicas SAM-T99 (KARPLUS et al., 2001) e JUFO (MEILER et al., 2001)). Inicialmente, a base de estruturas conhecidas é filtrada, excluindo-se modelos de baixa resolução (mais de 2,5 *ângströms* (Å)) e que tenham estrutura primária similar à estrutura a ser predita (mais de 50% de similaridade). Em seguida, refina-se a base filtrada, realizando pequenas perturbações em seus valores no intuito de minimizar sobreposições estéricas, mas mantendo-a dentro dos limites permitidos pela obtenção via raios-x. Uma vez que a base foi processada, todas as estruturas são comparadas com todas as janelas de tamanho 3 e 9 da sequência a ser predita. Baseado-se na similaridade das estruturas primária e secundária, os valores ocorridos na base são ranqueados, onde por fim selecionam-se os 200 primeiros fragmentos de cada tamanho (3 e 9).

Existem também bases que trabalham com valores individuais ao invés de fragmentos maiores. A APL (*Angle Probability List*) (BORGUESAN et al., 2015b), disponibilizada pelo NIAS-*Server* (*Neighbors Influence of Amino Acids and Secondary Structures - Server*) [2] é um exemplo disso. Esta base contém a preferência conformacional de resíduos de aminoácidos de uma proteína-alvo, considerando ambas estruturas primárias e secundárias (BORGUESAN; INOSTROZA-PONTA; DORN, 2017). Diferentemente dos fragmentos do *Rosetta*, esta base possui maior quantidade de informação, agindo com uma lista de pares ocorrência/probabilidade de cada combinação de aminoácido e estrutura secundária. Apesar de informar ocorrência de tamanho 1, a APL possui variações que consideram a influência dos aminoácidos vizinhos, permitindo ao usuário decidir qual a configuração melhor se encaixa ao seu problema (BORGUESAN; INOSTROZA-PONTA; DORN, 2017).

É importante ressaltar que não existe forma única de se utilizar essas bases de co-

[1] <http://robetta.bakerlab.org/>
[2] <http://sbcb.inf.ufrgs.br/npas>

nhecimento, entretanto, esse conhecimento é comumente utilizado na inicialização dos algoritmos, aplicando algum processo estocástico dentro das bases para se montar as soluções ao invés de inicializá-las de modo totalmente aleatório, reduzindo assim o espaço de busca (DORN et al., 2013b; CORRÊA et al., 2016).

2.2.2 Representação Computacional

Devido ao fato de que a modelagem computacional do problema PSP caracterizar-se pela busca de uma conformação de menor valor na função objetivo (função de energia) (ANFINSEN, 1973), a predição da estrutura tridimensional de proteínas pode ser modelada como um problema de otimização (LEUNG; WANG, 2001). Representar computacionalmente a estrutura tridimensional de uma proteína de modo a ser otimizada por meta-heurísticas não é uma tarefa simples. Modelos com maior nível de detalhamento descrevem não somente a informação estrutural do polipeptídeo como também interações com seu meio (e.g., ambiente de solvatação) (STILL et al., 1990). Entretanto, quanto mais detalhes e informação, maior a complexidade computacional do modelo. Mesmo em modelos onde se considera apenas a informação estrutural da proteína (solvatação implícita), trabalhar com um alto nível de detalhamento ainda representa um custo computacional considerável (CORRÊA; DORN, 2016). No modelo *all-atom*, todos os átomos que constituem a proteína são representados como pontos em um plano cartesiano no espaço tridimensional, resultando em 3*n* dimensões, sendo *n* o número total de átomos (SCHEEF; FINK, 2005). Ao considerarmos o problema PSP, apesar do alto nível de discriminação, estes modelos podem ser preteridos uma vez que modelos com menor dimensionalidade possuem capacidade satisfatória de representação estrutural.

Como alternativa, muitos trabalhos utilizam a representação por ângulos diedrais (ϕ, ψ, ω e χ). Este modelo parte do princípio de que o comprimento das ligações atômicas são quase constantes, tornando possível a reconstrução do modelo *all-atom* a partir destes ângulos. Isso resulta em uma menor complexidade, uma vez que a dimensionalidade deste modelo não está mais relacionada ao total de átomos da proteína, e sim ao número de ângulos em cada resíduo de aminoácidos (NEUMAIER, 1997). A Equação 2.1 descreve a dimensionalidade deste modelo:

$$dimensionalidade : \sum_{i=1}^{r} 3 + \|\chi_i\| \tag{2.1}$$

sendo *r* o total de resíduos de aminoácidos e $\|\chi_i\|$ o total de ângulos χ do *i*-ésimo resíduo. Há ainda modelos que não utilizam os ângulos χ, resultando em 3*r* dimensões.

Nesta dissertação foi utilizado o pacote de rotinas de modelagem molecular *PyRosetta* [3], que é baseado no método *Rosetta*. Este pacote possui implementações de modelos de representação de proteínas, rotinas de atribuição de estrutura secundária, cálculo de métricas de similaridade entre estruturas, funções de energia, entre outras inúmeras utilidades (CHAUDHURY; LYSKOV; GRAY, 2010). Como modelo de representação estrutural, utilizou-se a opção *centroid*, que trabalha com os ângulos ϕ, ψ e ω. Assim, um polipeptídeo é representado por um vetor $\vec{P}$ de ângulos conforme descrito pela Equação 2.2:

$$\vec{P} = [\vec{r}_1, \vec{r}_2, \cdots, \vec{r}_n] \tag{2.2}$$

onde $\vec{r}_n$ também é um vetor composto pelos ângulos ϕ, ψ e ω do *i*-ésimo resíduo. Nota-se então que o problema PSP pode ser tratado como um problema de minimização de uma função de energia (função objetivo), onde cada variável (ângulos diedrais) do nosso vetor solução está contida no intervalo [-180°, 180°]. É válido lembrar que a representação estrutural por ângulos consiste apenas em uma abstração do mesmo, sendo que, para o cálculo da energia na função objetivo, tal modelo é traduzido para a representação cartesiana, onde cada átomo é reposicionado respeitando os valores dos ângulos diedrais (NEUMAIER, 1997).

2.2.3 Função Objetivo

Como já apresentado, o *PyRosetta* é um pacote de rotinas de modelagem molecular e, entre várias tarefas automatizadas, possui a implementação de funções de energia. Dentre elas, esta dissertação utilizou a função *Score3*, que consiste em uma função compatível ao modelo de representação estrutural definido (*centroid*) e que possui maior robustez na descrição de estruturas de proteínas, uma vez que as outras funções são aplicáveis a propósitos específicos (e.g., detectar interferências estéricas) ou são compatíveis apenas com outros modelos de representação. De acordo com a documentação do *Rosetta* [4], a *Score3* é uma função formada por dez termos ponderados. A Tabela 2.2 apresenta estes termos:

[3] <http://www.pyrosetta.org/>
[4] <https://www.rosettacommons.org/docs/>

Tabela 2.2: Termos que compõem a função de energia *Score3*.

Termo	Descrição	Referência
env	Energia de solvatação	Rohl et al. (2004)
pair	Interações eletrostáticas e de ligações dissulfeto	Rohl et al. (2004)
cbeta	Energia de solvatação	Rohl et al. (2004)
vdw	Forças de *van der Waals* - repulsão estérica	Rohl et al. (2004)
rg	Favorece o empacotamento	Rohl et al. (2004)
cenpack	Favorece o empacotamento	Rohl et al. (2004)
hs_pair	Empacotamento entre hélices e fitas	Rohl et al. (2004)
ss_pair	Ligação de hidrogênio de folhas	Rohl et al. (2004)
rsigma	Descreve o posicionamento de pares de fitas	Shmygelska e Levitt (2009)
sheet	Favorece a formação de folhas	Rohl et al. (2004)

Fonte: Adaptado de *Rosetta Commons - User guide*[5].

Corrêa et al. (2016) ainda propõem um termo adicional, chamado Reforço de Estrutura Secundária (RES). Este termo avalia cada um dos resíduos de aminoácidos individualmente, comparando a estrutura secundária da proteína analisada com a estrutura secundária desejada (informada como parâmetro). Sua pontuação consiste no somatório dos reforços (θ) de cada resíduo de aminoácido sendo que, -θ é somado quando houver correspondência de estrutura secundária e, caso contrário, soma-se θ. Assim, esse termo busca auxiliar o empacotamento global de proteínas.

Dessa forma, por se tratar de um problema de minimização, este termo é capaz de quantificar a similaridade entre a estrutura secundária do modelo analisado em relação à desejada. A Equação 2.3 apresenta a função objetivo final utilizada neste trabalho:

$$Energia = \textit{Score3} + RES \tag{2.3}$$

[5]<https://goo.gl/1DZryX> - Acessado em: 29-12-2017

2.3 Meta-heurísticas Aplicadas ao Problema PSP

Devido ao fato de que a modelagem computacional do problema PSP caracterizar-se pela busca de uma conformação de menor valor na função objetivo (função de energia), a predição da estrutura tridimensional de proteínas pode ser modelada como um problema de otimização (LIWO et al., 1999). Várias técnicas tem sido empregadas na tentativa de resolver o problema PSP. Entretanto, devido a sua complexidade, ainda existem grandes desafios no sentido de encontrar soluções ótimas ou quase ótimas para o problema (DORN et al., 2014b). Uma alternativa consiste na utilização de meta-heurísticas (TANTAR et al., 2007a; ABUAL-RUB et al., 2012), que são abordagens de propósito geral que não necessitam de conhecimentos específicos do problema (MANIEZZO; STÜTZLE; VOSS, 2009) e que empregam otimização estocástica (LUKE, 2013a). Devido a esse fator estocástico, meta-heurísticas não possuem garantias de que encontrarão a solução ótima, todavia, espera-se que seja possível encontrar soluções satisfatórias em tempo razoável (TALBI, 2009; YANG, 2010).

De acordo com Boussaïd, Lepagnot e Siarry (2013) e Birattari et al. (2001), meta-heurísticas podem ser classificadas entre abordagens baseadas em solução única ou baseadas em população de soluções. Como exemplos da classe baseada em solução única temos o Recozimento Simulado (KIRKPATRICK; GELATT; VECCHI, 1983), Busca Tabu (GLOVER, 1986), busca em Vizinhança Variável (MLADENOVIC, 1995), já da classe baseada em população de soluções temos a Optimização por Enxame de Partículas (KENNEDY; EBERHART, 1995), Colônia de Formigas (DORIGO; MANIEZZO; COLORNI, 1996), AG (HOLLAND, 1975a), entre outros.

Ao longo dos últimos anos, diferentes meta-heurísticas foram aplicadas no problema PSP. No trabalho de Fonseca, Paluszewski e Winter (2010), uma variação do algoritmo *Bee Colony Optimization* (KARABOGA; BASTURK, 2007), que baseia-se no comportamento de forrageamento das abelhas, foi aplicado no problema PSP, pela primeira vez considerando proteínas com tamanho superior a 50 resíduos de aminoácidos. Saleh, Olson e Shehu (2013) propuseram um Algoritmo Memético (AM) composto por duas abordagens evolutivas, baseadas em fragmentos estruturais de aminoácidos, para tratar o problema dos múltiplos mínimos locais presentes na função de energia. Elofsson, Grand e Eisenberg (1995) apresenta um AG com busca local no espaço conformacional de ângulos diedros de uma proteína, buscando predizer a sua estrutura nativa. Tantar et al. (2007b) apresentam uma versão híbrida de um AG paralela combinado com uma busca

local de descida do gradiente para predizer a estrutura de polipeptídeos. De modo geral, a utilização de técnicas baseadas em AG tem sido amplamente empregadas no problema PSP, alcançando boas taxas de sucesso (UNGER, 2004).

2.3.1 Algoritmos Genéticos

O AG, proposto por Holland (1975b), consiste em uma técnica inspirada pela Teoria da Evolução, com foco na sobrevivência dos mais aptos. Trazendo isto para uma visão computacional, temos que cada possível solução representa um individuo que, em conjunto, representam uma população. Cada indivíduo possui um vetor de valores, chamados de material genético, onde cada valor representa um alelo de DNA. O AG propõem que os indivíduos sejam submetidos a operadores genéticos (Seleção, Recombinação, Mutação), buscando gerar-se indivíduos aprimorados. Para tal, o *fitness* (valor de aptidão da solução na função objetivo) de cada indivíduo é levado em conta no operador de Seleção, de modo que os melhores tem maior chances de serem selecionados. Inicialmente os indivíduos são ordenados pelo seus respectivos valores de *fitness* e, em seguida, pares dos mesmos são selecionados para participar da Recombinação e da Mutação. De acordo com o esquema de seleção por Roleta, cada indivíduo possui uma probabilidade de seleção proporcional ao seu valor de *fitness* (WHITLEY, 1994; GOLBERG, 1989; HOLLAND, 1975b). A Equação 2.4 descreve isso:

$$\text{Probabilidade\_seleção}(i) : \frac{Fitness(i)}{\sum_{j=1}^{n} Fitness(j)} \tag{2.4}$$

onde *i* representa o $i_{\text{-ésimo}}$ indivíduo e *n* representa o tamanho da população. Uma vez selecionados, o par de indivíduos (chamados de pais) passam pela recombinação (*crossover*), onde dois novos indivíduos, chamados de prole (*offspring*), são gerados a partir da permutação de material genéticos dos pais. Para tal, um ponto de corte *k* é aleatoriamente definido entre 1 e $l - 1$, sendo *l* o número de alelos do indivíduo. Então, a primeira prole recebe material genético de 1 a *k* do primeiro pai, e de *k + 1* até *l* do segundo pai, sendo a segunda prole gerada pelo inverso disso. Em seguida, sob certa probabilidade *m*, estas proles podem passar por um processo de mutação, onde um alelo é aleatoriamente selecionado e alterado. Todo esse processo se repete até que se gerem *n* novos indivíduos, que irão formar a nova população. Este ciclo, chamado de geração, se repete até que um critério de parada seja alcançado (SRINIVAS; PATNAIK, 1994; LUKE, 2013b).

Uma pequena variação consiste em se copiar o melhor indivíduo (Elite) para a população seguinte, garantindo assim que a melhor solução encontrada pelo AG ao longo de sua execução será sempre mantida (DEB et al., 2002). O Algoritmo 1 sintetiza este processo:

Algoritmo 1 AG canônico - *m* representa a probabilidade de mutação e *n* representa o tamanho da população. Adaptado de Alixandre e Dorn (2017).

Entrada: *m*, *n*
Saída: o indivíduo de melhor *fitness*

```
Inicializa a população
enquanto não atingir o critério de parada faça
    Ordena a população
    Copia o melhor indivíduo da população atual para a próxima (Elitismo)
    para i = 0 até n − 1 faça
        Seleciona dois indivíduos de acordo com suas probabilidades no
        esquema de seleção por Roleta
        Cria uma novo indivíduo (prole) através do operador de recombinação
        Aplica o operador de mutação sob uma probabilidade m
        Envia a prole para a população da próxima geração
    fim para
fim enquanto
```

Apesar de ser bastante utilizado, o AG pode enfrentar algumas dificuldades em problemas muito complexos. Um dos principais obstáculos encontrados é a convergência prematura, que é observada quando os indivíduos convergem para uma única solução não satisfatória. Isso ocorre principalmente em decorrência da perda de diversidade na população, deixando o algoritmo preso em ótimos locais (WHITLEY, 1994; MAULIK; BANDYOPADHYAY, 2000). Para evitar tal situação, é importante que o método seja capaz de combinar de modo equilibrado dois tipos comportamento: *exploration*, que são ações que visam explorar o espaço de busca; e *exploitation*, que são ações de refinamento, onde os indivíduos sofrem alterações pontuais (i.e., populações com maior nível de similaridade tendem a resultar em novas populações com pouca variação). Ambos os comportamentos são necessários, uma vez que o *exploration* pode evitar ótimos locais, mas em certo ponto o algoritmo precisa convergir para uma solução satisfatória, caracterizando ações de *exploitation* (ČREPINŠEK; LIU; MERNIK, 2013). Uma maneira de equilibrá-los é através do controle de diversidade, o qual será abordado na próxima seção.

2.3.2 Diversidade

Alguns problemas de otimização, como o PSP, apresentam espaços de busca dinâmicos ou com múltiplos ótimos locais, que podem deixar algoritmos baseados em população presos em certas regiões, levando-o a uma convergência prematura. Como forma de se evitar isto, muitos trabalhos buscam realizar a manutenção da diversidade dos indivíduos da população (WHITLEY et al., 1989; LOZANO; HERRERA; CANO, 2008). Diversidade é um fator diretamente relacionado ao desempenho da otimização. Populações pouco diversificadas aceleram a difusão do material genético de suas melhores soluções entre os demais indivíduos, resultando em um processo de convergência. Apesar disso ser algo natural e desejado para um algoritmo de otimização, algumas vezes esse processo ocorre antes do esperado, convergindo em soluções pouco satisfatórias. Dessa forma, manter uma certo nível de diversidade pode trazer a possibilidade da população explorar novos locais no espaço de busca e consequentemente melhorar os resultados encontrados (URSEM, 2002; HUANG; CHEN, 2001).

Uma métrica de diversidade adequada pode ter grande impacto na tomada de decisão durante a execução de uma meta-heurística (TAKAHASHI, 2016). Nesse trabalho utilizou-se a métrica de diversidade proposta por Alixandre e Dorn (2017), que utiliza como referência a melhor solução encontrada até o momento de seu cálculo. A Equação 2.5 descreve esta métrica:

$$Diversidade(Pop) : \frac{\sum_{i=2}^{N} Dist(Pop_1, Pop_i)}{N-1} \tag{2.5}$$

onde *Pop* é a população final ordenada pela função objetivo, *N* é o tamanho da população, *Dist* é a Distância utilizada no problema (e.g., Distância Euclidiana) e *1* é o índice da primeira (e melhor) solução da população.

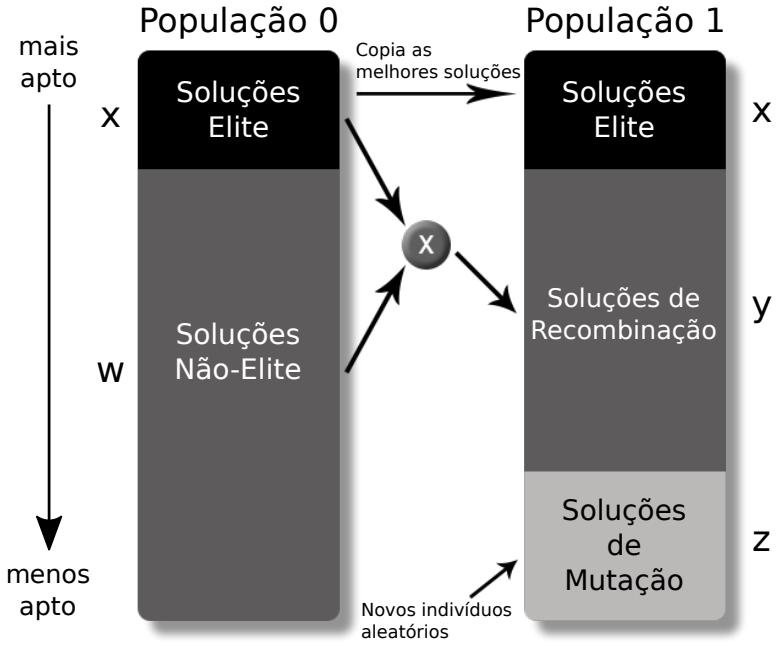
Diferentes formas de manter a diversidade foram propostas nas últimas décadas, indo desde versões de AG mais elaboradas, como o BRKGA, até modelos distribuídos. As próximas seções irão detalhar esses casos.

2.3.3 BRKGA

Proposto por Gonçalves e Resende (2011), o BRKGA é uma técnica que possui um mecanismo de codificação/decodificação que o torna aplicável à qualquer problema.

Enquanto o AG canônico foi originalmente desenvolvido para casos binários (WHITLEY, 1994), o BRKGA codifica o problema em um vetor de chaves dentro do intervalo $[0, 1)$. Essa normalização o torna independente da aplicação e, quando necessário, o operador de decodificação é aplicado à solução encontrada, trazendo de volta ao domínio do problema (BEAN, 1994). Outra particularidade do BRKGA é sua forma de estruturar a população de acordo com o valor de *fitness*. Após criar a população inicial e ordená-la, os indivíduos são divididos em dois grupos, Elite e Não-Elite, de tamanhos x e w respectivamente, sendo $x + w = n$. Para gerar a próxima geração, o grupo de Elite é inteiramente copiado para a nova população. Em seguida, y novos indivíduos são gerados pelo operador de recombinação. Por fim, fazendo o papel de mutação, z novos indivíduos são criados aleatoriamente e inseridos na nova população, sendo $y + z = w$ (GONÇALVES; RESENDE, 2011; ALIXANDRE; DORN, 2017). A Figura 2.7 apresenta esse esquema.

Figura 2.7: Estrutura da população do BRKGA. Transição da população inicial para a geração 1.

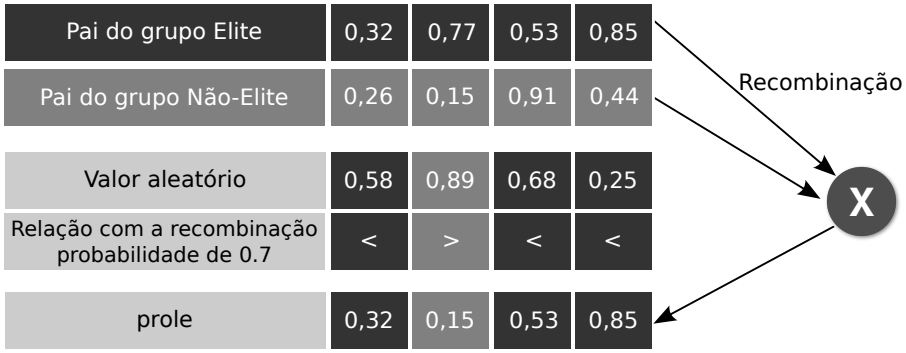


Fonte: Adaptado de Alixandre e Dorn (2017).

O modo com que o BRKGA realiza a seleção de indivíduos para a recombinação também é diferente do AG canônico. Visando explorar a diversidade contida na população, o BRKGA garante que os indivíduos selecionados sejam de grupos distintos. O primeiro pai é selecionado do grupo Elite, enquanto o segundo é escolhido entre os grupos Não-Elite e Mutação, ambos de forma aleatória. Além disso, a recombinação resulta em apenas uma prole, diferente do AG canônico que resulta em dois. Durante a recombinação, valores aleatórios entre 0 e 1 são definidos para cada alelo do vetor de chaves do

primeiro pai e, caso seja menor ou igual à uma probabilidade *p*, o valor do alelo é transmitido à prole. Caso contrário, o segundo pai é quem transmite seu material genético. Este processo ocorre de maneira uniforme por todo o vetor de chaves (GONÇALVES; RESENDE, 2011). Os autores ainda propõem que essa probabilidade *p* seja entre 50% e 70% (0.5 e 0.7), uma vez que o primeiro pai possui garantidamente um *fitness* melhor que o segundo. A Figura 2.8 apresenta um exemplo onde, com uma probabilidade *p* igual a 70%, a prole recebe o material genético do primeiro pai nos alelos 1, 3 e 4, e o restante do segundo pai. O Algoritmo 2 sintetiza todo o processo.

Figura 2.8: Recombinação uniforme parametrizada.



Fonte: Adaptado de Alixandre e Dorn (2017).

Algoritmo 2 BRKGA - *x*, *y*, *z* representam o tamanho dos grupos Elite, Não-Elite e Mutação respectivamente, *p* representa a probabilidade do pai 1 repassar seu material genético no operador de recombinação. Adaptado de (ALIXANDRE; DORN, 2017).

Entrada: *x*, *y*, *z*, *p*
Saída: o indivíduo de melhor *fitness*

```
Inicializa a população
enquanto não atingir o critério de parada faça
    Ordena a população
    Copia os x melhores indivíduos da população atual para a próxima
    para i = 0 até y faça
        Selecionado aleatoriamente um indivíduo do grupo Elite
        Selecionado aleatoriamente um indivíduo do grupo Não-Elite ou Mutação
        Cria uma novo indivíduo (prole) através do operador de
        recombinação com probabilidade p
        Envia a prole para a população da próxima geração
    fim para
    Cria z novos indivíduos aleatórios e envia-os para a população da próxima geração
fim enquanto
```

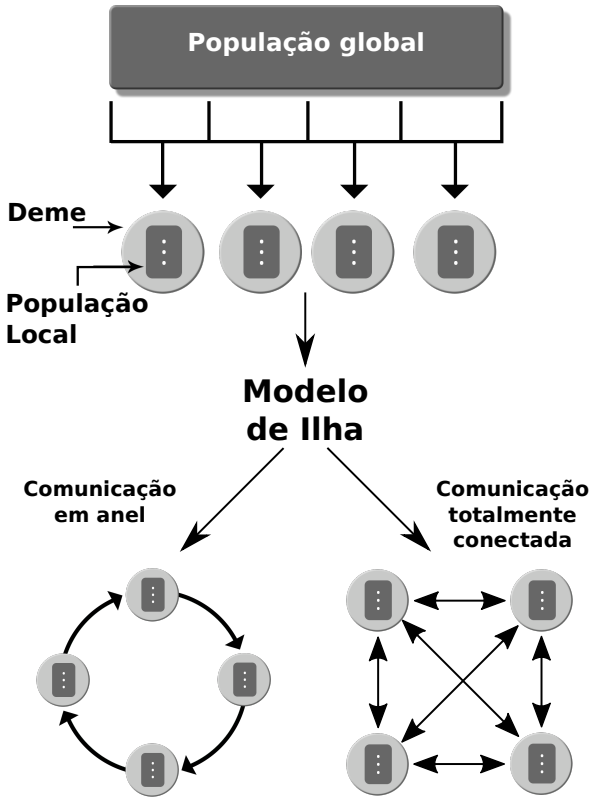
2.3.4 Algoritmos Genéticos Distribuídos

Os AGDs foram inicialmente estudados por Pettey, Leuze e Grefenstette (1987) e Cohoon et al. (1987), se tornando também muito utilizado em diversos trabalhos ao longo dos anos (TANESE, 1989; LORASCHI et al., 1995; STARKWEATHER; WHITLEY; MATHIAS, 1990; COHOON et al., 1991; WHITLEY; STARKWEATHER, 1990). A ideia principal consiste em dividir a população global em subpopulações (também chamadas de *demes*) que, de forma independente e isolada (Arquitetura de Ilha), executam algum AG. Entretanto, periodicamente ocorre uma troca de indivíduos entre as *demes*, caracterizando assim o processo de migração (BELDING, 1995; CANTÚ-PAZ, 1998). Dessa forma, ao passo que se divide a população em um maior número de *demes*, o AGD é capaz de desacelerar a convergência, além de proporcionar o aumento da diversidade ao realizar a migração de indivíduos com material genético diferente ao da subpopulação de destino (WHITLEY, 1994; TANESE, 1989; BELDING, 1995).

Um fator essencial a se definir ao utilizar um AGD é a topologia de comunicação, sendo anel, estrela e totalmente conectada as opções comumente utilizadas (LUQUE; ALBA, 2011). A topologia descreve o sentido da comunicação entre as *demes*. Neste trabalho, as técnicas abordadas utilizam as topologias anel e totalmente conectada, que são ilustradas na Figura 2.9.

Outro fator igualmente importante é a política de migração, que define quantos indivíduos irão migrar (*migration size - ms*) e com que frequência isso irá ocorrer (*migration rate - mr*). Além disso, a política também estabelece regras para selecionar aqueles que irão migrar e aqueles que serão substituídos, sendo o *envia os melhores/substitui os piores* a versão canônica, onde envia-se sempre as *ms* melhores soluções, substituindo as *ms* piores na *deme* de destino (ALBA; TROYA, 1999; LUKE, 2013b). Os Algoritmos 3, 4 e 5 apresentam todos os componentes de um AGD. Neste esquema, um processo principal cria k instâncias de algum AG (e.g., AG canônico, BRKGA) que serão executados de forma independente e, a cada *mr* gerações realiza uma etapa de migração de acordo com a política utilizada (e.g., Canônica *envia os melhores/substitui os piores*). Em futuras referências, chamaremos a política *envia os melhores/substitui os piores* de Política 1.

Figura 2.9: AGD - Arquitetura de Ilha e topologias de comunicação.



Fonte: Adaptado de Alixandre e Dorn (2017).

Algoritmo 3 Processo principal - *k* representa o número de *demes*.

Entrada: *k*
Saída: o melhor entre todos os *k* resultados *// Ótimo global*
1: Cria *k* instâncias do Algoritmo 4
2: Aguarda até que todas as *k* instâncias tenham sido finalizadas
3: Recolhe e ordena os indivíduos retornados *// Ótimo local*

Algoritmo 4 AGD_i

Saída: o indivíduo de melhor *fitness* para o processo principal
1: Inicializa a população
2: **enquanto** não atingir o critério de parada **faça**
3: Realiza a seleção, recombinação e/ou mutação
// De acordo com as características do AG usado (e.g., Algoritmo 1 ou 2)
4: **se** está no momento de migrar **então**
// Baseado na taxa de migração
5: Seleciona e troca indivíduos de acordo com a topologia
e a política de migração usada *// e.g., Algoritmo5*
6: **fim se**
7: **fim enquanto**

Algoritmo 5 Política 1 - Canônica - *envia os melhores/substitui os piores* (ALBA; TROYA, 1999)

Entrada: *ms* *// tamanho dos migrantes*

1: Envia os *ms* melhores indivíduos de acordo com a topologia
2: Aguarde receber novos indivíduos
3: Substitua os *ms* piores indivíduos por aqueles recebidos

2.4 Resumo do Capítulo

Neste capítulo foram apresentados os conceitos fundamentais para o entendimento do problema PSP, abordando a definição de proteína, seus níveis de representação estruturais, as características do problema de predição, bem como sua modelagem computacional. O capítulo também introduziu o conceito de meta-heurística, além dos métodos AG, BRKGA e AGD. No próximo capítulo serão apresentados os trabalhos que se relacionam à abordagem proposta e que se aplicam ao problema PSP.

3 TRABALHOS RELACIONADOS

Ao longo dos últimos anos, diversos métodos foram propostos para o problema PSP. Diversas abordagens, foram estudadas e aprimoradas de modo a se explorar certas características do problema visando a obtenção de melhores resultados. Neste capítulo serão apresentadas abordagens envolvendo o controle e manutenção de diversidade em AGD aplicados em funções de teste (funções genéricas sem área de aplicação explícita), bem como abordagens de otimização aplicadas no problema PSP. Tais técnicas foram igualmente importantes no desenvolvimento deste trabalho uma vez que se identificou a necessidade de explorar as características do AGD em um ambiente sem viés (i.e., função de energia com certa imprecisão) para que em seguida possa ser empregado ao domínio de predição da estrutura de proteínas com maior rigor metodológico.

3.1 Meta-heurísticas Distribuídas Baseadas em População e sua Utilização em Problemas de Otimização

Nesta seção serão apresentados quatro abordagens para manutenção da diversidade em AGD. No trabalho de Alixandre e Dorn (2017), um política de migração baseada no *fitness* é proposta, enquanto que Denzinger e Kidney (2003) propõem realizar a escolha dos migrantes considerando tanto o *fitness* quanto a similaridade entre indivíduos, de modo a ranqueá-los. Já Power, Ryan e Azad (2005) e Araujo e Merelo (2011), apesar de também levarem em conta o *fitness* na tomada de decisão, apresentam políticas que focam principalmente na noção de distância entre as populações com os indivíduos que sejam mais similares a elas.

3.1.1 Migração Baseada em *Fitness*

No trabalho de Alixandre e Dorn (2017) uma política de migração foi proposta juntamente com uma versão distribuída do BRKGA (GONÇALVES; RESENDE, 2011). A abordagem utiliza a topologia de Anel com comunicação unidirecional. No processo de migração, a política proposta baseia-se na própria distribuição da população, que se dá de forma estruturada em função do *fitness*. Como no BRKGA a população é composta por três grupos (Elite, Não-Elite e Mutação), a seleção dos migrantes obrigatoriamente

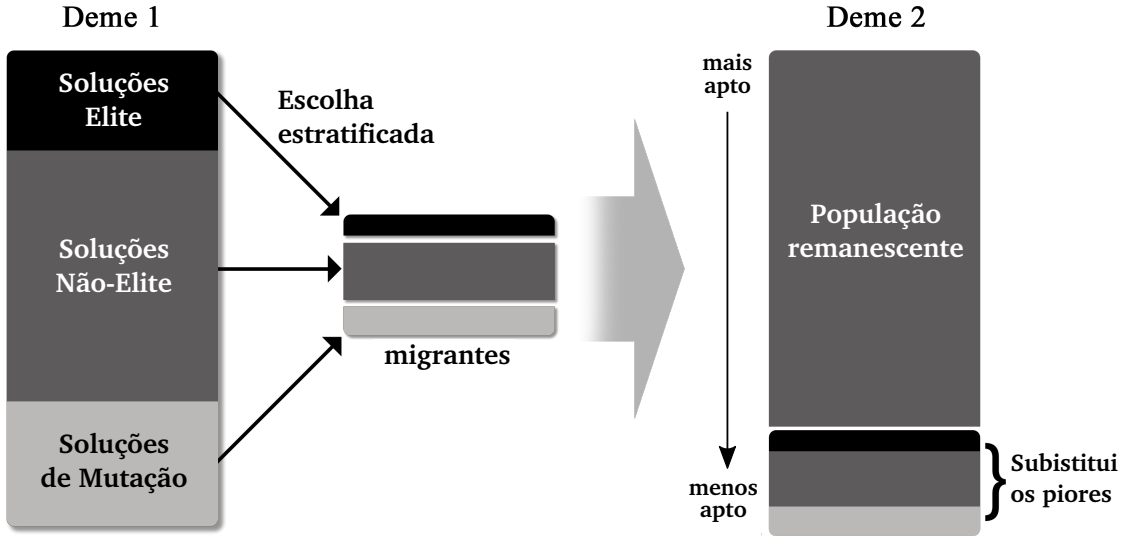
seleciona indivíduos de todos os grupos. Para manter o equilíbrio, esta escolha ocorre de forma estratificada, ou seja, cada grupo contribui com uma quantidade de indivíduos para a migração proporcional a sua própria representatividade na população. O Algoritmo 6 detalha esse processo:

Algoritmo 6 Política 2 - (ALIXANDRE; DORN, 2017). *ms* representa o tamanho da migração, *m1*, *m2* e *m3* representam a quantidade de migrantes de cada grupo (Elite, Não-Elite e Mutação) respectivamente.

Entrada: *ms*, *m1*, *m2*, *m3*

1: Escolha *m1* indivíduos aleatórios do grupo Elite
2: Escolha *m2* indivíduos aleatórios do grupo Não-Elite
3: Escolha *m3* indivíduos aleatórios do grupo Mutação
4: Encaminha-se os indivíduos escolhidos para migração
5: Espera-se até receber os migrantes de outra *deme*
6: Substitua os *ms* piores indivíduos pelos recebidos

Figura 3.1: Política de migração baseada em *fitness*.



Fonte: Imagem adaptada de Alixandre e Dorn (2017).

Esta política de migração busca diversificar a escolha dos migrantes considerando selecionar indivíduos de diferentes intervalos de *fitness*, todavia a mesma se torna exclusivamente aplicável ao BRKGA, impedindo sua utilização com demais variações.

3.1.2 Migração Mesclando a Similaridade ao Indivíduo Médio

Visando reduzir a convergência prematura, Power, Ryan e Azad (2005) propõem uma política de migração que, apesar de considerar o *fitness* da população, baseia-se prin-

cipalmente na similaridade entre os indivíduos. Para tal, calcula-se o *indivíduo médio*, que consiste no indivíduo que tenha a menor distância média para todos os demais. A Equação 3.1 descreve esse cálculo. Após isso, são selecionados três conjuntos de indivíduos para migração:

- O *indivíduo médio* e o melhor indivíduo.
- Os *m* indivíduos mais distantes e que tenham *fitness* menor que o *indivíduo médio*.
- Os *m* indivíduos mais distantes e que tenham *fitness* maior que o *indivíduo médio*.

sendo $m = \frac{ms}{2} - 1$ e *ms* o tamanho dos migrantes.

Após enviar os migrantes, como detalhado no Algoritmo 7, a política prioritariamente substitui os indivíduos que sejam idênticos a outros contidos na população, começando por aqueles de pior *fitness*. Em seguida, se necessário, os piores indivíduos são substituídos.

$$\text{Distância\_}mdia(ind_i) : \frac{\sum_{j=1}^{n} \text{Distância}(ind_i, ind_j)}{n-1}, \tag{3.1}$$

sendo ind_i o *i*-ésimo indivíduo, *n* o tamanho da população e Distância() a métrica de similaridade adequada ao problema.

Algoritmo 7 Política 3 - (POWER; RYAN; AZAD, 2005). *ms* representa o tamanho da migração.

Entrada: *ms*

1: $m \leftarrow \frac{ms}{2} - 1$
2: Calcula-se o *indivíduo médio* *// Aquele que tem a menor distância Euclidiana média para todos os outros indivíduos*
3: Dentre os indivíduos que tenham o *fitness* maior que o *indivíduo médio*, selecione os *m* indivíduos mais distantes dele
4: Dentre os indivíduos que tenham o *fitness* menor que o *indivíduo médio*, selecione os *m* indivíduos mais distantes dele
5: Envia-se para migração o *indivíduo médio*, os dois conjuntos selecionados baseados no *fitness* e na distância, e o indivíduo com melhor *fitness*
6: Espera-se até receber os migrantes de outra *deme*
7: Primeiramente substitua indivíduos replicados, então substitua os piores

Para garantir um conjunto de migrantes diversificado, esta política busca enviar indivíduos que estejam distantes da média da população, escolhendo parte com *fitness* maior e parte com *fitness* menor que o *indivíduo médio*. Entretanto, esta política possui um custo computacional extremamente sensível ao tamanho da população e à dimensão do problema, uma vez que se faz necessário calcular a distância entre todos os indivíduos para encontrar aquele mais similar aos demais.

3.1.3 Migração Minimizando a Semelhança ao Indivíduo Médio

No trabalho de Araujo e Merelo (2011), os autores propõem uma técnica que tem como objetivo migrar indivíduos que sejam diferentes da população de destino. Para isso, cada *deme* deve calcular o indivíduo médio da população, chamado de *representante*. Após isso, cada *deme* seleciona e envia um dentre os x melhores indivíduos de sua população (chamados de *Elite*), tal que este indivíduo seja o mais diferente do *representante* da população de destino (baseado em alguma métrica de similaridade adequada ao problema). Dessa forma, esta política busca realizar a migração considerando a similaridade dos indivíduos, sem desconsiderar por completo seus respectivos valores de *fitness*. Para se calcular o *representante* de uma população, os autores propõem duas maneiras: (a) a utilização do conceito de *consensus sequence* ou (b) selecionar o melhor indivíduo. Para aplicações futuras usaremos a *consensus sequence*, que consiste em calcular um indivíduo formado pelos valores de maior ocorrência dentro da população. O Algoritmo 8 apresenta esse processo:

Algoritmo 8 Política 4 - (ARAUJO; MERELO, 2011). x representa o tamanho da *Elite*.

Entrada: x

1: Calcula-se o *representante* da população e envia-o para a *deme* anterior
2: Receba o *representante* da *deme* subsequente
3: Dentre os x melhores indivíduos (*Elite*), escolha o que estiver mais distante do *representante* recebido e envie-o para migração
4: Espera-se até receber o migrante da *deme* anterior
5: Substitua o pior indivíduo pelo novo recebido

Os autores ainda destacam que o valor de x tem grande impacto nessa política, sendo que, quanto menor o tamanho da *Elite*, menos opções de indivíduos haverá, aproximando-se do que ocorre na política *envia os melhores/substitui os piores* (resultando no mesmo comportamento para o caso de $X = 1$). Entretanto, à medida que o tamanho da *Elite* aumenta (e.g., x igual ao tamanho da população), a qualidade do *fitness* perderá influência, uma vez que deve-se escolher para migrar aquele indivíduo que for mais diferente da população de destino. Os autores testaram vários valores de x, concluindo que apesar da variação de valores ótimos, em populações de aproximadamente 100 indivíduos, utilizar *Elites* de tamanho 8 foi em média a melhor opção.

3.1.4 Migração por Ranqueamento em *Fitness* + Similaridade

Para garantir uma migração com diversidade suficiente, Denzinger e Kidney (2003) apresentam uma política que refaz o ranqueamento dos indivíduos de uma população, baseando-se no valor de *Qualidade* apresentado pela Equação 3.2. Esta equação tem por objetivo equilibrar a influência do *fitness* e da similaridade dos indivíduos na escolha dos migrantes. Basicamente, dois pesos, *Wfit* e *Wdiv*, são vinculados ao valor de *fitness* e *diversidade* de cada indivíduo respectivamente. Estes dois pesos tem como papel ponderar cada um dos fatores envolvidos, sendo *Wfit* $= 1 -$ *Wdiv*. O valor de *diversidade* consiste na distância deste indivíduo para o melhor de todos dentro da população. Após toda a população ser reordenada, selecionam-se os *ms* melhores indivíduos para migrar. O Algoritmo 9 detalha esse processo:

$$Qualidade(ind) : Wfit \times \frac{fitness(ind)}{fitness(ind_{melhor})} + Wdiv \times \frac{\text{Distância}(ind, ind_{melhor})}{n} \tag{3.2}$$

onde ind_{melhor} representa o indivíduo com o melhor *fitness* na população, *Wfit* e *Wdiv* são respectivamente os pesos associados ao *fitness* e a distância entre o indivíduo e o ind_{melhor}, *Wfit* + *Wdiv* $= 1,0$ e Distância() a métrica de similaridade adequada ao problema.

Algoritmo 9 Política 5 - (DENZINGER; KIDNEY, 2003). *ms* representa o tamanho da migração; *Wfit* e *Wdiv* são os pesos utilizados pela função de *Qualidade*.

Entrada: *Wfit*, *Wdiv*, *ms*

1: Calcula-se e ordena-se a população pelo seu valor de *Qualidade* // (Veja Eq. 3.2)
2: Envia-se os *ms* melhores indivíduos
3: Espera-se até receber os migrantes
4: Substitua os *ms* piores indivíduos pelos recebidos

Os autores realizaram um estudo aprofundado acerca dos valores ideais para *Wfit* e *Wdiv*. Apesar de algumas variações, os valores de 0.7 e 0.3 respectivamente se sobressaíram.

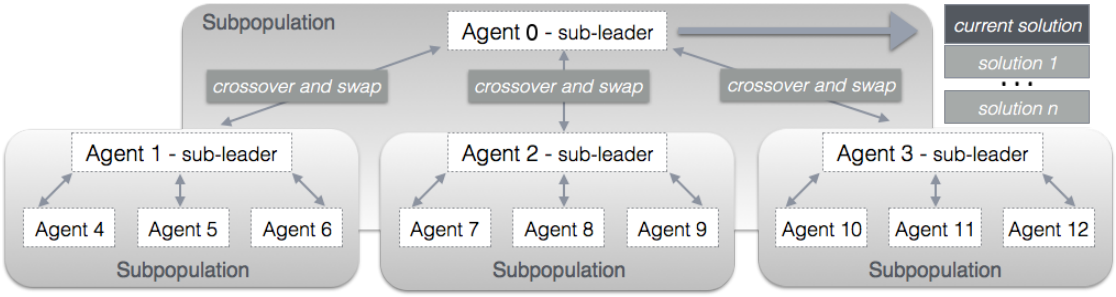
3.2 Métodos Baseados em População e sua Utilização no Problema PSP

Nesta seção serão apresentados dois métodos de primeiros princípios que utilizam informação experimental no problema PSP. O primeiro, proposto por Corrêa et al. (2016), utiliza a APL (BORGUESAN et al., 2015b) como base de conhecimento, enquanto que (ROHL et al., 2004) utilizam fragmentos (KIM; CHIVIAN; BAKER, 2004).

3.2.1 Método de Predição Utilizando APL

Corrêa et al. (2016) propõem uma versão de AM utilizando uma população organizada em uma estrutura de árvore ternária, conforme ilustrado na Figura 3.2. A população é inicializada utilizando informação da APL no intuito de restringir o espaço de busca conformacional e consequentemente partir de soluções com certo nível de qualidade. Contendo treze subpopulações, chamadas de *agentes*, o método estabelece restrições de comunicação entres elas. Cada *agente* se comunica apenas com seu líder no nível superior e com seus três *agentes* do nível inferior. A ideia é que boas soluções sejam enviadas para os níveis superiores à medida que estas sejam melhores do que alguma solução contida no *agente* líder. Além disso, os operadores de recombinação utilizam indivíduos selecionados do *agente* em questão e de seu líder, buscando diversificar o processo. Em seguida, uma técnica de busca local é aplicada ao melhor indivíduo de cada *agente*, visando refinar tais soluções. O Algoritmo 10 apresenta o funcionamento do método, onde, sol_{best} é a melhor solução encontrada ao final da execução e *agente.atual* representa a melhor solução do *agente*.

Figura 3.2: População do AM - Estrutura em Árvore ternária. *Agentes* se comunicam apenas com seus líderes e suas ramificações do nível inferior. A recombinação envolve indivíduos de *agentes* distintos e novos indivíduos com *fitness* melhores são enviados para seus respectivos *agentes* líderes.



Fonte: Corrêa et al. (2016).

Algoritmo 10 AM - Adaptado de Corrêa et al. (2016). sol_{best} é a melhor solução encontrada.

Entrada: número máximo de cálculos de função de energia
Saída: $sol_{best} \leftarrow$ melhor solução encontrada

```
Inicializa-se a população
sol_best ← melhor solução do agente_0
enquanto critério de parada não é satisfeito faça
    para cada agente faça
        par_1 ← solução aleatória do agente sub-líder
        par_2 ← solução aleatória do agente
        se operador de probabilidade ≤ rand(0, 1) então
            agente.atual ← uniformCrossoverSS(par_1, par_2)
        senão
            agente.atual ← crossoverSS(par_1, par_2)
        fim se
    fim para
    para cada agente faça
        agente.atual ← SA(agente.atual)    // Busca Local
    fim para
    Atualiza a população
    se melhor solução do agente_0 < sol_best então
        sol_best ← melhor solução do agente_0
    fim se
    se Não atingiu o limiar de melhoria então
        Reinicializa a população
    fim se
fim enquanto
```

O método possui dois operadores de recombinação: *uniformCrossoverSS* e *crossoverSS*, ambos baseados na informação da estrutura secundária. No *uniformCrossoverSS*, para cada resíduo, é verificado se a estrutura secundária dos indivíduos selecionados são iguais à informada. Caso ambas sejam iguais ou diferentes, realiza-se um sorteio com probabilidade de 60% de se escolher o indivíduo de melhor *fitness* (um número aleatório é definido no intervalo $(0, 1]$, escolhendo-se o indivíduo de melhor *fitness* caso seja menor que 0,6). Caso apenas um dos indivíduos tenha a estrutura secundária igual à informada, este será imediatamente selecionado. Já o *crossoverSS* realiza um processo semelhante, exceto que ao invés de analisar resíduo por resíduo, este age por segmentos de estrutura secundária semelhantes (Ex.: Estrutura:{CHHHCCEE} $\rightarrow$ segmentos:{C; HHH; CC; EE}). Para cada segmento, verifica-se a estrutura secundária do primeiro resíduo de cada indivíduo e compara-se à estrutura informada. Como no *uniformCrossoverSS*, caso ambos sejam iguais ou diferentes, realiza-se um sorteio com maior probabilidade para o indivíduo de melhor *fitness*, caso contrário seleciona-se aquele que tenha estrutura secundária

igual à informada.

Como técnica de busca local, utilizou-se o algoritmo de Recozimento Simulado (KIRKPATRICK; GELATT; VECCHI, 1983) combinado com informação da APL. O método percorre cada resíduo separadamente, com probabilidade de 90% de refiná-los (pertubações locais visando melhorar o indivíduo). Dessa forma, caso um resíduo seja selecionado para refinamento, uma mutação ocorre, alterando-o por outro valor contido na APL e em seguida aplica-se Recozimento Simulado. Por fim, o indivíduo resultante será mantido apenas se houver uma melhora de seu *fitness*. Além disso, o método possui um mecanismo de reinicialização da população. Tal procedimento ocorre se não houver melhora na melhor solução global durante 250 gerações, mantendo-se apenas as melhores soluções de cada *agente* e reinicializando todo o resto.

3.2.2 Método de Predição Baseado em Fragmentos

Utilizado pelo *BAKER-ROSETTASERVER* [1] (KIM; CHIVIAN; BAKER, 2004), o método de predição da estrutura de proteínas do *Rosetta* [2] (ROHL et al., 2004) consiste em um conjunto de protocolos tanto para predição por primeiros princípios utilizando conhecimento experimental, quanto para modelagem comparativa (SONG et al., 2013).

Na abordagem baseada em conhecimento, o método utilizada uma base de fragmentos gerada de forma personalizada para cada proteína-alvo (Veja Sec. 2.2.1). Essa base possui fragmentos do tamanho de 3 e 9 resíduos de aminoácidos. Utilizando estratégia de *Monte Carlo* (*Replica Exchange Monte Carlo*), o método inicialmente busca formar modelos de baixa precisão estrutural (*low-resolution*), de modo a se buscar apenas um empacotamento global (nesta etapa o método ainda não visa ajustes locais dos modelos). Para tal, realizam-se etapas de inserção de fragmentos de tamanho 9, onde estes são avaliados por uma função de energia composta apenas por um termo que avalia a repulsão estérica dos átomos. Dessa forma, buscando a minimização da função, os fragmentos são rejeitados ou não de acordo com o impacto na estrutura geral. À medida que modelos completos são gerados, novas inserções são realizadas e novamente analisadas, aumentando-se o número de termos da função de energia a cada estágio, elevando o nível de precisão dos modelos. Durante estas etapas, os valores e limiares de análise do método variam de modo a proporcionar o que os autores chamam de *"relaxamento"*,

[1] <http://robetta.bakerlab.org/>
[2] <https://www.rosettacommons.org/>

permitindo que se explorem novas conformações e consequentemente aperfeiçoando os modelos (CONWAY et al., 2014).

Em seguida, o método realiza tentativas de inserção de fragmentos de tamanho 3, buscando refinar os modelos por meio da minimização de uma função de energia composta por todos os termos (Apresentados pela Tab. 3.1. Mais detalhes em (SIMONS et al., 1997; SIMONS et al., 1999)). Por fim, estas estruturas são agrupadas com técnicas de *clustering* baseando-se em similaridade estrutural. Como resultado, o método analisa e seleciona modelos em relação a seus respectivos valores de energia e aos grupos mais numerosos.

Tabela 3.1: Termos que compõem a função de energia do método *Rosetta*.

Termo	Descrição
`env`	Ambiente de resíduos (solvatação)
`pair`	Interações de resíduos pareados (eletrostática, dissulfetos)
`SS`	Pareamento de fitas (ligação de hidrogênio)
`sheet`	Arranjo de fitas em folhas
`HS`	Empacotamento hélice-fita
`rg`	Raio de giro (atração vdw ; solvatação)
`cbeta`	Cβ densidade (solvatação, correção para o efeito de volume excluído introduzido pela simulação)
`vdw`	Repulsão estérica

Fonte: Adaptado de Rohl et al. (2004).

3.3 Resumo do Capítulo

Neste capítulo foram apresentadas diferentes políticas de migração em AGD que podem ser aplicadas em quaisquer problemas. Tais políticas foram implementadas e utilizadas como comparação à política proposta, buscando encontrar um modelo que melhor se adeque a problemas de otimização complexos, fazendo uso de mecanismos de controle de diversidade. Também apresentaram-se dois métodos de predição da estrutura de proteínas (Um AM estruturado em árvore ternária e o método *Rosetta*), ambos classificados como métodos de primeiros princípios que utilizam conhecimento experimental. Tais métodos foram utilizados nas comparações que serão apresentadas no Capítulo 5. O próximo capítulo apresentará o desenvolvimento da abordagem proposta, bem como a metodologia empregada. Também serão apresentados as configurações dos experimentos realizados.

4 MATERIAIS E MÉTODOS

O desenvolvimento do método proposto foi dividido em duas etapas: (I) definição da política de migração utilizada na construção do AGD e; (II) incrementos ao método de melhor desempenho na etapa anterior utilizando conhecimento específico do problema PSP. Devido à complexidade do problema PSP, sabe-se que os diversos modelos de funções de energia encontram dificuldades para representar de modo preciso o comportamento de estruturas de proteínas durante o processo de busca pela estrutura nativa da mesma (KIM et al., 2009). Dado este viés, optou-se por realizar na etapa I o desenvolvimento do método em um nível global por meio de experimentação em funções de teste, evitando assim quaisquer possíveis interferências. Na etapa II, novos componentes foram elaborados, incorporando o conhecimento específico do problema (i.e., operadores genéticos com tomada de decisão utilizando informação biológica).

Como forma de avaliação, as abordagens utilizadas na etapa I foram submetidas a um conjunto de funções de testes, presentes no pacote de *benchmark* do CEC (AWAD M. Z. ALI; QU, 2016). Na etapa II, os componentes desenvolvidos foram testados em conjunto com a melhor abordagem da etapa I em um conjunto de 6 proteínas com diferentes padrões de estrutura secundária. Por fim, a versão final do método proposto (melhor abordagem da etapa I em conjunto dos componentes de melhor resultado na etapa II) foi comparada com dois métodos de predição de estrutura tridimensional de proteínas (Veja Capítulo 3), por meio de um conjunto de teste contendo um número maior de proteínas. Todos os testes foram executados em um servidor IBM X3650 M5 - *Intel Xeon* E5-2650V4 30 MB, 4 CPUs, 2.2Ghz, 96 *cores* / *threads*, 128 GB ram e disco com 4 TB.

4.1 Etapa I - Desenvolvimento da Abordagem Aplicada em Funções de Teste

Nesta seção, iremos descrever uma nova política de migração que visa manter a diversidade populacional através de um mecanismo de compartilhamento de indivíduos com posições semelhantes no ranqueamento pelo *fitness* em suas próprias populações. Da mesma forma que o BRKGA, a política proposta tira vantagem do fato de a população ser estruturada, explorando assim a diversidade existente e compartilhando-a com as demais *demes* em uma topologia totalmente conectada. Diferentemente de outras abordagens, a política proposta baseia-se apenas no *fitness*, não sendo necessário o cálculo de similaridade entre os indivíduos das subpopulações. Em futuras referências, a chamaremos de

Política 6. Para melhor compreensão, temos abaixo algumas definições:

Definição 1. *Seja* $\boldsymbol{n}$ *o número de indivíduos na população de uma única deme e* $\boldsymbol{k}$ *o número total de demes, a matriz* $\boldsymbol{M_{n,k}}$ *representa a união dos indivíduos ordenados de cada deme, como definida abaixo:*

$$\boldsymbol{M_{n,k}} = \begin{pmatrix} ind_{1,1} & ind_{1,2} & \cdots & ind_{1,\boldsymbol{k}} \\ ind_{2,1} & ind_{2,2} & \cdots & ind_{2,\boldsymbol{k}} \\ \vdots & \vdots & \ddots & \vdots \\ ind_{\boldsymbol{n},1} & ind_{\boldsymbol{n},2} & \cdots & ind_{\boldsymbol{n,k}} \end{pmatrix}, \tag{4.1}$$

onde $ind_{i,j}$ *representa o i-ésimo indivíduo da j-ésima deme e o fitness de* $ind_{i,j}$ *é melhor ou igual ao fitness de* $ind_{i+1,j}$*.*

Definição 2. *Seja* $\boldsymbol{w} = \frac{\boldsymbol{n}}{\boldsymbol{k}}$*, a matriz* $\boldsymbol{G_{k,k}}$ *é obtida através da função de agrupamento* $f_1(\boldsymbol{M}) \rightarrow \boldsymbol{G}$ *tal que:*

$$\boldsymbol{G_{k,k}} = \begin{pmatrix} g_{1,1} & g_{1,2} & \cdots & g_{1,\boldsymbol{k}} \\ g_{2,1} & g_{2,2} & \cdots & g_{2,\boldsymbol{k}} \\ \vdots & \vdots & \ddots & \vdots \\ g_{\boldsymbol{k},1} & g_{\boldsymbol{k},2} & \cdots & g_{\boldsymbol{k,k}} \end{pmatrix}, \tag{4.2}$$

onde $g_{p,q}$ *representa o vetor coluna* $\vec{v} = [\boldsymbol{M}_{p_\alpha,q} \cdots \boldsymbol{M}_{p_\beta,q}]^T$*,* $p_\alpha = (p-1) \times \boldsymbol{w} + 1$ *e* $p_\beta = p \times \boldsymbol{w}$*.*

Definição 3.

A abordagem proposta consiste em calcular a matriz transposta $\boldsymbol{G^T_{k,k}}$ *e submete-la à função de desagrupamento* $f_2(\boldsymbol{G}^T) \rightarrow \boldsymbol{F}$*, obtendo a matriz* $\boldsymbol{F_{n,k}}$*, tal que:*

$$\boldsymbol{F_{n,k}} = \begin{pmatrix} ind_{1,1} & ind_{1,2} & \cdots & ind_{1,\boldsymbol{k}} \\ ind_{2,1} & ind_{2,2} & \cdots & ind_{2,\boldsymbol{k}} \\ \vdots & \vdots & \ddots & \vdots \\ ind_{\boldsymbol{n},1} & ind_{\boldsymbol{n},2} & \cdots & ind_{\boldsymbol{n,k}} \end{pmatrix}, \tag{4.3}$$

onde $ind_{i,j}$ *representa o r-ésimo elemento do vetor coluna* $\vec{v}_{j,i_\alpha}$ *da matriz agrupada* $\boldsymbol{G_{k,k}}$*, sendo:*

$$r = \begin{cases} \boldsymbol{w} & \textit{se i mod } \boldsymbol{w} \textit{ = 0} \\ \textit{i mod } \boldsymbol{w} & \textit{caso contrário} \end{cases} \tag{4.4}$$

e $i_\alpha = \lceil \frac{i}{\boldsymbol{w}} \rceil$.

Após isso, a *i*-ésima coluna da matriz **F** representa a nova população da *i*-ésima *deme*. É importante ressaltar que devido ao fato dos indivíduos estarem ordenados, este processo resulta em novas populações formadas por indivíduos provenientes de todas as populações anteriores e contidos no mesmo intervalo de *fitness*. Para garantir que as novas populações tenham o mesmo tamanho que as anteriores, a seguinte igualdade deve ser verdadeira:

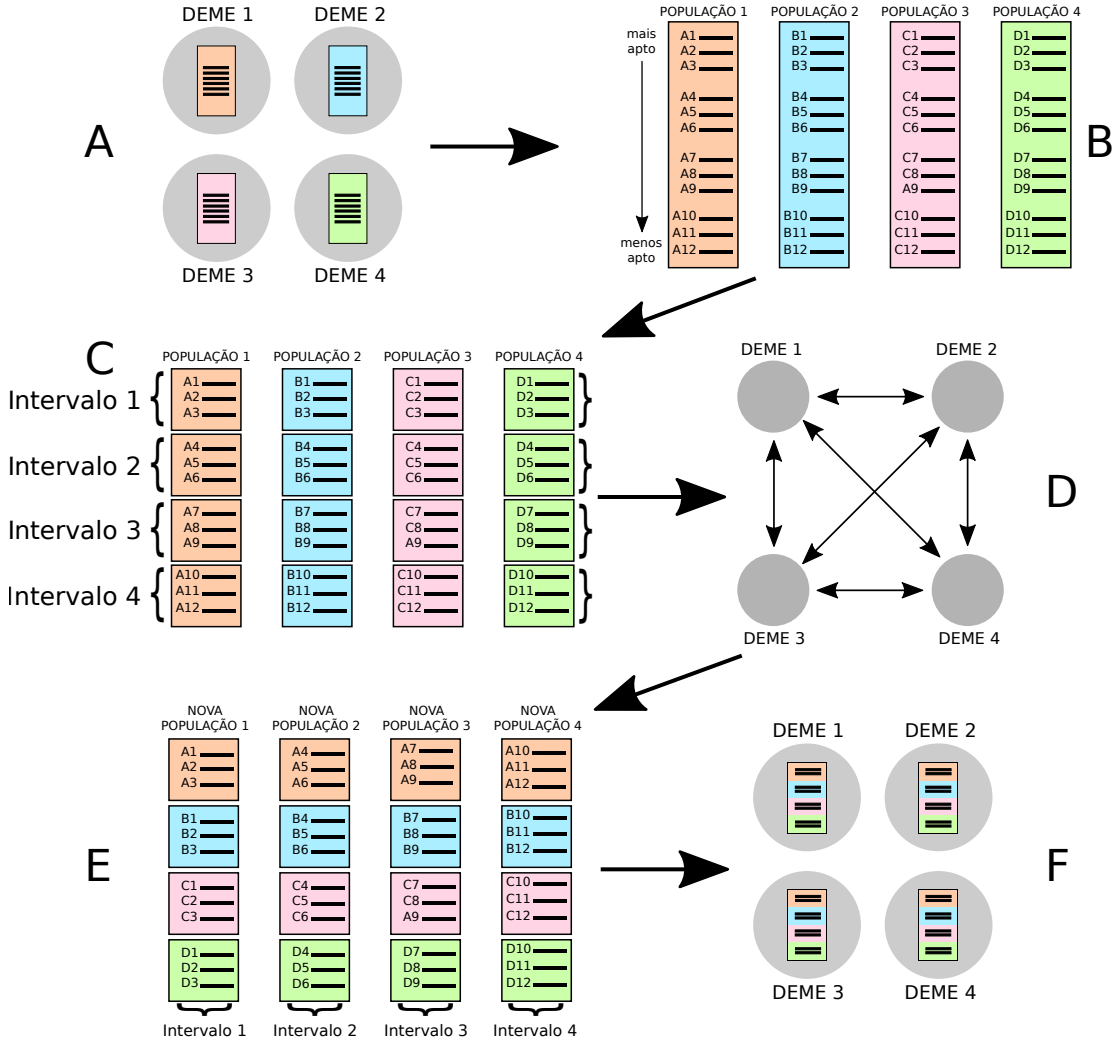
$$\mathbf{t}\ mod\ \mathbf{k}^2 = 0, \tag{4.5}$$

onde **t** representa o tamanho total da população do algoritmo dado por $\mathbf{t} = \mathbf{m} \times \mathbf{k}$. A Figura 4.1 ilustra todo o processo.

Como apresentado pela Figura 4.1, a permutação dos indivíduos configura uma comunicação totalmente conectada. Até a próxima migração, cada população terá disponível material genético proveniente de todas as outras *demes* possibilitando assim a chance de explorar combinações no espaço de busca que antes poderiam estar inalcançáveis devido à convergência local. Entretanto, para evitar uma grande disparidade entre indivíduos após a migração, a nova população é formada por indivíduos que estejam localizados no mesmo intervalo de *fitness* em suas respectivas populações, mas não necessariamente que possuam valores de *fitness* semelhantes. Isso garante que, apesar de em um primeiro momento possuir um comportamento de *exploitation*, ao longo prazo a política oscila entre *exploitation* e *exploration*. A Figura 4.2 exemplifica este processo.

Na Figura 4.2, temos três *demes* com seis indivíduos cada, resultando em intervalos de tamanho dois (***n*=6, *k*=3 e *w*=2**). O valor $\lambda_{i,j}$ representa a magnitude da diferença de *fitness* entre o pior e o melhor indivíduo da população *i* em uma etapa de migração *j*. Podemos observar uma forte queda nos valores de λ da etapa de migração i para a etapa $i + 1$, o que configura um processo de *exploitation*, devido ao fato de indivíduos com *fitness* semelhantes serem reunidos na mesma população (os melhores na primeira *deme*, os médios na segunda e o restante na terceira), permitindo assim que os operadores genéticos possam alcançar melhorias pontuais. Já a transição entre as etapas $i + 1$ e $i + 2$ configura um processo de *exploration*, uma vez que os indivíduos semelhantes de cada *deme* são novamente separados e enviados para todas as demais. Isso resulta em novas populações mais diversificadas e consequentemente possibilita uma maior exploração do espaço de busca ou mesmo escapar de ótimos locais. O aumento dos valores de λ da etapa de migração $i + 1$ para a etapa $i + 2$ explicitam esse comportamento. É importante ressal-

Figura 4.1: O fluxo da política de migração proposta. Em **A**, as quatro *demes* estão prontas para a migração. Em **B**, temos a as subpopulações unificadas como uma população global. Em **C**, a população é dividida em grupos de tamanho $[\boldsymbol{w} \times 1]$. **D** apresenta a comunicação totalmente conectada. **E** apresenta a disposição dos indivíduos após a migração, formando as novas populações e em seguida, em **F** temos a população global novamente dividida em subpopulações.



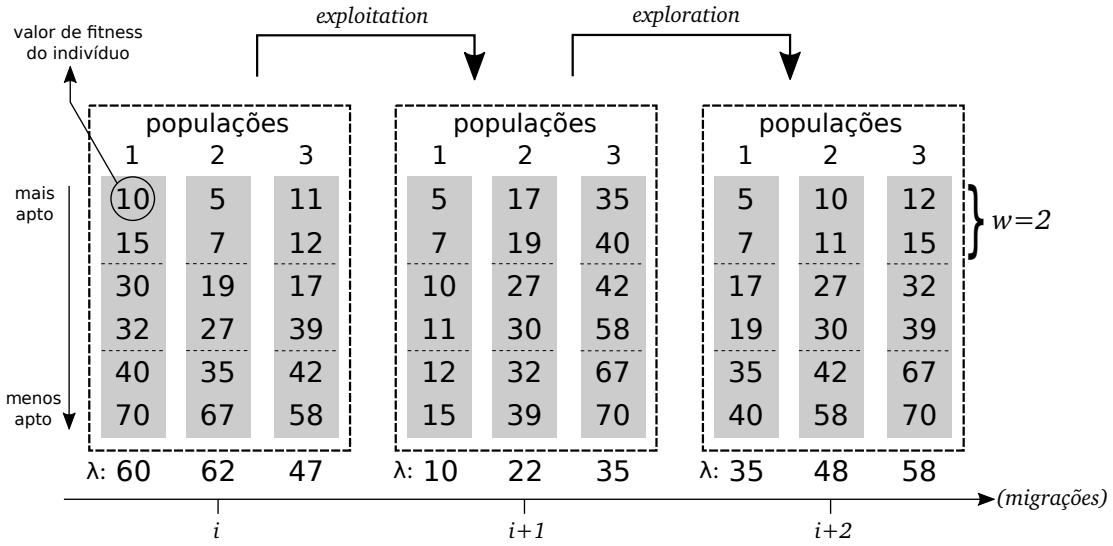
Fonte: Do Autor.

tar que o fator estocástico dos operadores genéticos influenciam este processo, sendo que quanto mais gerações ocorrerem entre as etapas de migração, maior o impacto.

4.2 Etapa II - Desenvolvimento da Abordagem para o Problema PSP

Esta seção apresentará melhorias nos operadores genéticos que exploram o conhecimento específico do problema PSP. Para tal, formas de aplicar informações de base de conhecimento foram propostas. Também foram desenvolvidas maneiras de utilizar infor-

Figura 4.2: O comportamento da abordagem proposta em relação ao tempo.



Fonte: Do Autor.

mação da estrutura secundária na tomada de decisão dos operadores de recombinação e mutação no intuito de aprimorar a qualidade da abordagem proposta.

Buscando aproveitar informações do problema, uma série de alterações foram elaboradas. A primeira delas é a maneira de se inicializar os indivíduos da população, buscando formá-los utilizando valores contidos nos fragmentos do *Rosetta* (Veja sec. 2.2.1). Além disso, diferentes operadores de recombinação e mutação foram desenvolvidos de modo a se explorar a informação da estrutura secundária. A seguir apresentaremos tais incrementos bem como os experimentos realizados com os mesmos.

4.2.1 Inicialização

A informação contida em bases de dados experimentais podem ser utilizadas na criação de novos indivíduos, com o objetivo de reduzir as possibilidades do espaço de busca da proteína-alvo (CORRÊA et al., 2016; CORRÊA; INOSTROZA-PONTA; DORN, 2017; OLIVEIRA; BORGUESAN; DORN, 2017; CORRÊA et al., 2018). De modo a aplicar os fragmentos do *Rosetta* nesse processo, dois modelos de inicialização foram propostos. Conforme será apresentado, em determinadas situações tais modelos necessitam de uma métrica que descreva a similaridade entre dois resíduos através de seus respectivos ângulos ϕ e ψ. Para tal, propomos a *Semelhança* ($\bar{S}$). Esta métrica consiste em somar a diferença entre os ângulos ϕ dos dois resíduos com a diferença de seus ângulos ψ. Dessa forma, $\bar{S}$ é expressa no intervalo $[0, 360]$, sendo quando menor seu valor,

maior a similaridade entre os resíduos. A Equação 4.6 apresenta a definição de $\bar{S}$:

$$\bar{S}((phi_1, psi_1), (phi_2, psi_2)) : \ |\Delta(phi_1, phi_2)| \ + \ |\Delta(psi_1, psi_2)| \tag{4.6}$$

sendo Δ a função que calcula a menor diferença entre dois ângulos conforme apresentado pela Equação 4.7:

$$\Delta(ang_1, ang_2) : min(\ \ (\alpha - \beta), \ \ (\gamma_2 - \alpha + \beta - \gamma_1)\ \ ) \tag{4.7}$$

onde α = *max*(ang_1, ang_2), β = *min*(ang_1, ang_2); γ_1 = -180,0 e γ_2 = 180,0 (respectivamente o intervalo de representação dos ângulos).

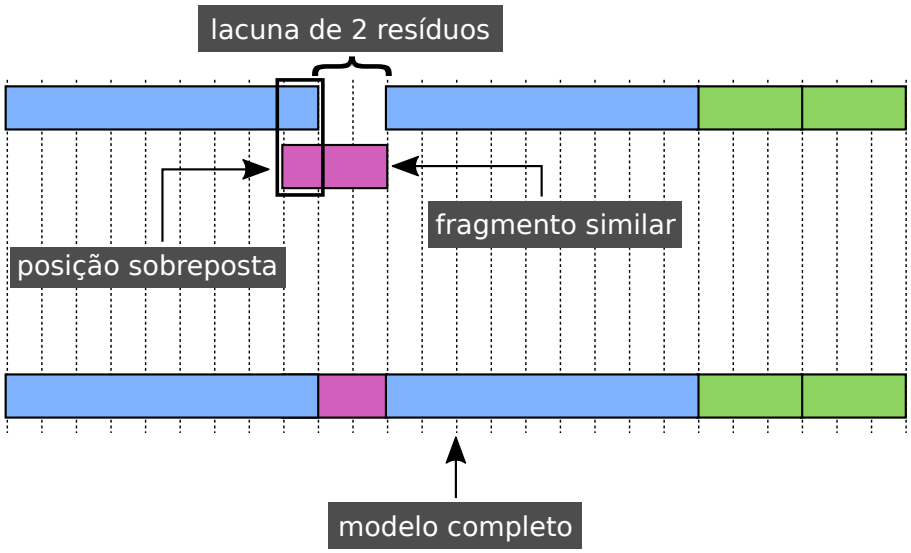
4.2.1.1 Modelo de Inicialização I

Este modelo inspira-se no modo de inserção dos fragmentos utilizados no método *Rosetta*. Para preencher um novo indivíduo (e.g., inicialização da população), fragmentos de tamanho 9 são aleatoriamente selecionados e inseridos em posições também aleatórias sem sobreposição. À medida que não houver mais a possibilidade de inserir fragmentos de tamanho 9, começa-se a inserir fragmentos de tamanho 3, igualmente selecionados e posicionados de maneira aleatória. Em seguida, caso haja espaços vazios de tamanho 2, procura-se por fragmentos de tamanho 3 que possam preencher tal lacuna de modo a haver sobreposição com o resíduo anterior (que já está preenchido). Dessa forma, este fragmento de tamanho 3 contribui com 2 valores (a prioridade na posição sobreposta é sempre do resíduo inserido primeiro). Devido ao fato deste fragmento sobrepor um outro fragmento já inserido em uma posição, buscam-se por fragmentos que minimizem o valor de $\bar{S}$ em relação àquela posição já preenchida, com o objetivo de se manter uma certa similaridade devido a esse processo realizar uma mescla de fragmentos. A Figura 4.3 ilustra essa situação.

A escolha do fragmento de tamanho 3 que irá preencher esta lacuna de tamanho 2 ocorre da seguinte forma:

1. Calcula-se a *semelhança* (Eq. 4.6) entre o resíduo anterior à lacuna e todos os fragmentos de tamanho 3 contidos na base que possam preencher aquela região.
2. Ordenam-se os fragmentos pelo valor da *semelhança*, escolhendo aleatoriamente um entre os 5 mais similares.

Figura 4.3: Exemplo de lacuna de tamanho 2. Para preenchê-la, um novo fragmento de tamanho 3 (rosa) é selecionado baseado na sua similaridade na posição de sobreposição. Em azul temos os fragmentos de tamanho 9, seguidos dos de tamanho 3, em verde.



Fonte: Do Autor.

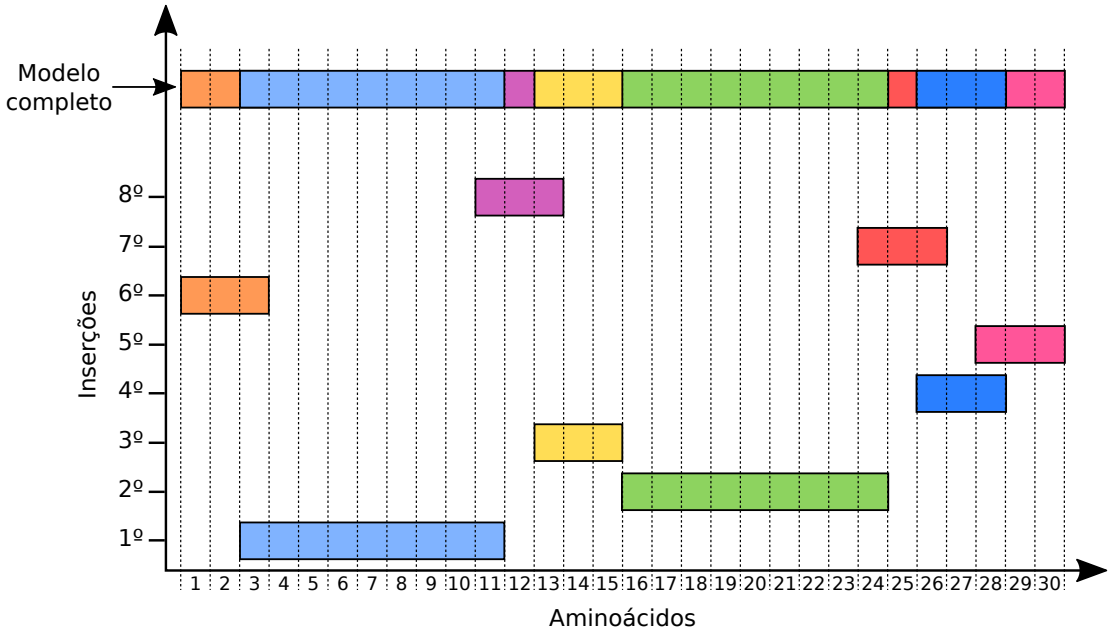
O mesmo ocorre para lacunas de tamanho 1, exceto que o fragmento de tamanho 3 irá contribuir com apenas uma posição. A Figura 4.4 ilustra um exemplo de inicialização. Este modelo busca explorar o fator estocástico durante as inserções, optando pela *semelhança* apenas em situações pontuais.

4.2.1.2 Modelo de Inicialização II

Com o objetivo de reduzir a influência do fator de aleatoriedade, este modelo adota de maneira mais concisa o critério de seleção de fragmentos de acordo com sua *semelhança*. Diferentemente do anterior, este modelo não dá prioridade para fragmentos maiores (modelo anterior insere fragmentos de tamanho 9, seguido de tamanho 3, para depois preencher lacunas de tamanho 2 e 1). Dessa vez, fragmentos de tamanho 9 e 3 possuem chances equiprováveis de serem selecionados, desde que a região na qual o mesmo é inserido possua espaço suficiente. Inicialmente este modelo seleciona um fragmento de maneira aleatória e o insere no indivíduo. Este fragmento é chamado de *pivô*. O método então procura preencher consecutivamente os espaços vazios tanto à esquerda quanto à direita do *pivô*, da seguinte maneira:

1. Para cada inserção, sorteia-se com probabilidades iguais entre inserir um fragmento de tamanho 9 ou 3.
2. Calcula-se a *semelhança* entre o *pivô* e todos os fragmentos da base do tamanho

Figura 4.4: Inicialização usando fragmentos - Modelo I. Inicialmente fragmentos de tamanho 9 são inseridos (azul claro e verde), seguidos pelos de tamanho 3 (amarelo e azul escuro). Lacunas de tamanho 2 e 1 são preenchidas com partes de fragmentos de tamanho 3 (rosa, laranja, vermelho e roxo), escolhidos de acordo com a semelhança entre eles e os valores já inseridos no indivíduo.



Fonte: Do Autor.

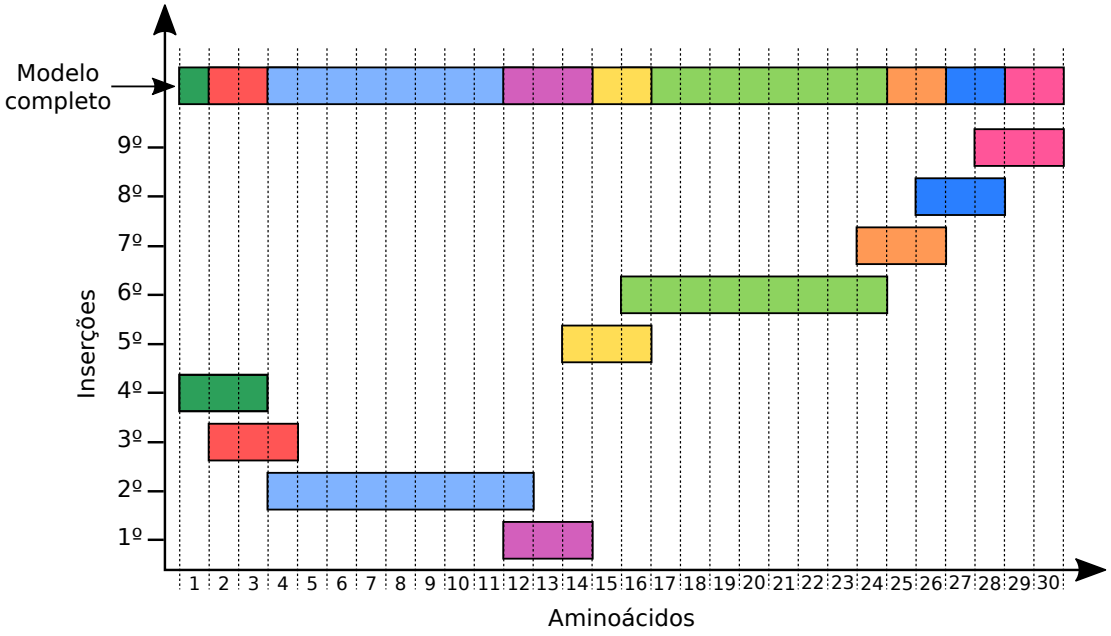
sorteado e que possuam sobreposição de 1 resíduo com o mesmo.

3. Escolha-se aleatoriamente um entre os 5 mais similares para ser inserido, preservando sempre o valor sobreposto do *pivô*.
4. O novo fragmento inserido passa a ser o *pivô*.

Após realizar os passos acima para um dos lados, volta-se o *pivô* para o primeiro fragmento inserido e repetem-se os passos para o lado restante.

Nota-se que apenas o primeiro fragmento inserido contribui com a totalidade de sua informação. As demais inserções representam n-1 valores, sendo n o tamanho do fragmento. Caso alguma das extremidades do indivíduo permaneça vazia em 1 posição, então o processo se repete de maneira semelhante, exceto que seleciona-se um fragmento semelhante da base de tamanho 3, na qual irá contribuir com apenas 1 valor. A Figura 4.5 apresenta um exemplo um exemplo de inicialização. Este modelo busca explorar a relação de dependência dos fragmentos durante as inserções. Por meio da métrica de *semelhança*, a influência do fator estocástico é reduzida em comparação ao modelo anterior.

Figura 4.5: Inicialização usando fragmentos - Modelo II. Ambos os tamanhos de fragmentos possuem probabilidades iguais de serem utilizados. Inicialmente um fragmento é escolhido (*pivô*, em roxo) e, a partir dele realizam-se inserções em ambas as adjacentes buscando sempre por fragmentos similares. Cada cor representa uma inserção distinta.



Fonte: Do Autor.

4.2.2 Recombinação

O operador de recombinação possui grande influência no desempenho de um AG. Com o objetivo de se analisar o impacto de diferentes modelos, nesta etapa utilizou-se a recombinação uniforme (BEAN, 1994) (Veja Fig. 2.8), seguido de um novo modelo proposto baseado na informação da estrutura secundária. Dessa forma os modelos testados possuem características para problemas de aspecto geral (recombinação uniforme) e especializadas no problema (recombinação por estrutura secundária). Além disso, um terceiro modelo combinando ambas também foi testado.

4.2.2.1 Recombinação Uniforme

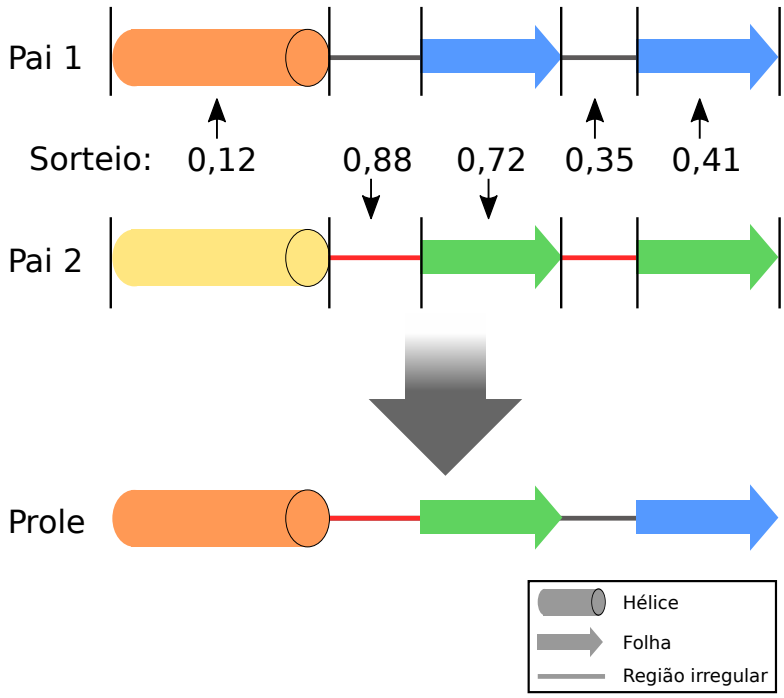
Este modelo, conforme já introduzido na Seção 2.3.3, busca tomar a decisão de qual valor utilizar a cada posição de maneira individual. Com uma maior probabilidade se escolher valores do indivíduo de melhor *fitness*, a recombinação uniforme trata cada valor do indivíduo de forma independente, anulando possíveis relações de dependência entre as variáveis em certos casos (e.g., a posição i do indivíduo possui um valor x sempre que a posição $i + 1$ possui um valor y). Ao utilizarmos informação da base de fragmentos para

inicializar nossos indivíduos, temos exatamente este cenário. Um fragmento contém valores de 3 ou 9 resíduos que ocorrem em conjunto, de forma dependente. Visando preservar ao menos em partes tais relações de dependência, um segundo modelo de recombinação foi proposto.

4.2.2.2 Recombinação Baseada na Estrutura Secundária

Neste modelo, a as posições do indivíduo são subdivididas de acordo com seus padrões de estrutura secundária (Veja Sec. 2.1.1.2), semelhante ao utilizado por Corrêa et al. (2016) na recombinação *crossoverSS* (e.g., sequência de estrutura s=CHHHCCEE resulta na subdivisão *d*:{C; HHH; CC; EE}). Em seguida, para cada elemento da sequência *d*, sorteia-se, sob uma probabilidade *p*, qual indivíduo deverá repassar sua informação genética à prole. A Figura 4.6 ilustra este modelo. Este processo é semelhante à recombinação uniforme, entretanto o mesmo é capaz de preservar, ao menos em partes, as relações de dependência entre os valores contidos em um indivíduo. Seguindo o padrão utilizado em Gonçalves e Resende (2011) e Alixandre e Dorn (2017), a probabilidade *p* empregada foi de 70% de chances para o indivíduo de melhor *fitness*.

Figura 4.6: Recombinação por estrutura secundária. Pai 1 é o indivíduo com melhor *fitness* e, transmite seu material genético com probabilidade de 70%.



Fonte: Do Autor.

4.2.2.3 Recombinação Mista

Devido à grande distinção entre as características dos modelos apresentados, optou-se por aplicar uma terceiro modelo combinando ambos. Sabendo que a recombinação uniforme possui grande capacidade de explorar os valores contidos nos indivíduos, dado que ela atua em cada posição separadamente, e que a recombinação por estrutura secundária é capaz de preservar certas relações de dependência, este modelo tem por objetivo utilizar ambas durante o processo de otimização. A ideia é que o algoritmo de otimização comece utilizando a recombinação por estrutura secundária até que metade da execução seja alcançada e, em seguida, passe a utilizar a recombinação uniforme.

A hipótese é que, devido ao fato do algoritmo começar com indivíduos aleatórios e pouco otimizados, preservar sub-regiões dependentes seja importante até que se atinja um maior nível de qualidade estrutural. Uma vez que a população esteja nesse nível, a recombinação uniforme é aplicada de modo a refinar tais indivíduos. O fato de aplicar a recombinação uniforme em uma população mais evoluída pode minimizar o impacto deste modelo em relação à quebra de dependência de variáveis, considerando que ambos os indivíduos selecionados estão mais avançados no processo de convergência, aumentando as chances de ambos conterem bons valores de material genético. Em suma, espera-se que aplicar a recombinação uniforme em indivíduos de qualidade aumente as chances de refiná-los em soluções ainda melhores.

4.2.3 Mutação

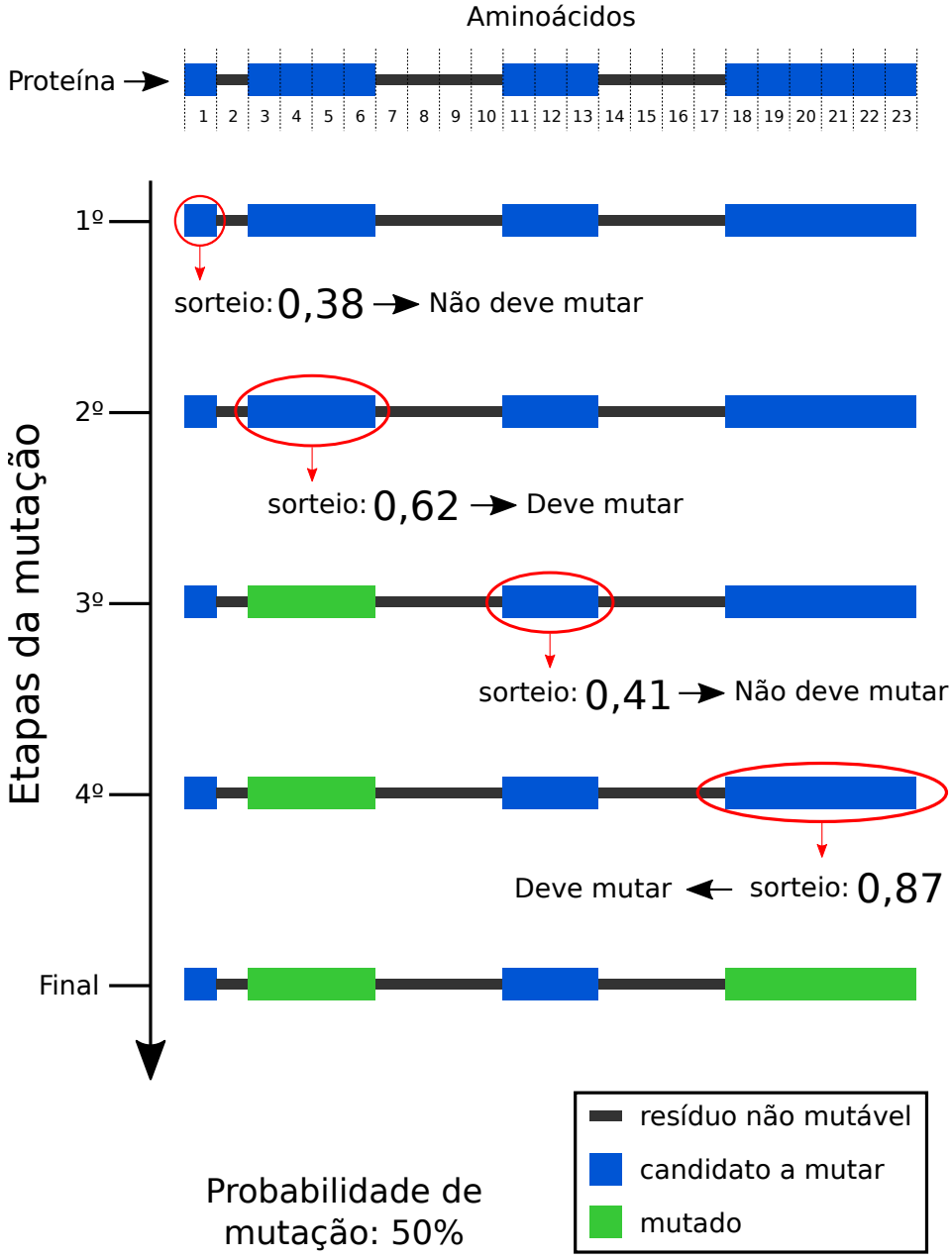
O operador de mutação está diretamente relacionado à capacidade do algoritmo de explorar novas regiões no espaço de busca. Tal característica tem papel fundamental no escape de ótimos locais. Sabemos também que certas tipos de estruturas secundárias possuem maior flexibilidade que outras, chamadas de regiões irregulares. Com o objetivo de combinar a ação exploratória da mutação com esta característica de flexibilidade de certas regiões estruturais das proteínas, dois modelos de mutação foram propostos.

4.2.3.1 Modelo de Mutação I

Este modelo permite mutações apenas em regiões irregulares. Inicialmente, deve-se calcular o conjunto de trechos que são irregulares (e.g., estrutura=CHHHCCEE $\rightarrow$ d:{C; CC}). Cada trecho contido no conjunto d tem uma probabilidade p = 50% de sofrer

mutação (condição equiprovável, sem priorizar nenhum cenário). A Figura 4.7 apresenta uma exemplo.

Figura 4.7: Mutação - Modelo I. Cada trecho de região irregular tem 50% de probabilidade de mutação.

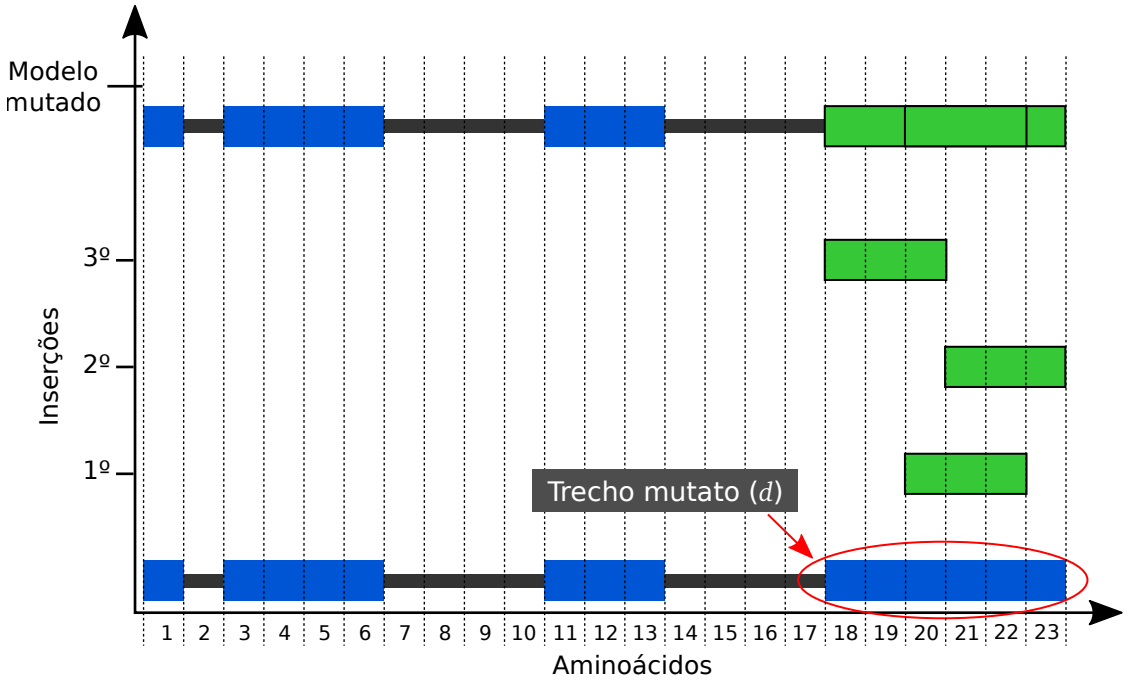


Fonte: Do Autor.

Quando um trecho é selecionado para mutação, todos os seus resíduos serão substituídos por novos valores contidos na base de fragmentos de tamanho 3. Inicialmente, fragmentos inteiros de tamanho 3 são inseridos enquanto houver espaço. Em seguida, caso haja lacunas de tamanho 2 ou 1, novas inserções ocorrem, selecionando um dentre os 5 com maior valor de *semelhança* (Veja Eq. 4.6) com o resíduo imediatamente adja-

cente. A Figura 4.8 apresenta um exemplo de mutação em um trecho com 6 resíduos de tamanho.

Figura 4.8: Exemplo de mutação em um trecho - Modelo I. Inicialmente fragmentos de tamanho 3 são inseridos. Lacunas de tamanho 2 e 1 são preenchidas em seguida, seguindo o critério de *semelhança*.



Fonte: Do Autor.

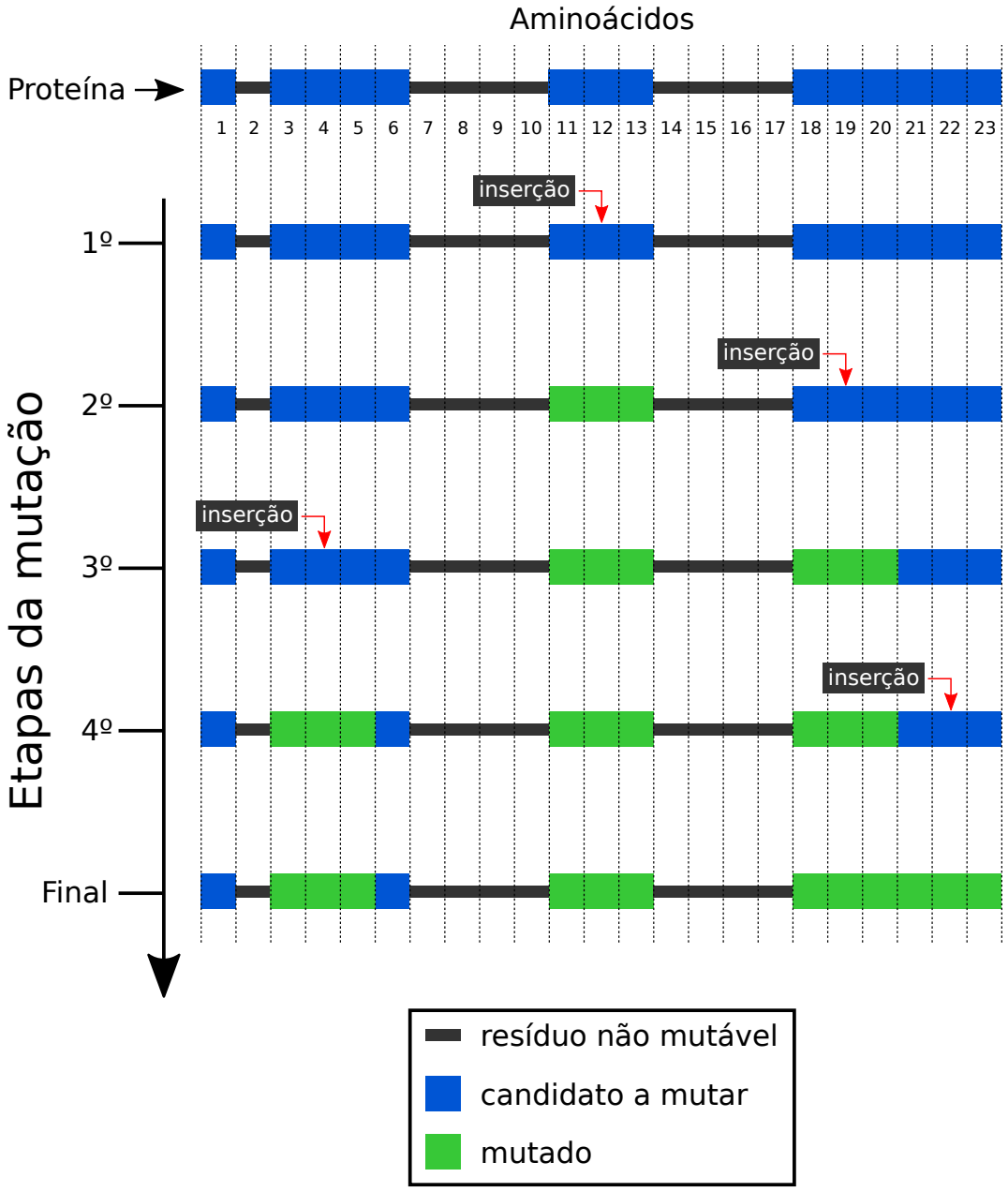
4.2.3.2 Modelo de Mutação II

Neste segundo modelo, com objetivo de se manter a integridade dos fragmentos inseridos na mutação, apenas regiões irregulares com espaço suficiente para receber fragmentos de tamanho 3 são alteradas. De modo parecido ao modelo anterior, identificam-se os trechos do indivíduo que são regiões irregulares e, de modo aleatório, realizam-se inserções de fragmentos de tamanho 3 enquanto houver espaço. Caso haja lacunas de tamanho 2 ou 1 que não sofreram mutação, seus valores originais são mantidos. Dessa forma, os novos fragmentos inseridos permanecem inteiros, diferente do modelo anterior que completava tais lacunas com partes de fragmentos. A Figura 4.9 ilustra esta situação.

4.3 Etapa III - Aplicando a Versão Final ao Problema PSP

Com o intuito de analisar o desempenho do método desenvolvido, utilizaram-se duas abordagens de predição de estrutura de proteínas como meio de comparação (Veja

Figura 4.9: Mutação - Modelo II. Apenas inserções de fragmentos de tamanho 3 são permitidas.



Fonte: Do Autor.

Sec. 3.2.1 e 3.2.2).

A primeira abordagem a ser comparada é o método *Rosetta* proposto por Rohl et al. (2004). Este método utiliza uma base de fragmentos para montar soluções e evoluí-las por meio de simulações de *Monte Carlo*. Este processo ocorre de maneira progressiva, começando com fragmentos de tamanho 9 e com funções de energia mais básicas, visando apenas desenvolver um empacotamento global, e vai aumentando a qualidade das soluções à medida que utilizam-se funções mais robustas e fragmentos de tamanho 3, re-

sultando em estruturas refinadas. O método utiliza a informação da estrutura secundária de modo a guiar seu processo, sendo tal informação obtida por meio da técnica de predição de estrutura secundária Psipred. Devemos ressaltar que os métodos como o Psipred, apesar de bons resultados, ainda não possuem total precisão em suas predições, incumbindo o método em questão de lidar com tal viés para que se alcance uma predição de estrutura tridimensional de qualidade.

A segunda abordagem, proposta por Corrêa et al. (2016), consiste em um AM dividido em subpopulações em uma estrutura de árvore ternária. Tal técnica, chamada pelos autores de M5, combina os operadores genético de um AG com uma busca local, de modo a refinar as soluções. O método também utiliza os dados da APL como forma de conhecimento visando reduzir o espaço de busca das soluções. Em seus experimentos, os autores utilizam a informação da estrutura secundária atribuída, ou seja, com total precisão da informação uma vez que ela provêm da análise sobre a estrutura experimental. Isso tem por objetivo eliminar quaisquer interferências que a predição da estrutura secundária traria caso a mesma fosse utilizada.

4.4 Experimentos - Etapa I

Esta seção apresentará as abordagens utilizadas na comparação à política proposta. Também serão apresentadas as configurações dos experimentos realizados juntamente com as métricas de avaliação.

4.4.1 Abordagens Testadas

De modo a obter uma melhor análise do comportamento da abordagem proposta bem como as demais políticas (Veja Tab. 4.1 para maiores detalhes), os experimentos foram realizados utilizando inicialmente o AG canônico e posteriormente o BRKGA. Os testes com o algoritmo genético canônico nos permitem analisar o impacto de cada política isoladamente e, ao aplicarmos o BRKGA, podemos observar seu potencial de contribuição com abordagens mais robustas. Devido a suas próprias características, a Política 2 foi executada apenas com o BRKGA, resultando assim em 11 combinações de pares Algoritmo/Política conforme apresentado pela Tabela 4.2.

Em algumas das abordagens utilizadas, a mensuração da distância entre indivíduos

Tabela 4.1: Políticas e suas características. P1 representa a política canônica, P2, P3, P4 e P5 são políticas baseadas on diversidade e P6 é a política proposta.

Política	**Baseado em**	**Arquitetura**	**Topologia**	**Ref.**
P1	*Fitness*	Ilha	Anel - unidirecional	Alg. 5
P2	*Fitness* estruturado	Ilha	Anel- unidirecional	Alg. 6
P3	*Fitness* + Similaridade	Ilha	Anel - unidirecional	Alg. 7
P4	*Fitness* + Similaridade	Ilha	Anel - unidirecional	Alg. 8
P5	*Fitness* + Similaridade	Ilha	Anel - unidirecional	Alg. 9
P6	*Fitness* estruturado	Ilha	Totalmente conectada	-

Fonte: Do Autor.

Tabela 4.2: Abordagens testadas. Combinação de Ga's e políticas.

Abordagem	**AG**	**Ref.**	**Política**	**Ref.**
A1	AG canônico	Alg. 1	P1	Alg. 5
A2	AG canônico	Alg. 1	P3	Alg. 7
A3	AG canônico	Alg. 1	P4	Alg. 8
A4	AG canônico	Alg. 1	P5	Alg. 9
A5	AG canônico	Alg. 1	P6	Sec. 4.1
A6	BRKGA	Alg. 2	P1	Alg. 5
A7	BRKGA	Alg. 2	P2	Alg. 6
A8	BRKGA	Alg. 2	P3	Alg. 7
A9	BRKGA	Alg. 2	P4	Alg. 8
A10	BRKGA	Alg. 2	P5	Alg. 9
A11	BRKGA	Alg. 2	P6	Sec. 4.1

Fonte: Do Autor.

se faz necessária para a tomada de decisão. Com o objetivo de padronizar, utilizaremos como métrica a distância Euclidiana, conforme apresentada:

$$\textit{Distância}(X, Y) : \sqrt{(X_1 - Y_1)^2 + (X_2 - Y_2)^2 + \cdots + (X_d - Y_d)^2}, \tag{4.8}$$

onde *X* e *Y* representam os dois indivíduos no qual será calculada a distância, e *d* representa a dimensão destes indivíduos.

4.4.2 Benchmark do CEC 2017

Ao longo dos últimos anos, além dos próprios competidores, muitos trabalhos tem utilizado as funções de *benchmark* propostas pelas competições do CEC (ELSAYED;

SARKER; ESSAM, 2014; BANITALEBI; AZIZ; AZIZ, 2016; VIKTORIN; PLUHACEK; SENKERIK, 2016; ALIXANDRE; DORN, 2017). Além de constantes incrementos, os organizadores disponibilizam este pacote em diversas linguagens de programação, tornando assim um meio consolidado para validação de diversas técnicas de otimização. Neste trabalho, utilizamos o pacote mais recente, disponibilizado em <<https://goo.gl/SwQGQh>> e que possui instruções descritas no relatório de *benchmark* do CEC 2017 (AWAD M. Z. ALI; QU, 2016). Foram escolhidas 10 funções para serem testadas, variando suas características e complexidades conforme apresentado pela Tabela 4.3 e ilustrado pela Figura 4.10.

Tabela 4.3: Conjunto de funções de *benchmark* do CEC 2017. N representa o número de funções básicas que compõem a função(apenas para híbridas e compostas) e F_i^* representa o valor de *fitness* ótimo da *i*-ésima função.

Funções							
ID	Característica	N	F_i^*		ID	Característica	N	F_i^*
1	Unimodal	-	100		**16**	Híbrida	4	1600
4	Multimodal Simples	-	400		**17**	Híbrida	5	1700
6	Multimodal Simples	-	600		**22**	Composta	3	2200
9	Multimodal Simples	-	900		**24**	Composta	4	2400
11	Híbrida	3	1100		**26**	Composta	5	2600
Intervalo de busca: $[-100, 100]^D$								

Fonte: Do Autor.

4.4.3 Métricas de Avaliação

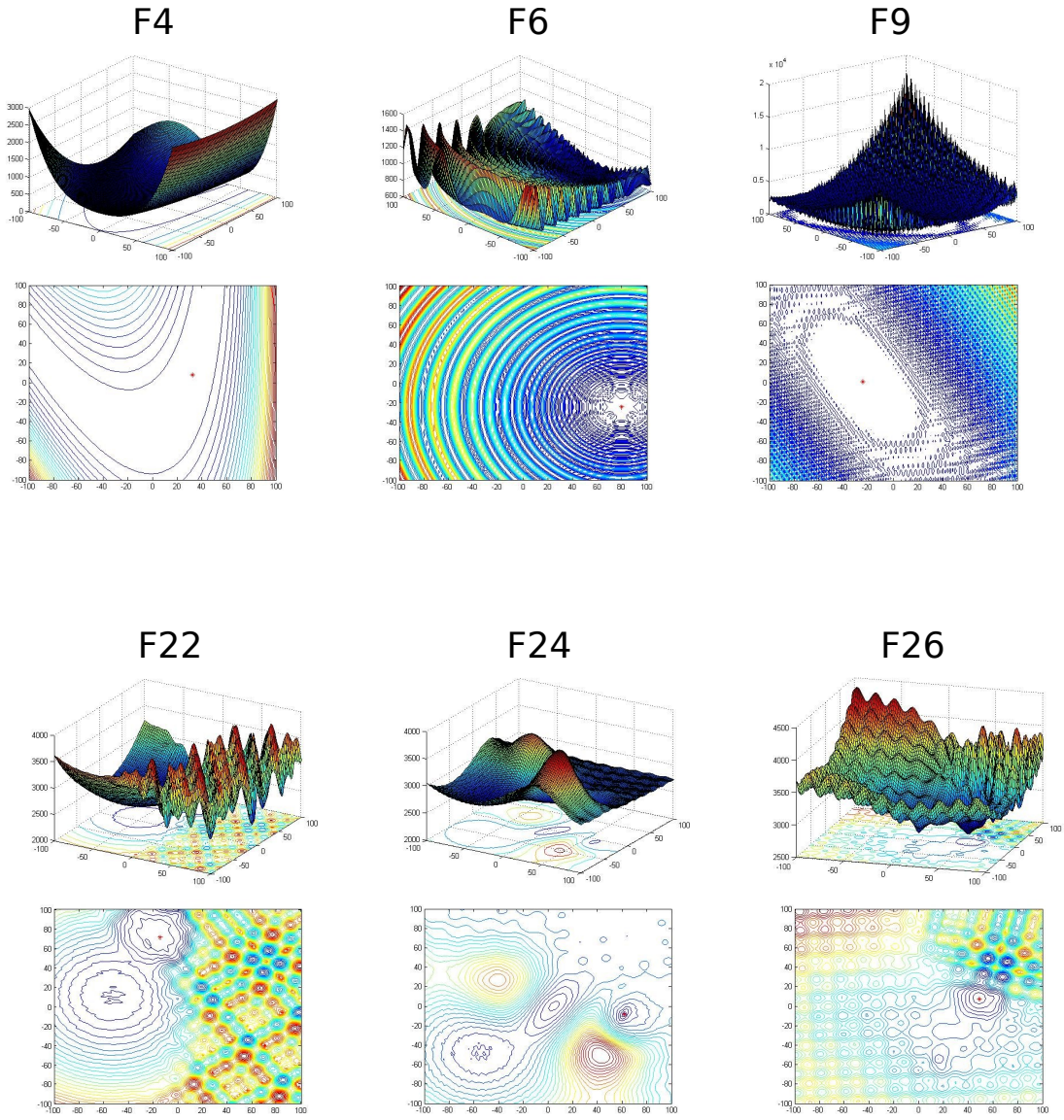
Nessa etapa, os resultados dos testes serão expressos em forma de Erro, conforme utilizado no *benchmark* do CEC 2017 (AWAD M. Z. ALI; QU, 2016) e apresentado pela Equação 4.9:

$$Erro(F_i) : F_i - F_i^* \tag{4.9}$$

sendo F_i^* o valor ótimo e F_i o resultado obtido, ambos para a *i*-ésima função respectivamente.

Apesar do objetivo principal em uma otimização ser encontrar o melhor resul-

Figura 4.10: Exemplo de funções usadas nos experimentos. Mapa 3D (superior) e mapa de contorno (inferior) para os casos 2D.



Fonte: Adaptado de Awad M. Z. Ali e Qu (2016).

tado possível, não podemos considerar apenas o melhor *fitness* ao julgarmos diferentes abordagens. Alguns fatores, tais como a complexidade, possuem grande relevância ao escolhermos qual abordagem utilizar. A seguir listaremos os demais critérios utilizados neste trabalho:

4.4.3.1 Diversidade

Como já explicado, a diversidade populacional tem um importante papel no processo de otimização. Para mensurarmos, utilizaremos uma das métricas abordadas por

Toffolo e Benini (2003) e aplicada em Alixandre e Dorn (2017), conforme a Equação 4.10:

$$Diversidade(Pop) : \frac{\sum_{i=2}^{n} Dist(Pop_1, Pop_i)}{n-1} \tag{4.10}$$

onde *Pop* é a população final ordenada pelo *fitness*, *n* é o número total de indivíduos em *Pop*, *Dist* é a distância Euclidiana e *1* é o índice do melhor indivíduo da população.

4.4.3.2 Convergência

Outro fator importante na comparação entre diferentes técnicas consiste em analisarmos a curva de convergência, expressa pelo melhor valor de *fitness* em toda a população ao longo do tempo. Esta curva nos permite visualizar o quão rápido um algoritmo converge para o resultado final bem como seu comportamento ao longo da execução. Nesta etapa, os valores foram computados a cada migração.

4.4.3.3 Complexidade

O relatório de *benchmark* do CEC 2017 (AWAD M. Z. ALI; QU, 2016) também provê uma forma de avaliarmos e compararmos a complexidade de diferentes técnicas. Para tal, deve-se calcular inicialmente o $T0$, que representa o tempo total de execução do Algoritmo 11. Esse termo será utilizado para eliminarmos a influência da máquina na qual a técnica foi executada. Em seguida, computamos em $T1$ o tempo total para se calcular 200000 avaliações de *fitness* da *Função 18* em uma certa dimensão *D*. Esse termo será utilizado para eliminarmos o tempo de execução utilizado pelos componentes do *benchmark* do CEC 2017. Por último, devemos calcular $T2$ cinco vezes e obtermos $\hat{T}2 = Mean(T2)$, sendo $T2$ o tempo total de execução da técnica para otimizar a *Função 18* com 200000 avaliações de *fitness* e com a mesma dimensão *D*. A Equação 4.11 apresenta o cálculo da complexidade proposta pelo *benchmark* do CEC 2017:

$$Complexidade : \frac{\hat{T}2 - T1}{T0} \tag{4.11}$$

Entretanto, devido ao fato de utilizarmos algoritmos distribuídos, se fez necessário a ponderação da complexidade pelo total de *demes* (*k*) de cada abordagem, uma vez que este valor variou entre as abordagens testadas. Assim sendo, a complexidade utilizada neste

trabalho é dada pela Equação 4.12:

$$Complexidade\_Final : \frac{\hat{T}2 - T1}{T0} \times k \tag{4.12}$$

Para uma melhor análise da relação da complexidade com as dimensões, deve-se calcular a *Complexidade_Final* para *D* = 10, 30 e 50. Para evitar distorções, o estilo de programação deve ser similar para todos os termos.

Algoritmo 11 Programa teste - T0

1: $x \leftarrow 0,55$
2: **para** $i = 1$ **até** 1000000 **faça**
3: $x \leftarrow x + x$
4: $x \leftarrow x \div x$
5: $x \leftarrow x * x$
6: $x \leftarrow x^2$
7: $x \leftarrow \log x$
8: $x \leftarrow \exp x$
9: $x \leftarrow x \div (x + 2)$
10: **fim para**

4.4.4 Configuração

Todas as abordagens foram implementadas em *Python*. Para garantir condições semelhantes, alguns parâmetros foram fixados, entretanto devido às características de cada abordagem, alguns parâmetros foram condicionados à essas particularidades. Conforme apresentado pela Tabela 4.4, os parâmetros fixos foram os mesmo utilizados em Alixandre e Dorn (2017), exceto pelo máximo de avaliações da função, onde foram utilizados os valores propostos pelo relatório de *benchmark* do CEC 2017 (AWAD M. Z. ALI; QU, 2016). Também de maneira semelhante ao aplicado por Alixandre e Dorn (2017), os testes com o BRKGA utilizaram os grupos Elite, Não-Elite e Mutação com os tamanhos 10%, 70% e 20%, respectivamente. No tamanho da migração e no número de *demes* novamente utilizaram-se valores como proposto em (ALIXANDRE; DORN, 2017), exceto pelas abordagens que utilizam as Políticas 4 e 6 (P4 propõem a migração de apenas um indivíduo e a P6 possui restrições quanto ao número de *demes* conforme apresentado pela Eq. 4.5).

Em nossos experimentos, todas as combinações de parâmetros foram executadas, buscando encontrar o melhor número de *demes* para cada abordagem em cada Função.

Tabela 4.4: Parâmetros dos experimentos - Etapa I. Parâmetros fixos são utilizados em todas as abordagens. Entretanto, alguns parâmetros variam de acordo com as características de cada política. O número de *demes* é um parâmetro a ser ajustado nos experimentos.

Parâmetros fixados		**Parâmetros condicionais**	
Arquitetura	*Ilha*	***Tamanho da migração***	• *10% da população para A1,A2,A4,A5,A6,A7,A8,A10,A11* • *1 indivíduo para A3, A9*
Taxa de migração	*25 gerações*		
Probabilidade de mutação	*1% - * Apenas para o AGD*	**Parâmetros a serem ajustados**	
Tamanho da população	*400*	***Número de demes***	• *5, 8, 16 para A1,A2,A3,A4,A6,A7,A8,A9,A10* • *5, 10 para A5, A11*
Máximo de avaliações da função	*10000 x D (dimensão)*		

Fonte: Do Autor.

Visando obter valores estatisticamente confiáveis, 30 execuções foram realizadas para cada um desses cenários (Abordagem $\times$ Função $\times$ Dimensão $\times$ número de *demes*). O resultado dos experimentos serão apresentados e discutidos no Capítulo 5, abordando cada cenário testado (Abordagem x Função x Dimensão) já com o número de *demes* ajustado, bem como o valor ideal em cada caso.

4.5 Experimentos - Etapa II

Esta seção apresentará a configuração dos testes dos operadores genéticos desenvolvidos, bem como métricas de avaliação adequadas ao problema PSP.

4.5.1 Componentes Testados

Nesta etapa de testes, o objetivo é analisar, de maneira construtiva, o impacto dos incrementos propostos. Devido ao fato de estarmos englobando componentes que foram desenvolvidos com base em informações do problema PSP, devemos considerar, nesse estágio de desenvolvimento do método, os valores encontrados na otimização juntamente com o aspecto estrutural das soluções.

Inicialmente, os modelos de inicialização foram testados, buscando definir a melhor opção de utilizar o conhecimento da base de fragmentos. Em seguida, os modelos de recombinação foram aplicados em conjunto com o modelo de inicialização escolhido, buscando encontrar a melhor combinação de maneira construtiva. Por fim, após definirmos os modelos de inicialização e recombinação, os modelos de mutação foram analisados, obtendo-se então a versão final do método proposto. A Tabela 4.5 apresenta os componentes propostos:

Tabela 4.5: Lista de incrementos propostos. Componentes desenvolvidos com base em conhecimento do domínio do problema.

Operador	Modelo	Descrição
Inicialização	Modelo I	Inserção estocástica. Segue a ordem decrescente de tamanho.
	Modelo II	Inserção por similaridade. Ambos os tamanhos são equiprováveis.
Recombinação	Uniforme	Processa cada resíduo individualmente.
	Por estrutura secundária	Processa cada trecho de possua o mesmo padrão.
	Mista	Começa pelo de estrutura secundária e após metade da execução utiliza o uniforme.
Mutação	Modelo I	Processa cada trecho irregular individualmente.
	Modelo II	Processa todas as regiões irregulares. Usa apenas com fragmentos inteiros.

Fonte: Do Autor.

4.5.2 Métricas de Avaliação

Sabemos que o problema PSP tem por característica buscar a minimização da função de energia e, conforme apresentado na Seção 2.2.3, utilizaremos a função *Score3* do *Rosetta*. Todavia, além de considerarmos os valores de finais de energia obtidos, devemos também analisar a qualidade estrutural dos resultados.

Uma métrica comumente utilizada é o desvio médio quadrático - RMSD (*Root Mean Square Deviation*) (ZHANG; SKOLNICK, 2004). O RMSD pode ser calculado utilizando todos os átomos de uma proteína (ENGH; HUBER, 1991) ou considerando apenas um subconjunto de átomos (e.g., C_α de cada resíduo) que, segundo (KABSCH; SANDER, 1983) é uma maneira eficiente para se expressar a similaridade da conformação global de duas proteínas, onde quanto menor seu valor, mais similaridade há entre as estruturas. A Equação 4.13 descreve esta métrica, considerando que as duas estruturas em questão já tenham sidas previamente sobrepostas de maneira ótima.

$$RMSD(a,b) : \sqrt{\frac{\sum_{i=1}^{n} \|v_{ai} - v_{bi}\|}{n}} \tag{4.13}$$

onde *a* e *b* são as duas estruturas a serem comparadas, n é o total de resíduos na sequência de aminoácidos, v_{ai} e v_{bi} representa o posicionamento do *i*-ésimo átomo dos vetores de posições cartesianas v_a e v_b, respectivamente contidos nas estruturas *a* e *b*.

4.5.3 Configuração dos Experimentos

Todos os componentes mantiveram o padrão da etapa anterior e foram implementados em *Python*. Conforme apresentado, os componentes foram testados na seguinte ordem: inicialização, recombinação e mutação. Para avaliarmos a inicialização, um processo de amostragem foi realizado com cada modelo proposto. Dessa forma, 100000 indivíduos foram criados de maneira aleatória, no intuito de avaliarmos a qualidade média dos indivíduos gerados, bem como a capacidade dos modelos de obter soluções com baixos valores de RMSD e energia. Em relação aos modelos de recombinação e mutação, um conjunto de 6 proteínas foi utilizado (Ver Tab. 4.6), onde executou-se 12 vezes cada combinação de componente/proteína em conjunto com o método selecionado na etapa anterior. Como critério de parada, utilizou-se um valor máximo de avaliações da função de energia, permitindo 1000000 por execução.

Tabela 4.6: Lista de proteínas para testes. Etapa II.

ID-Proteína	Tamanho	Conteúdo da estrutura secundária
2P5K	64	1 folha / 3 hélices
2MR9	44	3 hélices
1ACW	29	1 folha / 1 hélice
1AB1	46	1 folha / 2 hélices
1K43	14	1 folha
1ZDD	34	2 hélices

Fonte: Do Autor.

Em relação aos parâmetros da abordagem de otimização, mantiveram-se os mesmos da etapa anterior. Nos caso do tamanho da migração, será utilizado o valor correspondente à abordagem escolhida na etapa I, além do número de *demes* que também será definido de acordo com o melhor valor baseado nos testes já realizados. A Tabela 4.7 apresenta os parâmetros fixos.

4.6 Experimentos - Etapa III

Para podermos comparar a abordagem proposta em relação aos dois métodos apresentados, utilizou-se o mesmo conjunto de proteínas testados em Corrêa et al. (2016). A Tabela 4.9 apresenta esse conjunto em detalhes. Apesar das diferentes características dos métodos apresentados, buscou-se equilibrar, dentro do possível, suas condições de execução. Seguindo o parâmetro utilizado pelo método M5, o número máximo de avaliações

Tabela 4.7: Parâmetros dos experimentos - Etapa II. Parâmetros utilizados em todos os testes de incrementos. Demais parâmetros estão condicionados à escolha da abordagem na etapa I.

Parâmetros	
Arquitetura	*Ilha*
Taxa de migração	*25 gerações*
Tamanho da população	*400*
Máximo de avaliações da função de energia	*1000000*

Fonte: Do Autor.

da função de energia foi de 1000000. Além disso, devido ao fato de que M5 e *Rosetta* utilizam informação da estrutura secundária obtidas de maneiras diferentes (atribuída e predita, respectivamente), a abordagem proposta foi testada com cada um desses cenários. A tabela 4.8 sumariza esta situação. Cada método foi executado 30 vezes visando uma maior precisão estatística. A abordagem proposta seguiu a mesma parametrização utilizada na Etapa II, variando apenas a informação da estrutura secundária (atribuída e predita).

Tabela 4.8: Informação da estrutura secundária dos métodos M5 e *Rosetta*.

Método	**Estrutura secundária**	**Modo de obtenção**
M5 (CORRÊA et al., 2016)	Atribuída	STRIDE
Rosetta (ROHL et al., 2004)	Predita	Psipred

Fonte: Do Autor.

4.6.1 Métricas de Avaliação

Ao submetermos o método desenvolvido aos testes das etapas anteriores, pôde-se avaliar suas características sob uma perspectiva geral. Uma vez finalizado, avaliou-se o método proposto principalmente em função de sua qualidade estrutural, que pode ser obtida por meio do cálculo de RMSD conforme utilizado na etapa anterior. Outra métrica de similaridade entre estruturas amplamente utilizada é o GDT_TS (*Global Distance Test - Total Score*). Esta métrica, de acordo com Kufareva e Abagyan (2011), é mais robusta que as demais por ser menos sensível à discrepâncias em regiões de estruturas secundárias irregulares. Expressa em uma escala que vai de 0 a 100% de similaridade, o GDT é

Tabela 4.9: Lista de proteínas para testes. Etapa III.

ID-Proteína	Tamanho	Conteúdo da estrutura secundária
1AB1	46	1 folha / 2 hélices
1ACW	29	1 folha / 1 hélice
1CRN	46	1 folha / 2 hélices
1D5Q	27	1 folha / 1 hélice
1ENH	54	3 hélices
1K43	14	1 folha
1L2Y	20	2 hélices
1Q2K	31	1 folha / 1 hélice
1ROP	63	2 hélices
1UTG	70	5 hélices
1WQC	26	2 hélices
1ZDD	35	2 hélices
2MR9	44	3 hélices
2MTW	20	1 hélice
2P5K	64	1 folha / 3 hélices
2P6J	52	3 hélices
2P81	44	2 hélices
2PMR	87	3 hélices
3V1A	48	2 hélices

Fonte: Do Autor.

calculado conforme apresentado pela Equação 4.14:

$$GDT : \frac{GDT_{p1} + GDT_{p2} + GDT_{p4} + GDT_{p8}}{4} \tag{4.14}$$

onde GDT_{p_n} representa a porcentagem de resíduos com distância menor que *n*Å entre as estruturas avaliadas. Assume-se que tais estruturas já estejam em uma sobreposição ótima.

Devido ao fator estocástico das abordagens utilizadas, utilizou-se o teste estatístico *Wilcoxon-Mann-Whitney* para complementar a análise dos resultados obtidos. Este é um teste não-paramétrico para amostras independentes que possui a capacidade de avaliar a heterogeneidade de duas amostras. Sua hipótese nula afirma que a distribuição de ambos os grupos são iguais, ao passo que a hipótese alternativa afirma que as amostras provém de populações distintas, rejeitando a possibilidade de igualdade das medianas.

4.7 Resumo do Capítulo

Neste capítulo apresentamos a metodologia do desenvolvimento da abordagem proposta bem como as técnicas utilizadas para comparação. Inicialmente, um modelo de política de migração com foco em manutenção e exploração de diversidade foi proposto e avaliado em um ambiente genérico, sem viés do domínio do problema. À medida que uma técnica de otimização foi estabelecida, uma série de incrementos baseados no problema PSP foram propostos e testados de modo incremental. Por fim, a abordagem final foi comparada com os métodos M5 (CORRÊA et al., 2016) e *Rosetta* (ROHL et al., 2004), ambos para predição de estrutura tridimensional de proteínas. No próximo capítulo serão apresentados os resultados obtidos em cada uma das etapas do desenvolvimento.

5 RESULTADOS E DISCUSSÃO

5.1 Resultados - Etapa I

Os resultados das abordagens apresentadas foram analisados em relação ao menor erro encontrado, além de fatores como diversidade, convergência e complexidade. Estas análises foram realizadas em duas etapas: a primeira considerando as abordagens A1, A2, A3, A4 e A5, que representam as políticas utilizando o AG canônico; e a segunda, que engloba A6, A7, A8, A9, A10 e A11, na qual utiliza-se o BRKGA. Ressaltamos que os testes com o AG canônico possuem uma abordagem a menos devido a Política 2 se aplicar apenas ao BRKGA. Os resultados completos, juntamente com todos os gráficos de convergência podem ser consultados no material suplementar através do link <<http://sbcb.inf.ufrgs.br/dbrkga.html>>.

5.1.1 Erro e Diversidade

As Tabelas 5.1 e 5.2 apresentam os erros (Eq. 4.8) e os valores de diversidade (Eq. 4.9) das políticas aplicadas no AG canônico, destacando-se os menores erros e os maiores valores de diversidade em verde e amarelo respectivamente.

Em relação ao erro, pode-se notar que a abordagem A5 (que utiliza a política proposta) foi quem obteve os melhores resultados (e.g., Funções 6, 9, 16, 17, 24, 26). Seguida por A1 que se destacou em alguns casos, especialmente nas funções de menor complexidade como F1 e F4. Devido ao fato de A1 utilizar a política de migração canônica (sem nenhum controle de diversidade), sua convergência ocorreu de maneira prematura em relação às demais. Dessa forma, ao ser aplicada em funções de menor complexidade, foi possível convergir para resultados melhores enquanto as demais abordagens mantiveram a diversidade para melhor explorar o espaço de busca e consequentemente desaceleraram a convergência. Já em funções mais complexas, principalmente apresentadas pela Tabela 5.2, este mesmo comportamento resulta em valores melhores para A5, uma vez que neste cenário a capacidade de explorar o espaço de busca é crucial para escapar de ótimos locais.

Ao considerarmos os valores de diversidade, A3 alcançou os melhores níveis em 90% dos casos. Isso se deve principalmente ao fato da política utilizada em A3 migrar apenas um indivíduo por vez, enquanto as demais migram 10% da população. Uma vez

que mais indivíduos são enviados na migração, mais rápido essas soluções irão se espalhar por toda a população, acelerando a convergência. Isso, combinado ao fato do AG canônico não possuir controle de diversidade, culminou em uma perda de diversidade mais acelerada do que o ocorrido em A3.

Já nos casos utilizando o BRKGA, apresentados pelas Tabelas 5.3 e 5.4, nota-se maior equilíbrio entre as abordagens testadas, com destaque para A11 que alcançou melhores resultados em 12 casos, seguido por A7 e A8 com 8 e 6 casos respectivamente. Este equilíbrio se deve à característica de manutenção de diversidade do BRKGA, que permitiu uma melhor exploração do potencial de cada abordagem testada. Apesar disso, novamente a abordagem que utilizou a política proposta (A11) foi superior, destacando-se nas funções mais complexas. Se considerarmos os resultados de A11 juntamente com A7, podemos observar que as abordagens com políticas baseadas no BRKGA foram majoritariamente superiores às demais (baseadas somente em *fitness* e em *fitness*+similaridade).

Em relação à diversidade, A11 obteve resultados superiores em 75% dos casos, seguida por A9 com quase 25%. Este comportamento era esperado uma vez que A9 utiliza a Política 3, que já havia mostrado bons resultados. Entretanto a política proposta foi baseada no BRKGA e com principal foco na exploração de diversidade, levando A11 a alcançar valores superiores de diversidade na grande maioria dos casos.

Em suma, podemos observar que A5 e A11, ambas utilizando a política proposta, alcançaram melhores resultados em nossos testes, lidando melhor com funções mais complexas (e.g., Funç. 17, 24 e 26 possuem o maior número de funções básicas em sua composição). Em relação às abordagens que utilizam a política canônica P1, A1 obteve sucesso em algumas funções de menor complexidade (básicas e/ou unimodais), entretanto A6 não conseguiu repetir os mesmos resultados. Ao observarmos as abordagens com políticas focadas em diversidade, temos resultados relevantes apenas quando aplicadas ao BRKGA. A8, A9 e A10 até conseguiram ser competitivas, entretanto suas versões com AG canônico (A2, A3 e A4) não obtiveram o mesmo desempenho. Por fim, A7 também provou seu potencial ao alcançar os menores erros em alguns casos, todavia sua limitação de ser vinculada diretamente ao BRKGA restringe sua utilização.

Tabela 5.1: Resultados dos experimentos com A1-A5 (políticas com o AG canônico) para as funções F1, F4, F6, F9 e F11. *D* representa a dimensão da função. *Erro* representa a média das diferenças entre o valor ótimo e o valor encontrado; e *div* representa a média da diversidade final da população, ambos em 30 execuções.

		A 1		**A 2**		**A 3**		**A 4**		**A 5**	
	D	Erro	div.	Erro	div.	Erro	div.	Erro	div.	Erro	div.
F1	10	2,683E+08	6,990E+01	2,645E+08	8,224E+01	5,674E+08	8,520E+01	3,238E+08	**8,983E+01**	**1,573E+08**	2,001E+01
	30	**1,739E+09**	6,085E+01	1,959E+09	6,990E+01	3,880E+09	**1,094E+02**	1,858E+09	7,032E+01	2,357E+09	7,822E+01
	50	**4,948E+09**	6,881E+01	5,166E+09	7,116E+01	1,075E+10	**1,390E+02**	5,096E+09	7,158E+01	7,829E+09	7,881E+01
	100	**1,035E+10**	7,497E+01	1,088E+10	7,887E+01	2,748E+10	**1,899E+02**	1,159E+10	7,543E+01	2,143E+10	1,247E+02
F4	10	4,566E+01	7,929E+01	4,280E+01	7,911E+01	5,414E+01	**9,405E+01**	4,569E+01	8,305E+01	**3,281E+01**	4,154E+01
	30	4,093E+02	1,063E+02	3,993E+02	1,094E+02	5,891E+02	**1,558E+02**	4,024E+02	1,071E+02	**3,584E+02**	2,633E+01
	50	**8,930E+02**	1,154E+02	9,276E+02	1,088E+02	1,665E+03	**1,957E+02**	9,060E+02	1,086E+02	9,808E+02	4,114E+01
	100	**1,655E+03**	8,987E+01	1,763E+03	1,003E+02	3,129E+03	**2,508E+02**	1,703E+03	9,546E+01	2,270E+03	5,474E+01
F6	10	5,808E+00	7,461E+01	6,317E+00	**8,877E+01**	7,935E+00	8,465E+01	6,237E+00	7,538E+01	**4,858E+00**	6,851E+01
	30	9,189E+00	7,600E+01	9,018E+00	7,849E+01	1,217E+01	**9,917E+01**	8,636E+00	7,671E+01	**7,848E+00**	7,222E+01
	50	9,973E+00	6,589E+01	9,891E+00	7,079E+01	1,299E+01	**1,091E+02**	1,012E+01	6,884E+01	**8,999E+00**	8,479E+01
	100	1,179E+01	7,103E+01	1,134E+01	7,119E+01	1,444E+01	**1,374E+02**	1,140E+01	7,021E+01	**9,440E+00**	9,100E+01
F9	10	1,486E+02	6,568E+01	1,538E+02	5,926E+01	2,368E+02	**7,701E+01**	1,802E+02	6,404E+01	**1,149E+02**	1,906E+01
	30	2,163E+03	1,079E+02	2,117E+03	1,056E+02	2,760E+03	**1,889E+02**	1,991E+03	1,063E+02	**1,191E+03**	7,694E+01
	50	6,487E+03	1,277E+02	6,393E+03	1,297E+02	9,667E+03	**2,738E+02**	6,456E+03	1,218E+02	**4,042E+03**	4,732E+01
	100	2,279E+04	1,254E+02	2,422E+04	1,266E+02	2,688E+04	**4,115E+02**	2,509E+04	1,440E+02	**1,534E+04**	6,029E+01
F11	10	**1,626E+02**	1,170E+02	1,876E+02	1,111E+02	2,222E+02	**1,353E+02**	1,736E+02	1,202E+02	2,437E+02	4,904E+01
	30	2,536E+03	1,468E+02	**1,731E+03**	1,379E+02	2,811E+03	**1,998E+02**	2,534E+03	1,597E+02	4,727E+03	1,244E+02
	50	**9,840E+03**	1,874E+02	9,937E+03	1,858E+02	1,290E+04	**2,634E+02**	9,953E+03	1,999E+02	1,169E+04	1,247E+02
	100	**7,639E+04**	2,176E+02	7,783E+04	2,260E+02	1,066E+05	**4,824E+02**	8,579E+04	2,419E+02	1,197E+05	6,567E+01

Fonte: Do Autor.

Tabela 5.2: Resultados dos experimentos com A1-A5 (políticas com o AG canônico) para as funções F16, F17, F22, F24 e F26. *D* representa a dimensão da função. *Erro* representa a média das diferenças entre o valor ótimo e o valor encontrado; e *div* representa a média da diversidade final da população, ambos em 30 execuções.

		A 1		**A 2**		**A 3**		**A 4**		**A 5**	
	D	Erro	div.	Erro	div.	Erro	div.	Erro	div.	Erro	div.
F16	10	1,997E+02	1,396E+02	2,165E+02	1,172E+02	2,534E+02	**1,472E+02**	1,959E+02	1,152E+02	**1,751E+02**	4,776E+01
	30	1,354E+03	1,434E+02	1,291E+03	1,472E+02	1,494E+03	**1,828E+02**	1,321E+03	1,585E+02	**1,217E+03**	1,648E+01
	50	2,357E+03	1,566E+02	2,275E+03	1,581E+02	2,602E+03	**2,042E+02**	2,329E+03	1,581E+02	**2,243E+03**	5,510E+01
	100	5,910E+03	1,492E+02	6,198E+03	1,621E+02	7,362E+03	**4,860E+02**	6,167E+03	1,480E+02	**5,393E+03**	6,461E+01
F17	10	1,070E+02	1,356E+02	1,007E+02	1,191E+02	1,189E+02	**1,671E+02**	9,208E+01	1,353E+02	**8,345E+01**	7,397E+01
	30	7,072E+02	1,804E+02	7,976E+02	1,878E+02	9,386E+02	**2,971E+02**	7,880E+02	2,159E+02	**6,507E+02**	4,616E+01
	50	2,018E+03	3,142E+02	2,029E+03	2,585E+02	2,193E+03	**3,759E+02**	2,100E+03	3,302E+02	**1,867E+03**	4,511E+01
	100	4,537E+03	1,605E+02	4,707E+03	1,906E+02	5,516E+03	**4,500E+02**	4,667E+03	1,961E+02	**4,536E+03**	2,177E+01
F22	10	1,416E+02	1,014E+02	**1,306E+02**	**1,162E+02**	1,677E+02	9,999E+01	1,389E+02	9,518E+01	1,497E+02	4,980E+01
	30	**5,727E+02**	7,650E+01	8,875E+02	1,156E+02	1,161E+03	**1,443E+02**	8,299E+02	8,120E+01	7,349E+02	6,571E+01
	50	8,695E+03	1,051E+02	7,894E+03	1,214E+02	9,257E+03	**4,169E+02**	8,969E+03	1,401E+02	**7,034E+03**	1,861E+01
	100	1,984E+04	5,633E+02	2,003E+04	5,588E+02	2,002E+04	**5,877E+02**	2,002E+04	1,215E+02	**1,621E+04**	5,478E+00
F24	10	**3,170E+02**	1,219E+02	3,401E+02	1,213E+02	3,558E+02	1,294E+02	3,374E+02	**1,346E+02**	3,442E+02	4,431E+01
	30	6,088E+02	1,397E+02	6,152E+02	1,394E+02	6,409E+02	**1,564E+02**	6,197E+02	1,389E+02	**5,368E+02**	4,203E+00
	50	9,116E+02	1,956E+02	9,349E+02	1,963E+02	9,798E+02	**2,299E+02**	9,404E+02	1,196E+02	**7,726E+02**	5,563E+00
	100	1,735E+03	9,330E+01	1,711E+03	9,059E+01	1,887E+03	**2,549E+02**	1,740E+03	9,480E+01	**1,689E+03**	1,116E+02
F26	10	5,040E+02	8,795E+01	5,017E+02	9,807E+01	5,316E+02	**1,079E+02**	5,106E+02	8,486E+01	**4,853E+02**	3,346E+01
	30	2,925E+03	1,189E+02	2,995E+03	1,121E+02	3,412E+03	**1,834E+02**	3,030E+03	1,149E+02	**2,742E+03**	4,456E+01
	50	4,944E+03	1,034E+02	4,855E+03	1,023E+02	5,524E+03	**2,044E+02**	4,985E+03	1,009E+02	**4,383E+03**	2,685E+01
	100	1,225E+04	9,154E+01	1,216E+04	9,046E+01	1,412E+04	**2,773E+02**	1,217E+04	9,216E+01	**1,207E+04**	9,866E+01

Fonte: Do Autor.

Tabela 5.3: Resultados dos experimentos com A6-A11 (políticas com o BRKGA) para as funções F1, F4, F6, F9 e F11. *D* representa a dimensão da função. *Erro* representa a média das diferenças entre o valor ótimo e o valor encontrado; e *div* representa a média da diversidade final da população, ambos em 30 execuções.

		A 6		**A 7**		**A 8**		**A 9**		**A 10**		**A 11**	
	D	Erro	div.	Erro	div.	Erro	div.	Erro	div.	Erro	div.	Erro	div.
F1	10	2,290E+05	1,120E+02	2,404E+05	1,119E+02	**1,546E+05**	1,130E+02	3,102E+05	1,152E+02	1,721E+05	1,140E+02	2,897E+05	**1,193E+02**
	30	1,281E+06	1,761E+02	1,499E+06	1,768E+02	**1,232E+06**	1,765E+02	1,269E+06	1,742E+02	1,293E+06	1,755E+02	2,652E+06	**1,915E+02**
	50	4,236E+06	2,271E+02	4,260E+06	2,269E+02	4,088E+06	2,289E+02	**4,082E+06**	2,257E+02	4,148E+06	2,286E+02	9,195E+06	**2,450E+02**
	100	1,338E+07	3,124E+02	1,372E+07	3,097E+02	1,384E+07	3,143E+02	1,377E+07	3,139E+02	**1,228E+07**	3,149E+02	4,292E+07	**3,995E+02**
F4	10	4,429E+00	1,190E+02	4,413E+00	**1,222E+02**	4,337E+00	1,099E+02	4,435E+00	1,220E+02	**3,886E+00**	1,212E+02	5,810E+00	1,178E+02
	30	9,898E+01	1,809E+02	9,826E+01	1,933E+02	**9,586E+01**	1,940E+02	9,917E+01	2,027E+02	1,020E+02	1,869E+02	1,077E+02	**2,139E+02**
	50	1,785E+02	2,465E+02	**1,701E+02**	2,461E+02	1,714E+02	2,450E+02	1,771E+02	2,442E+02	1,849E+02	2,467E+02	2,025E+02	**2,837E+02**
	100	3,408E+02	3,241E+02	3,213E+02	3,293E+02	3,446E+02	3,255E+02	3,308E+02	3,346E+02	**3,209E+02**	3,356E+02	3,800E+02	**4,260E+02**
F6	10	2,565E-01	1,097E+02	**2,345E-01**	1,093E+02	2,476E-01	1,096E+02	2,747E-01	1,088E+02	2,603E-01	1,097E+02	2,686E-01	**1,203E+02**
	30	4,752E-01	1,907E+02	4,806E-01	1,892E+02	4,239E-01	1,902E+02	4,452E-01	1,878E+02	**4,000E-01**	1,894E+02	5,688E-01	**2,088E+02**
	50	5,620E-01	2,397E+02	5,627E-01	2,383E+02	**5,325E-01**	2,391E+02	5,856E-01	2,389E+02	5,625E-01	2,376E+02	7,516E-01	**2,635E+02**
	100	7,261E-01	3,199E+02	7,412E-01	3,202E+02	**7,185E-01**	3,239E+02	7,417E-01	3,234E+02	7,331E-01	3,249E+02	1,038E+00	**4,225E+02**
F9	10	3,862E-01	9,411E+01	**3,232E-01**	9,276E+01	4,756E-01	9,397E+01	4,081E-01	9,395E+01	5,309E-01	9,391E+01	5,585E-01	**1,019E+02**
	30	1,454E+02	1,754E+02	1,052E+02	1,681E+02	1,334E+02	1,662E+02	9,442E+01	1,769E+02	9,216E+01	1,724E+02	**5,791E+01**	**1,872E+02**
	50	9,184E+02	2,416E+02	7,177E+02	2,367E+02	8,311E+02	2,351E+02	7,409E+02	2,344E+02	7,147E+02	2,376E+02	**4,240E+02**	**2,607E+02**
	100	5,949E+03	3,272E+02	5,052E+03	3,225E+02	5,705E+03	3,257E+02	5,304E+03	3,284E+02	5,307E+03	3,270E+02	**2,389E+03**	**4,249E+02**
F11	10	9,440E+00	1,007E+02	7,755E+00	1,041E+02	8,825E+00	1,033E+02	**5,897E+00**	1,029E+02	7,939E+00	1,017E+02	8,309E+00	**1,294E+02**
	30	2,569E+02	1,759E+02	2,454E+02	1,816E+02	2,621E+02	1,758E+02	2,710E+02	1,897E+02	2,581E+02	1,800E+02	**1,838E+02**	**2,365E+02**
	50	5,634E+02	2,386E+02	7,140E+02	2,408E+02	6,569E+02	2,371E+02	6,512E+02	2,349E+02	**4,876E+02**	2,463E+02	5,180E+02	**3,026E+02**
	100	**1,015E+04**	3,309E+02	1,096E+04	3,309E+02	1,187E+04	3,466E+02	1,024E+04	3,455E+02	1,054E+04	3,382E+02	1,387E+04	**5,070E+02**

Fonte: Do Autor.

Tabela 5.4: Resultados dos experimentos com A6-A11 (políticas com o BRKGA) para as funções F16, F17, F22, F24 e F26. *D* representa a dimensão da função. *Erro* representa a média das diferenças entre o valor ótimo e o valor encontrado; e *div* representa a média da diversidade final da população, ambos em 30 execuções.

		A 6		**A 7**		**A 8**		**A 9**		**A 10**		**A 11**	
	D	Erro	div.	Erro	div.	Erro	div.	Erro	div.	Erro	div.	Erro	div.
F16	10	7,123E+00	1,295E+02	1,062E+01	1,198E+02	**6,777E+00**	1,290E+02	1,449E+01	**1,375E+02**	8,540E+00	1,312E+02	1,212E+01	1,332E+02
	30	6,716E+02	2,250E+02	**6,516E+02**	2,379E+02	6,915E+02	2,208E+02	6,886E+02	**2,841E+02**	7,130E+02	2,303E+02	7,579E+02	2,559E+02
	50	1,356E+03	2,839E+02	1,316E+03	3,021E+02	1,370E+03	2,441E+02	**1,298E+03**	2,970E+02	1,405E+03	2,870E+02	1,493E+03	**3,214E+02**
	100	3,623E+03	3,802E+02	**3,351E+03**	4,097E+02	3,549E+03	3,918E+02	3,518E+03	**5,021E+02**	3,462E+03	3,948E+02	3,984E+03	4,983E+02
F17	10	**3,843E+00**	1,163E+02	4,559E+00	1,130E+02	4,985E+00	1,149E+02	5,523E+00	1,256E+02	4,100E+00	1,217E+02	4,650E+00	**1,321E+02**
	30	1,887E+02	2,317E+02	1,877E+02	2,441E+02	1,994E+02	2,342E+02	2,204E+02	**3,046E+02**	2,020E+02	2,326E+02	**1,798E+02**	2,681E+02
	50	9,393E+02	3,037E+02	**8,921E+02**	3,320E+02	9,033E+02	3,015E+02	9,372E+02	**3,865E+02**	9,760E+02	3,090E+02	1,004E+03	3,411E+02
	100	2,641E+03	4,117E+02	2,524E+03	4,413E+02	2,663E+03	4,097E+02	**2,520E+03**	**5,440E+02**	2,625E+03	4,318E+02	2,774E+03	5,101E+02
F22	10	**6,866E+01**	1,249E+02	7,847E+01	1,208E+02	7,454E+01	1,226E+02	7,419E+01	**1,616E+02**	7,269E+01	1,254E+02	8,251E+01	1,371E+02
	30	**1,117E+02**	1,756E+02	1,119E+02	1,749E+02	1,120E+02	1,757E+02	1,120E+02	1,754E+02	1,131E+02	1,773E+02	1,141E+02	**1,908E+02**
	50	4,260E+03	2,543E+02	4,281E+03	2,493E+02	4,499E+03	2,577E+02	3,946E+03	2,504E+02	4,332E+03	2,531E+02	**3,852E+03**	**3,276E+02**
	100	**1,146E+04**	3,505E+02	1,147E+04	3,483E+02	1,175E+04	3,508E+02	1,174E+04	3,540E+02	1,175E+04	3,514E+02	1,225E+04	**4,744E+02**
F24	10	1,643E+02	1,494E+02	1,702E+02	1,479E+02	1,691E+02	1,339E+02	1,804E+02	**1,824E+02**	1,654E+02	1,387E+02	**1,260E+02**	1,761E+02
	30	4,870E+02	2,198E+02	4,836E+02	2,170E+02	4,873E+02	2,219E+02	4,852E+02	2,165E+02	4,867E+02	2,221E+02	**4,726E+02**	**2,536E+02**
	50	6,343E+02	2,768E+02	6,348E+02	2,776E+02	6,419E+02	2,775E+02	6,370E+02	2,760E+02	6,407E+02	2,766E+02	**6,154E+02**	**3,295E+02**
	100	1,267E+03	3,748E+02	1,252E+03	3,775E+02	1,265E+03	3,785E+02	1,242E+03	3,839E+02	1,256E+03	3,788E+02	**1,225E+03**	**4,920E+02**
F26	10	2,451E+02	1,078E+02	**2,245E+02**	1,281E+02	2,945E+02	1,080E+02	2,890E+02	**1,432E+02**	2,488E+02	1,202E+02	2,682E+02	1,309E+02
	30	1,204E+03	1,813E+02	**1,110E+03**	1,838E+02	1,274E+03	1,818E+02	1,413E+03	2,085E+02	1,137E+03	1,927E+02	1,551E+03	**2,226E+02**
	50	2,666E+03	2,557E+02	2,575E+03	2,561E+02	2,679E+03	2,551E+02	2,646E+03	2,565E+02	2,666E+03	2,557E+02	**2,507E+03**	**2,863E+02**
	100	7,321E+03	3,614E+02	7,117E+03	3,608E+02	7,309E+03	3,614E+02	7,436E+03	3,649E+02	7,268E+03	3,594E+02	**6,941E+03**	**4,729E+02**

Fonte: Do Autor.

5.1.2 Ajustando o Número de *Demes*

Como apresentado na Seção 4.4.4, o número de *demes* foi ajustado, buscando o valor que resultasse em menor erro para cada função em cada abordagem. A Tabela 5.5 mostra esses valores. Lembramos que, devido a restrições quanto ao número de *demes*, as abordagens que utilizam a Política 6 (A5 e A11) foram testadas para os casos com 5 e 10 *demes*, enquanto as demais utilizaram 4, 8 e 16.

Ao analisarmos os resultados, podemos perceber que as abordagens usando o AG canônico (de A1 até A5) tiverem melhor desempenho com uma maior quantidade de *demes*. Isso se deve, de modo geral, ao fato de o AG canônico não possuir controle de diversidade, um número maior de *demes* possibilitou um retardo na convergência, e uma melhor exploração do espaço de busca.

Em relação às abordagens usando o BRKGA (A6 até A11), notamos dois comportamentos. O primeiro diz respeito às funções menos complexas, onde houve uma predominância de um menor número de *demes*. No segundo, a situação se inverte, havendo uma preferência por um número maior de *demes*. Pode-se observar que, no primeiro caso, o controle de diversidade do BRKGA foi suficiente para evitar a convergência prematura e, consequentemente, alcançar um equilíbrio com um menor número de *demes*. Entretanto, quando aplicadas em funções mais complexas, A6, A7, A8 e A10 foram melhores com 16 *demes*, devido ao espaço de busca com múltiplos ótimos locais, sendo tal cenário muito complexo para 4 e 8 *demes*. Curiosamente, nestas mesmas funções, A9 obteve melhores resultados com 4 *demes*, o que nos leva a crer que a migração de apenas um indivíduo por vez (Política 4), juntamente com o BRKGA, foi suficiente para evitar a convergência prematura e lidar com tais funções. Por fim, A11 alcançou melhores resultados com 5 *demes* em quase todos os testes. Como já explicado, A11 utiliza a política proposta, que tem como base o controle da diversidade. Esse aspecto, em conjunto com o BRKGA, que também emprega tais princípios, permitiu com que a abordagem lidasse com os obstáculos dos espaços de busca sem a necessidade de se dividir a população global em uma maior quantidade de *demes*.

5.1.3 Curva de Convergência

Devido ao número de abordagens e casos testados (11 abordagens $\times$ 10 funções $\times$ 4 dimensões), optou-se por apresentar apenas alguns casos variados daquelas que ob-

Tabela 5.5: O melhor número de *demes* para cada abordagem em cada função.
Fonte: Do autor.

	A 1	**A 2**	**A 3**	**A 4**	**A 5**	**A 6**	**A 7**	**A 8**	**A 9**	**A 10**	**A 11**
F1	16	16	16	16	10	4	4	4	4	4	5
F4	16	16	16	16	10	16	4	8	8	16	5
F6	16	16	16	16	10	4	4	4	4	4	5
F9	16	16	4	16	10	8	4	4	4	8	5
F11	16	16	16	16	10	8	8	8	4	4	10
F16	16	8	16	8	10	16	16	16	16	16	5
F17	16	16	16	16	10	16	16	16	16	16	5
F22	16	16	4	16	10	16	4	16	4	16	10
F24	4	4	4	16	10	16	16	16	4	4	5
F26	16	16	16	16	10	4	16	16	4	16	5
Mode	16	16	16	16	10	16	4	16	4	16	5

tiveram os melhores resultados quanto ao erro. A Figura 5.1 apresenta a curva de convergência de A11, A7 e A8 para as funções F9, F11, F17 e F26 nas dimensões 100, 50, 30 e 30 respectivamente. Estes casos foram escolhidos visando representar diferentes aspectos presentes nos testes realizados e, conforme anteriormente explicado, o restante dos gráficos podem ser consultados no material suplementar.

Como pode-se notar, na Função 9 com 100 dimensões, A11 (Fig. 5.1a) possui um caminho de otimização com menor variação que A7 e A8 (Fig. 5.1b e 5.1c). Além disso, a área abaixo das curvas de A11 e A7 são notoriamente menores do que em A8, o que caracteriza uma convergência mais rápida. Tal comportamento poderia contradizer a premissa de se evitar a convergência prematura, entretanto, juntamente com os resultados apresentados na Seção 5.1.1, percebe-se que de fato A11 e A7 convergiram de modo mais adequado que A8, alcançando menores erros de modo mais rápido. Comportamento semelhante ocorre na Função 11 com 50 dimensões. Pode-se observar que, apesar das três abordagem apresentarem curvas coesas, A8 (Fig. 5.1f) está em uma escala aproximadamente duas vezes maior que A7 e três vezes maior que A8 (Fig. 5.1e e 5.1d).

Ao observarmos a Função 17 com 30 dimensões, nota-se que todas as 3 abordagens tiveram dificuldades em manter curvas com pouca variação. Isso se deve ao fato desta função possuir um espaço de busca mais complexo, deixando as abordagens presas em ótimos locais. Apesar disso, A11 (Fig. 5.1g) se mostrou mais robusta que A7 e A8 (Fig. 5.1h e 5.1i), concentrando seus resultados em um intervalo de *fitness* menor que as demais (e.g., $[1700, 2100]$ em A11; $[1700, 2250]$ em A7 e A8). Na Função 26, A11 e A7 (Fig. 5.1j e 5.1k) novamente possuem curvas semelhantes, enquanto A8 (Fig. 5.1l)

enfrenta maior dificuldade em lidar com ótimos locais, resultando em um caminho de otimização altamente variado.

Podemos perceber que à medida em que a complexidade das funções aumentam, todas as abordagens encontram certas dificuldades. Entretanto, pode-se observar que A11 atua de forma mais determinística, resultando em um caminho de otimização coeso. Como resultado, temos uma abordagem capaz de escapar de ótimos locais sem atrasar a convergência da população global.

5.1.4 Complexidade

Conforme apresentado pela Equação 4.11, a complexidade foi calculada para cada abordagem e ponderada pelo número de *demes*. Os resultados apresentados pela Figura 5.2 foram obtidos utilizando 5 *demes* em A5 e A11, e 4 *demes* nas demais. Na Figura 5.2a temos os valores referentes às abordagens utilizando AG canônico e, na Figura 5.2b, os valores utilizando o BRKGA.

Em relação às abordagens utilizando o AG canônico, exceto por A2, que possui complexidade aproximadamente duas vezes maior, notamos que as demais abordagens possuem resultados semelhantes. Podemos notar ainda uma leve vantagem de A5, que utiliza a política proposta. Ao observarmos os resultados obtidos com o BRKGA, temos o mesmo padrão, com A8 atingindo a maior complexidade (que utiliza a mesma política que A2), e as demais com valores menores. Novamente a abordagem utilizando a política proposta (A11) obteve uma sutil vantagem em relação às demais.

O fato de se basear apenas em *fitness* permite à política proposta operar com um menor custo computacional, sem a necessidade de computar a similaridade entre indivíduos. Políticas como P3, P4 e P5 utilizam a informação da similaridade para a tomada de decisão. P3, que foi utilizada por A2 e A8, além de calcular esta similaridade entre todos os indivíduos da população, ainda considera a similaridade de indivíduos próximos ao indivíduo médio para realizar o processo de migração, fazendo desta a política de maior complexidade.

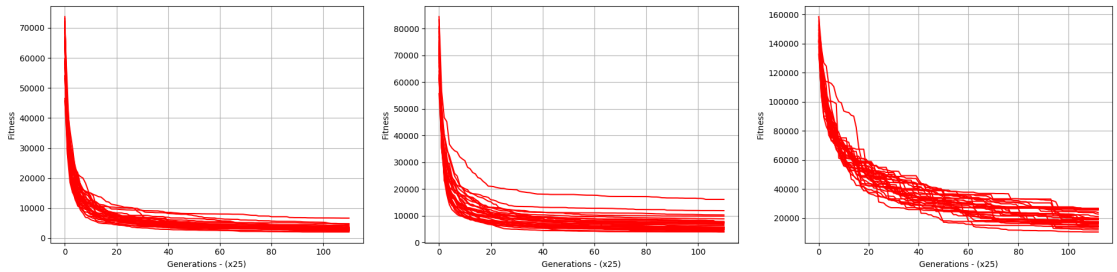
A Figura 5.2 ainda nos permite observar as diferenças entre AG canônico e BRKGA. Apesar de obter uma menor complexidade para 10 dimensões, o BRKGA possui um fator crescimento mais agressivo que o AG canônico, dobrando de complexidade nos casos de 100 dimensões para quase todos os casos. Entretanto esse comportamento era esperado, uma vez que o BRKGA possui mecanismos mais elaborados que visam o controle da

diversidade.

Figura 5.1: Curva de convergência das melhores abordagens (A11, A7, A8) em alguns casos. Eixo x representa a quantidade de gerações ($\times$ 25), já o eixo y representa o valor de *fitness*.

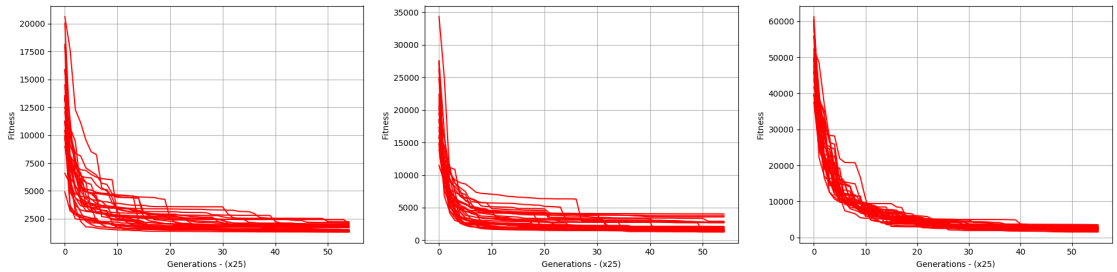
Função 9 - Dimensão 100

(a) A11 (b) A7 (c) A8



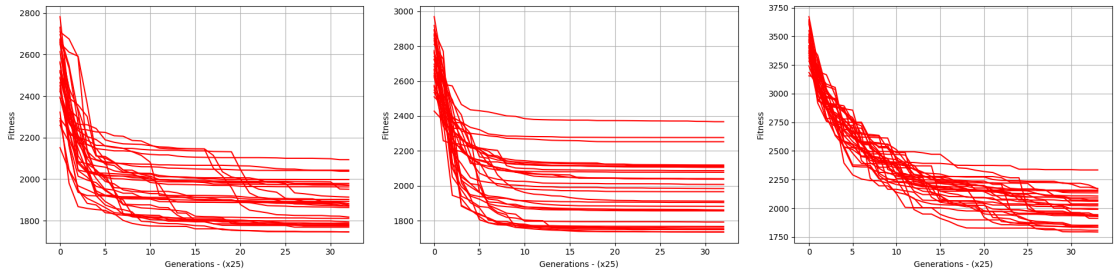
Função 11 - Dimensão 50

(d) A11 (e) A7 (f) A8



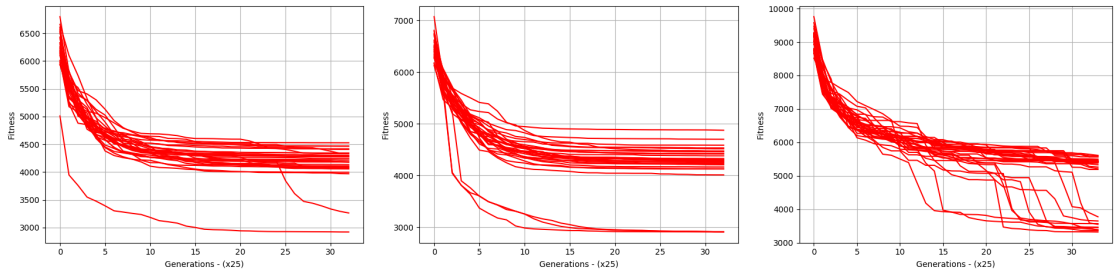
Função 17 - Dimensão 30

(g) A11 (h) A7 (i) A8



Função 26 -Dimensão 30

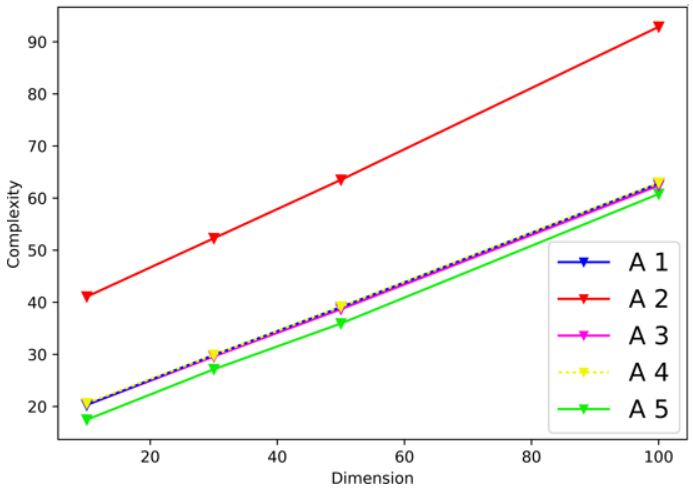
(j) A11 (k) A7 (l) A8



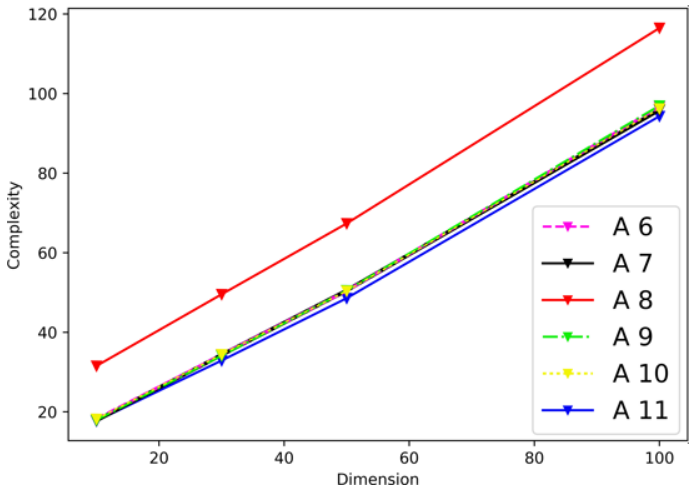
Fonte: Do Autor.

Figura 5.2: Curva de complexidade - Dimensões 10, 30, 50 e 100.

(a) Complexidade de A1-A5 (Políticas com o AG Canônico).



(b) Complexidade de A6-A11 (Políticas com o BRKGA)



Fonte: Do Autor.

5.1.5 Abordagem Selecionada

Como apresentado, a política proposta foi a que obteve maior sucesso, tanto em conjunto com o AG canônico, quanto utilizada com o BRKGA. Esta política alcançou um comportamento consistente, como apresentado pelas curvas de convergência, além de atingir melhores resultados em muitos casos. Também obteve os menores valores de complexidade. Em conjunto com o BRKGA, a política proposta alcançou o melhor resultado absoluto entre todas testadas, sendo então escolhida como a abordagem utilizada nas etapas seguintes.

5.2 Resultados - Etapa II

Nesta seção apresentaremos os resultados dos experimentos da etapa II, que consiste na incorporação de conhecimento do problema na forma de novos operadores genéticos e utilização de bases de dados experimentais. Estes operadores foram testados em conjunto com a abordagem proposta na etapa I, cuja a qual foi testada em funções de testes (sem domínio explícito). Dessa vez, as combinações abordagem + operadores foram testadas diretamente no problema PSP.

5.2.1 Inicialização

De modo a apresentar a distribuição dos indivíduos gerados pelos dois modelos de inicialização propostos, os valores de RMSD (eixo y) e energia (eixo x) foram dispostos em gráficos, conforme apresentado pela Figura 5.3. A linha vertical preta representa o valor de energia da estrutura experimental (em alguns casos a mesma não aparece por possuir valor inferior aos valores de energia presentes nas amostras). Pode-se observar que de maneira geral, ambos os modelos obtiveram distribuições semelhantes. Apesar disso, algumas diferenças pontuais podem ser percebidas, como no gráfico da 1AB1 do modelo I (Fig. 5.3g), onde há uma leve tendência à concentração de soluções em regiões de menor energia em comparação ao modelo II (Fig. 5.3h). Comportamento semelhante também ocorre com a proteína 1K43, exceto que dessa vez o modelo II (Fig. 5.3j) é quem possui sutil vantagem sobre o modelo I (Fig. 5.3i). Podemos também notar a complexidade do problema PSP quando observada a grande concentração de indivíduos (região em vermelho) com baixo valor de *fitness* mas com grande variação de RMSD, caracterizando um espaço de busca com múltiplos mínimos globais.

No intuito de estender a análise, a Tabela 5.6 apresenta as médias do RMSD e energia, além de um comparativo entre ambos. O primeiro caso (coluna "Ordenado por *Fitness*") apresenta os melhores casos ao ordenarmos as amostras pelo seus respectivos valores de *fitness* (que representa o que ocorre durante a execução da abordagem proposta). Esta análise visa ilustrar o que um dado método consideraria como melhor solução, uma vez que durante a execução se tem apenas a informação do *Fitness*. Em seguida, tem-se os melhores casos ao se ordenar as amostras pelo valor de RMSD. Por sua vez, esta análise representa o potencial que cada modelo tem em criar soluções com boa qualidade estrutural. É válido ressaltar que este caso é apenas ilustrativo, devido ao fato de que a

abordagem utilizada não tem conhecimento do valor de RMSD das soluções em relação à estrutura experimental. Apesar disso, a análise é valida uma vez que, apesar de que tais melhores soluções (em valores de RMSD) não poderem ser distinguidas, o fato das mesmas estarem contidas na população pode ser determinante para a obtenção de soluções de qualidade (e.g., uma solução com qualidade estrutural pode ser selecionada para recombinação e possivelmente gerar uma prole com bons valores de RMSD e *fitness*).

A Tabela 5.6 também provê uma comparação entre ambos os modelos, onde os melhores valores foram destacados em verde. Como pode-se perceber, o modelo II obteve melhores resultados em relação ao modelo I, principalmente nos valores médios de *fitness* e RMSD. Apesar de não haver diferenças em grande magnitude, os valores médios do modelo II foram majoritariamente melhores do que o modelo I. Ao encontro disso, temos a comparação ao ordenarmos as amostras pelo *fitness*, onde novamente o modelo II foi superior. Nos valores de *fitness*, o modelo II alcançou melhores resultados em todas as seis proteínas. Já em relação ao RMSD, apesar do modelo I alcançar melhores valores nas proteínas 2P5K e 2MR9, o modelo II foi melhor no restante do conjunto, havendo casos com diferenças significativas (e.g., proteína 1ACW, onde o modelo II obteve uma média menor em 2,4Å, ou na 1AB1, com diferença de 1,19Å). Já na comparação ao ordenarmos as amostras por RMSD, houve um maior equilíbrio, com certa vantagem do modelo I em relação ao modelo II, principalmente nos valores de *fitness*.

Apesar dos dois modelos apresentarem certa semelhança de resultados em alguns casos, o modelo II foi consistentemente superior ao modelo I, principalmente nos valores médios das amostras, o que julga-se extremamente importante no que tange a inicialização de uma população. Visando proporcionar indivíduos de qualidade à abordagem proposta, o modelo II foi escolhido como forma de inicialização utilizada. Entende-se que o modelo II tenha alcançado melhores resultados por possuir menor influência do fator estocástico em relação ao modelo I, uma vez que este modelo utiliza informação de similaridade em quase todas as inserções de fragmentos (apenas a primeira inserção ocorre totalmente aleatória), resultando assim em soluções de maior qualidade.

5.2.2 Recombinação

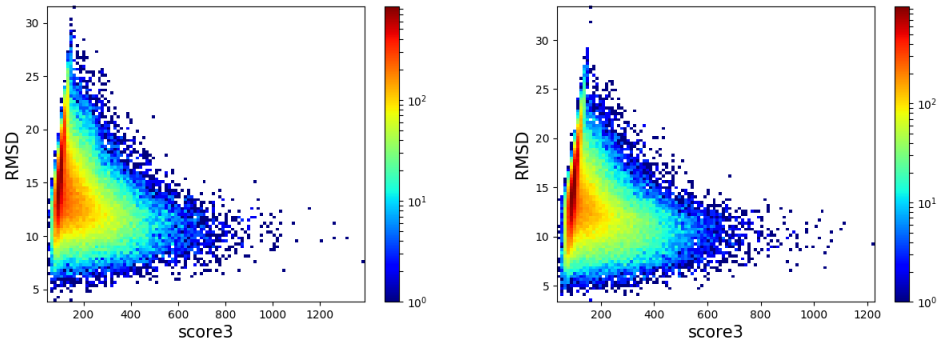
Os modelos de recombinação foram testados em conjunto com a abordagem selecionada na etapa anterior. A Tabela 5.7 apresenta os resultados de RMSD dos 3 modelos testados. Já a Tabela 5.8 apresenta os resultados em relação ao *fitness*. Em ambas as

Figura 5.3: Análise dos modelos de inicialização - energia x RMSD. A linha preta na vertical indica o valor de energia referente à proteína determinada experimentalmente. Alguns casos esta linha não aparece devido ao seu valor ser inferior ao obtidos nas presentes amostras.

2P5K

(a) Modelo I

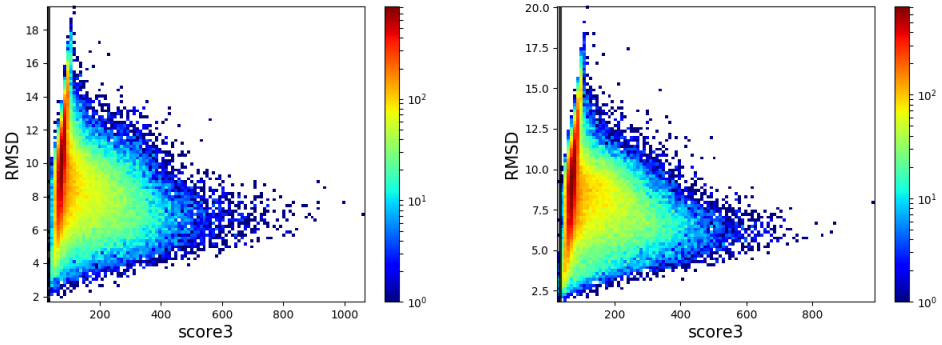
(b) Modelo II



2MR9

(c) Modelo I

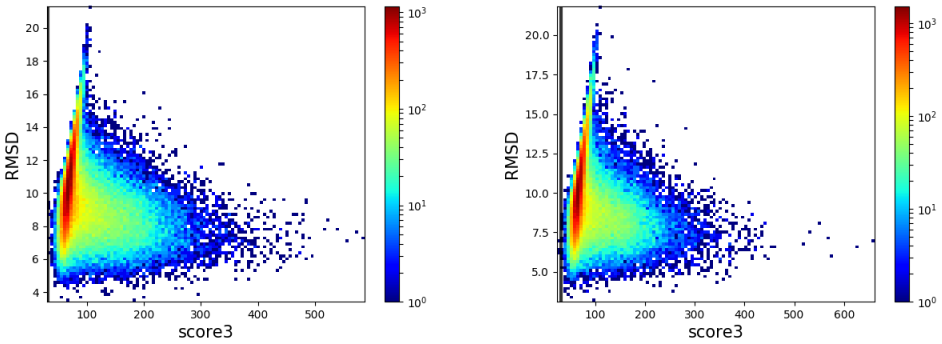
(d) Modelo II



1ACW

(e) Modelo I

(f) Modelo II

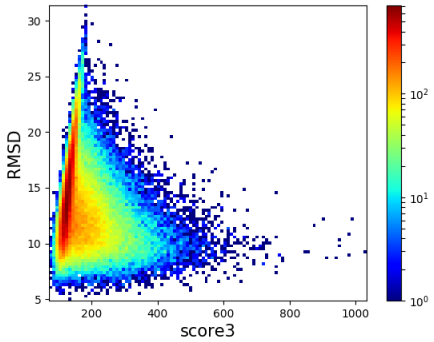


Fonte: Do Autor.

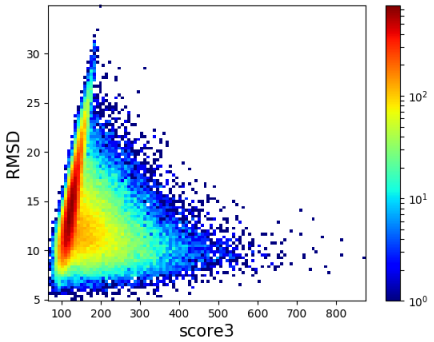
Figura 5.3: Continuação da Figura 5.3.

1AB1

(g) Modelo I

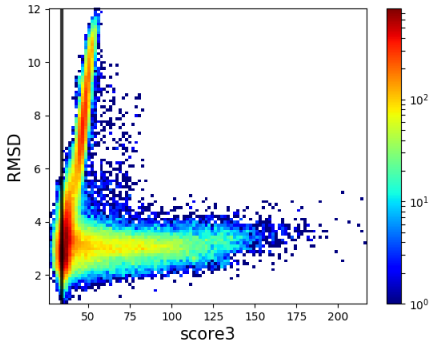


(h) Modelo II

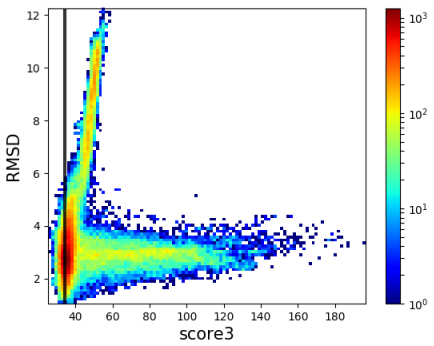


1K43

(i) Modelo I

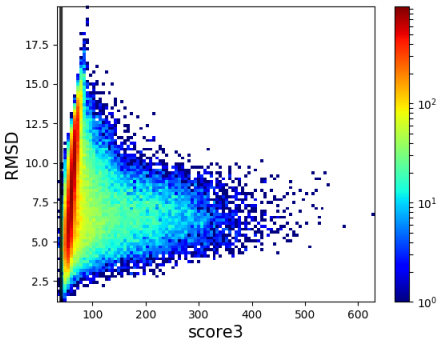


(j) Modelo II

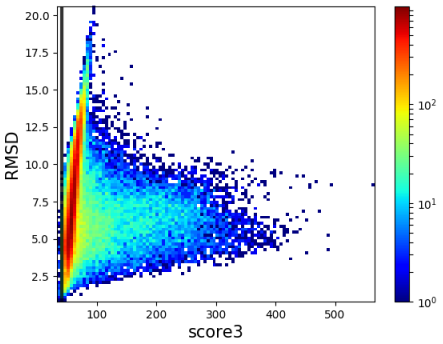


1ZDD

(k) Modelo I



(l) Modelo II



Fonte: Do Autor.

tabelas destacou-se em verde os melhores resultados.

Pode-se notar que, em relação ao RMSD, o modelo baseado em estrutura secun-

Tabela 5.6: Resultados dos métodos de inicialização utilizando conhecimento.

(a)

	Modelo I					
			Ordenado por Fitness		Ordenado por RMSD	
	Fitness medio	RMSD medio	menor Fitness	menor RMSD	menor Fitness	menor RMSD
2P5K	171,304	14,103	45,761	**9,244**	**81,573**	3,830
2MR9	128,915	8,922	28,533	**2,883**	**38,480**	**1,681**
1ACW	89,976	9,933	30,017	9,313	**67,087**	3,420
1AB1	167,565	**13,522**	70,430	8,124	**128,661**	**4,860**
1K43	49,380	4,423	26,410	3,231	36,597	**0,915**
1ZDD	81,668	8,065	33,286	4,646	47,472	1,196

(b)

	Modelo II					
			Ordenado por Fitness		Ordenado por RMSD	
	Fitness medio	RMSD medio	menor Fitness	menor RMSD	menor Fitness	menor RMSD
2P5K	**156,978**	**13,563**	**35,829**	10,530	155,831	**3,386**
2MR9	**123,034**	**8,348**	**25,401**	3,114	45,019	1,821
1ACW	**83,919**	**9,751**	**23,252**	**6,894**	96,822	**3,098**
1AB1	**158,001**	13,676	**65,933**	**6,931**	295,729	4,880
1K43	**46,197**	**3,924**	**25,251**	**2,997**	**35,162**	1,049
1ZDD	**74,634**	**7,381**	**32,794**	**4,546**	**39,491**	**0,789**

Fonte: Do Autor.

dária (II) e o modelo misto (III) foram os que obtiveram melhores resultados, enquanto que o modelo I (recombinação uniforme) obteve resultados inferiores. Percebe-se que as diferenças dos valores de RMSD entre os modelos II e III foram relativamente pequenas, ocorrendo casos em que, para uma mesma proteína, um dos métodos obteve o menor valor enquanto que o outro método alcançou a menor média (e.g., proteínas 1ACW e 1AB1). De modo geral, a diferença dos resultados dos métodos II e III são em torno de 0,3Å. Já o modelo I alcançou valores de menor qualidade, sendo que em algumas proteínas a diferença de RMSD médio chega a mais de 1,5Å (e.g., proteínas 2P5K e 1ZDD).

Entretanto, ao analisarmos os valores de *fitness*, o modelo I foi majoritariamente superior, seguido de perto pelo modelo III. Já o modelo II, obteve valores um pouco piores em relação ao demais modelos, não alcançando superioridade em nenhum dos casos testados. O modelo III, apesar de também não alcançar os menores valores em muitos casos (apenas na proteína 2P5K), obteve desempenho similar ao modelo I em algumas proteínas (e.g., proteínas 2MR9, 1K43 e 1ZDD).

Em se tratando dos componentes de recombinação propostos, nenhum modelo obteve superioridade absoluta. Houveram casos específicos onde cada modelo se provou eficiente. Todavia, o modelo III alcançou bons valores em relação ao RMSD, se provando igualmente eficiente ao modelo II, além de se mostrar satisfatório no quesito

fitness. Dessa forma, o modelo III foi o escolhido como operador de recombinação da abordagem proposta. Entende-se que este modelo obteve tal desempenho devido ao fato do mesmo utilizar ambos os modelos I e II de forma conjunta, explorando de maneira eficiente as melhores características de ambos (i.e., modelo I foi o melhor nos valores de *fitness* e o modelo II obteve melhor desempenho em relação ao RMSD).

Tabela 5.7: Resultados dos métodos de recombinação - (RMSD). Melhores resultados destacados em verde.

	Abordagem: A11								
	Uniforme			**Estrutura Secundária**			**Mista**		
	Menor	Médio	Desv.	Menor	Médio	Desv.	Menor	Médio	Desv.
2P5K	2,236	5,949	3,128	**1,567**	**3,207**	1,102	1,976	3,539	1,277
2MR9	1,614	2,184	0,515	1,664	2,257	0,488	**1,315**	**1,871**	0,352
1ACW	1,826	**2,738**	0,611	**1,535**	3,297	1,035	1,761	2,976	0,885
1AB1	3,856	5,147	1,100	3,103	**4,883**	1,193	**2,839**	4,954	1,775
1K43	1,218	1,453	0,197	0,920	1,196	0,154	**0,895**	**1,184**	0,165
1ZDD	1,222	3,711	1,727	**1,122**	**1,977**	1,054	1,349	2,076	0,969
Média	1,996	3,530		1,652	2,803		1,689	2,767	

Fonte: Do autor.

Tabela 5.8: Resultados dos métodos de recombinação - (*Fitness*). Melhores resultados destacados em verde.

	Abordagem: A11								
	Uniforme			**Estrutura Secundária**			**Mista**		
	Menor	Médio	Desv.	Menor	Médio	Desv.	Menor	Médio	Desv.
2P5K	-4,746	9,215	11,538	3,867	9,749	4,166	**-7,774**	**2,701**	6,218
2MR9	**-1,788**	**4,062**	3,865	5,210	9,919	2,305	-0,953	4,594	3,277
1ACW	5,195	**12,884**	6,343	15,740	24,138	8,891	**4,288**	14,339	6,710
1AB1	**25,624**	**38,476**	7,917	29,226	44,122	7,902	27,858	46,653	12,906
1K43	**15,950**	**17,652**	1,110	19,774	20,861	0,653	17,452	18,344	0,674
1ZDD	**14,612**	**18,809**	2,653	21,866	26,237	1,941	15,586	20,852	3,136
Média	9,141	16,850		15,947	22,504		9,410	17,914	

Fonte: Do autor.

5.2.3 Mutação

Seguindo uma linha de desenvolvimento construtiva, os componentes de mutação propostos foram testados em conjunto com a abordagem selecionada na etapa I (A11) e os modelos de inicialização (modelo II) e recombinação (modelo III) definidos nas seções anteriores. A Tabela 5.9 apresenta os resultados em função do RMSD. Já os valores de *fitness* podem ser consultados na Tabela 5.10. Nota-se que o BRKGA, que compõem a abordagem selecionada na etapa I, não possui explicitamente um operador de mutação. De modo a cumprir este papel, o algoritmo realiza a inserção de 20% de novos indivíduos a cada geração, conforme proposto por Alixandre e Dorn (2017). De modo a incorporar

os modelos propostos, e buscando não descaracterizar o BRKGA, optou-se por utilizar tanto a inserção de novos indivíduos quanto a mutação de indivíduos existentes, sendo cada opção responsável por gerar 10% da nova população. Ainda a respeito dos modelos de mutação, os mesmos foram aplicados a indivíduos contidos no grupo Elite, visando o refinamento das soluções de melhor qualidade.

Tabela 5.9: Resultados dos métodos de mutação - (RMSD). Melhores resultados destacados em verde.

	Abordagem: A11 - Recombinação: Mista					
	Modelo I			**Modelo II**		
	Menor	Médio	Desv.	Menor	Médio	Desv.
2P5K	2,015	**3,232**	0,844	**1,624**	3,318	1,096
2MR9	1,314	2,086	0,530	**1,293**	**1,920**	0,327
1ACW	1,876	2,796	0,750	**1,719**	**2,395**	0,457
1AB1	**1,821**	**5,246**	1,775	2,642	6,118	1,784
1K43	**0,622**	**1,083**	0,232	0,887	1,244	0,199
1ZDD	1,393	2,449	1,400	**1,092**	**2,154**	1,107
Média	1,507	2,815		1,543	2,858	

Fonte: Do autor.

Tabela 5.10: Resultados dos métodos de mutação - (*Fitness*). Melhores resultados destacados em verde.

	Abordagem: A11 - Recombinação: Mista					
	Modelo I			**Modelo II**		
	Menor	Médio	Desv.	Menor	Médio	Desv.
2P5K	-0,322	**10,282**	5,913	**-1,900**	11,120	8,139
2MR9	**-1,394**	**3,070**	2,575	-1,017	4,544	2,523
1ACW	**-1,405**	**14,833**	6,582	11,822	19,142	7,249
1AB1	**25,993**	37,017	9,608	26,101	**33,581**	5,583
1K43	**19,045**	**21,017**	1,170	19,855	21,393	1,202
1ZDD	16,737	**20,581**	2,346	**14,833**	20,820	3,753
Média	9,776	17,800		11,616	18,433	

Fonte: Do autor.

Em relação ao RMSD, percebe-se que de modo geral os resultados foram equilibrados entre ambos os modelos. O modelo I alcançou melhores resultados nas proteínas 1AB1 e 1K43. Já o modelo II foi superior nas proteínas 2MR9, 1ACW e 1ZDD, sendo a 2P5k dividida entre ambos. Apesar disso, ao olharmos com mais atenção, podemos notar que há pouca diferença de RMSD em alguns casos onde o modelo II foi superior (e.g., 2MR9 e 1ZDD, diferença de $\pm$ 0,2Å). Ao analisarmos os casos onde o modelo I foi superior, também pode-se notar certo equilíbrio, sendo a proteína 1AB1 o caso de maior disparidade (diferença de $\pm$ 0,8Å).

Todavia, ao analisarmos os valores de *fitness*, notam-se os bons resultados alcançados pelo modelo I. Apesar de alcançar melhores resultados em alguns casos (e.g., menor

valor nas proteínas 2P5K e 1ZDD), o modelo II não teve capacidade de acompanhar o desempenho do modelo I de maneira semelhante aos resultados de RMSD. Houve até certo equilíbrio em proteínas como a 2P5K e a 1K43, entretanto, Há casos de maior discrepância, como na proteína 1ACW. Dessa forma, podemos afirmar que o modelo I foi superior em relação aos valores de *fitness* obtidos.

De modo semelhante aos componentes de recombinação, nenhum modelo se destacou de maneira absoluta. Os resultados de RMSD se mostraram bastante equilibrados. Já os valores de *fitness* apresentaram certa vantagem ao modelo I. Ambas as métricas são importantes para a tomada de decisão, mas considerando o fato de que a abordagem utiliza apenas informação de *fitness* em seu processo de otimização, o modelo I foi escolhido como operador de mutação, uma vez que o mesmo se mostrou capaz de equilibrar bons valores de RMSD e *fitness*. Entende-se que tal fato tenha ocorrido devido ao fator de dependência dos fragmentos inseridos em seu processo de mutação (certos trechos são preenchidos com base na similaridade entre fragmentos), garantido assim estruturas com maior coesão e qualidade estrutural.

5.2.4 Operadores Selecionados

No intuito de agregar conhecimento específico do problema à abordagem proposta, uma série de operadores genéticos foram propostos e testados, buscando a combinação que resultasse nos melhores valores de *fitness*, bem como em relação à qualidade estrutural (medida pelo RMSD). Na inicialização da população, o modelo II obteve melhores resultados e se provou capaz de gerar, em média, boas soluções. Na recombinação, o modelo III, que é a conjunção dos dois outros modelos propostos, foi quem obteve mais sucesso, aproveitando as melhores características de ambos. Por fim, na mutação, optou-se pelo modelo I, que se mostrou eficiente tanto em relação aos valores de *fitness* obtidos, quanto na qualidade estrutural. Dessa forma, a versão final proposta é composta pelos seguintes componentes:

- **Versão de AG:** BRKGA (Veja Sec. 2.3.3)
- **Política de migração:** Política proposta (Veja Sec. 4.1)
- **Modelo de inicialização:** Modelo II (Veja Sec. 5.2.1)
- **Modelo de recombinação:** Modelo III (Veja Sec. 5.2.2)
- **Modelo de mutação:** Modelo I (Veja Sec. 5.2.3)

5.3 Resultados - Etapa III

Nesta seção serão apresentados os resultados da versão final da abordagem proposta, testada em um conjunto maior de proteínas. A mesma foi comparada com o método *Rosetta* (Veja Sec. 3.2.2), onde utilizou-se informação da estrutura secundária de forma predita, além de também compará-la como o método M5 de Corrêa et al. (2016) (Veja Sec. 3.2.1), dessa vez utilizando a estrutura secundária atribuída. Tanto o *Rosetta* quanto o AM utilizam informação experimental no processo de otimização (Fragmentos e APL respectivamente). Destaca-se que A11 representa o método proposto (no qual a abordagem 11 foi escolhida na etapa I) juntamente com os componentes desenvolvidos e selecionados na etapa II.

5.3.1 Comparando com o *Rosetta*

Nesta etapa de testes, comparou-se a abordagem proposta com o método *Rosetta*, um dos principais método de predição de estrutura tridimensional de proteína atualmente (KINCH et al., 2016). A Tabela 5.11 apresenta os resultados de ambos os métodos em relação ao RMSD e GDT, onde os melhores resultados foram destacados em negrito. A tabela também apresenta o resultado do teste estatístico aplicado no intuito de se verificar possíveis equivalências populacionais dos resultados obtidos. Casos onde a hipótese foi rejeitada (amostras não similares) foram destacados em negrito, com uma confiabilidade de 95%.

Observa-se que, de modo geral o método *Rosetta* foi superior ao método proposto (A11). Todavia algumas análises evidenciam uma certa vantagem de A11 em comparação ao *Rosetta* em determinados casos. Em relação ao RMSD, o método proposto alcançou melhores valores mínimos em 42% dos casos, e melhores médias em 36%. Apesar de grande parte dos resultados não terem similaridade estatística (de acordo com o teste realizado), em 26% dos casos a diferença de RMSD entre ambos os métodos, tanto em relação ao valor mínimo quanto à média, esteve próxima a 1Å (proteínas 1AB1, 1CRN, 1ROP, 1WQC e 1ZDD). Este número também se repete ao considerarmos uma diferença de $\pm$ 0,5Å (proteínas 1ENH, 1K43, 1L2Y, 1UTG e 2MTW), totalizando aproximadamente metade dos casos testados com diferença igual ou inferior a 1,0Å.

Se analisarmos os resultados abaixo de 3,0Å (limiar aceitável de similaridade em relação à estrutura experimental (CARUGO, 2003)), podemos notar que, em se tratando

do menor valor alcançado, ambos os métodos obtiveram êxito em aproximadamente 65% dos casos (A11 e *Rosetta* em 12 e 13 proteínas respectivamente). Contudo o mesmo não se repete ao considerarmos os valores médios, onde o método proposto obteve 4 casos abaixo de 3,0Å, seguido do *Rosetta* com 3. Essa diferença aponta para uma certa instabilidade dos métodos, que pode ser devido tanto aos próprios algoritmos quanto às funções de energia que ainda não são capazes de guiar a busca de maneira totalmente precisa. Esta dispersão se torna clara ao analisarmos os gráficos de caixa apresentados pela Figura 5.4. Apesar de alguns casos se mostrarem coesos (e.g., 1K43, 2MTW), ambos os métodos também tiveram casos com grande dispersão dos resultados (e.g., proteínas 1ROP e 2PMR no método proposto, 1UTG e 2P5K no método *Rosetta*). Tal dispersão caracteriza as dificuldades dos métodos em repetirem o caminho de otimização de maneira eficiente, todavia pode-se tirar de positivo a possibilidade de alcançar mínimos de qualidade, apontando para uma capacidade promissora de ambos os métodos.

Ao observarmos os resultados em função do GDT, o desempenho do *Rosetta* em comparação ao método proposto se mostra novamente superior. Apesar de 37% dos casos serem considerados estatisticamente similares, o *Rosetta* obteve melhores médias e máximos em, respectivamente, 14 e 15 casos. Vale destacar que, em alguns casos o melhor resultado de acordo com o RMSD foram obtidos pelo método proposto, entretanto em relação ao GDT, tais casos foram melhores no método *Rosetta* (e.g., proteínas 1AB1, 2PMR, 1ROP).

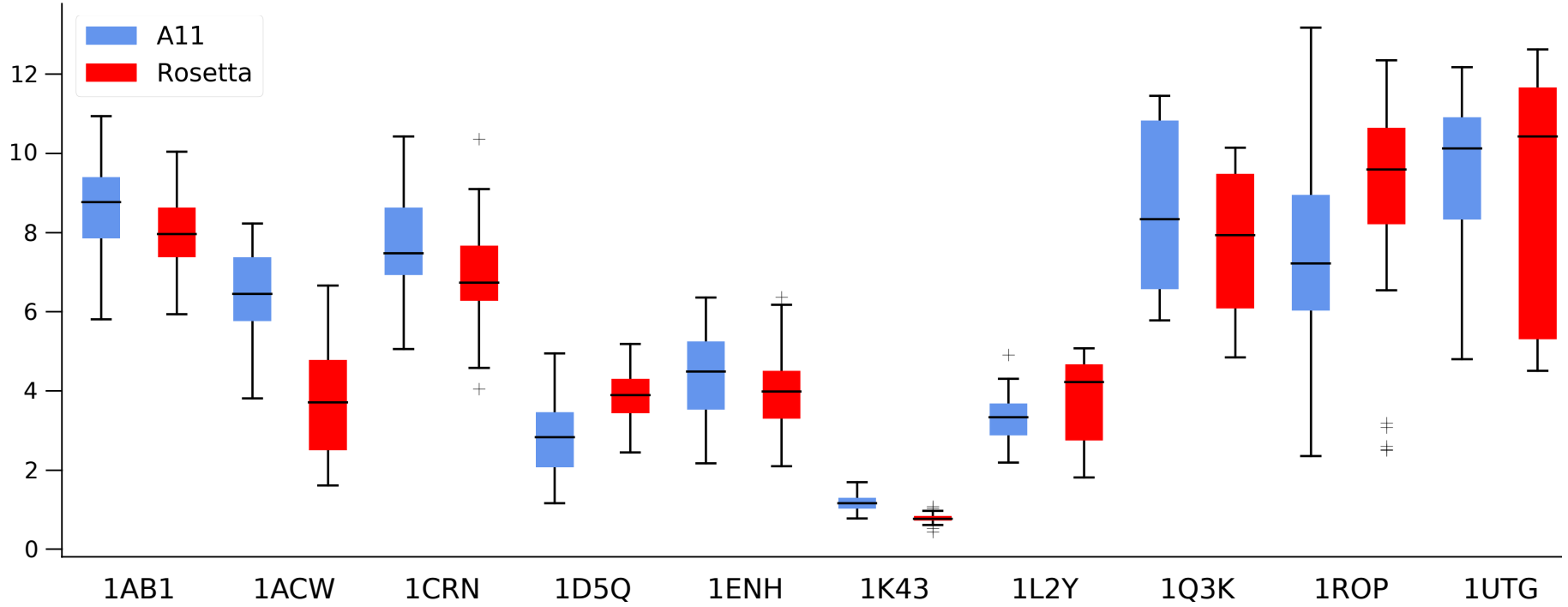
Entende-se que, como um todo, o método proposto se mostrou competitivo, uma vez que, se considerarmos os casos onde o mesmo foi superior ou alcançou resultados estatisticamente similares ao método *Rosetta*, temos uma parcela considerável dos casos testados. A Figura 5.5 apresenta de modo gráfico as estruturas de menor RMSD de ambos os métodos em cada um dos casos testados. Para melhor observação, ambas foram sobrepostas de maneira ótima em relação à estrutura experimental (A11 em vermelho, *Rosetta* em azul e a estrutura experimental em verde). Pode-se observar que, com algumas exceções (e.g., proteína 1AB1), ambos os métodos foram capazes de alcançar resultados com empacotamento semelhantes às estruturas nativas (e.g., proteínas 1K43, 2MR9, 3V1A).

Tabela 5.11: Resultados - A11 em comparação ao *Rosetta*. Valores obtidos após 30 execuções de cada método para cada proteína. Melhores resultados destacados em negrito. Valor-*p* representa o valor obtido pelo teste estatístico *Wilcoxon-Mann-Whitney* utilizando um intervalo de confiança de 95%, onde os valores em negrito representam casos de rejeição da hipótese de similaridade.

PDB ID	**Métodos**	**RMSD (Å)**				**GDT (%)**			
		Mín.	**Médio**	**(Desv. p.)**	**Valor-*p***	**Máx.**	**Médio**	**(Desv. p.)**	**Valor-*p***
1AB1	A11	**5,798**	8,603	1,167	**1,304E-02**	41,850	35,001	3,701	2,671E-01
	Rosetta	5,924	**7,969**	0,875		**45,650**	**36,050**	4,137	
1ACW	A11	3,805	6,461	1,116	**7,790E-09**	57,760	43,305	4,085	**5,057E-09**
	Rosetta	**1,607**	**3,750**	1,412		**80,170**	**61,868**	11,157	
1CRN	A11	5,046	7,569	1,295	**3,073E-02**	49,460	**44,765**	3,459	2,014E-01
	Rosetta	**4,035**	**6,902**	1,329		**55,430**	44,329	5,368	
1D5Q	A11	**1,159**	**2,962**	1,059	**7,212E-04**	**85,190**	**68,117**	7,219	**2,290E-03**
	Rosetta	2,438	3,805	0,648		73,150	62,685	5,622	
1ENH	A11	2,166	4,439	1,115	6,299E-02	44,440	39,121	2,677	**2,134E-03**
	Rosetta	**2,095**	**4,052**	1,085		**47,690**	**41,420**	2,969	
1K43	A11	0,778	1,184	0,240	**2,731E-09**	89,290	74,761	4,559	**8,533E-10**
	Rosetta	**0,440**	**0,781**	0,135		**91,070**	**85,238**	2,802	
1L2Y	A11	2,188	**3,331**	0,644	**4,534E-03**	71,250	60,458	5,169	2,425E-01
	Rosetta	**1,812**	3,830	1,070		**77,500**	**60,833**	8,026	
1Q2K	A11	5,769	8,596	2,076	**3,177E-02**	45,160	34,166	5,351	8,076E-02
	Rosetta	**4,830**	**7,707**	1,801		**54,840**	**37,742**	8,407	
1ROP	A11	**2,347**	**7,296**	2,410	**6,366E-03**	72,320	47,128	9,977	3,557E-01
	Rosetta	2,497	8,598	2,897		**75,000**	**47,456**	11,482	
1UTG	A11	4,788	9,444	2,115	4,794E-01	54,290	40,274	5,210	7,994E-02
	Rosetta	**4,493**	**9,049**	2,994		**55,710**	**43,142**	7,249	
1WQC	A11	2,814	3,427	0,236	**2,804E-05**	73,080	69,231	2,276	**1,221E-02**
	Rosetta	**1,657**	**2,556**	0,742		**78,850**	**71,314**	4,921	
1ZDD	A11	1,671	4,880	1,007	**5,614E-03**	43,380	40,025	1,467	**1,865E-04**
	Rosetta	**1,193**	**3,809**	1,545		**48,530**	**42,060**	2,397	
2MR9	A11	**1,284**	**1,998**	0,406	**2,109E-04**	**85,800**	**75,379**	5,227	**4,848E-05**
	Rosetta	1,651	3,472	2,173		79,550	66,136	9,426	
2MTW	A11	**4,410**	**5,506**	0,719	2,846E-01	**56,250**	49,167	3,359	2,350E-01
	Rosetta	4,787	5,597	0,552		**56,250**	**49,583**	3,810	
2P5K	A11	**1,275**	**3,454**	1,505	**3,101E-04**	**55,160**	**48,822**	3,094	**1,717E-04**
	Rosetta	2,059	7,267	4,047		51,980	43,426	5,454	
2P6J	A11	2,592	6,307	2,457	**1,579E-05**	65,870	51,075	8,450	**7,860E-05**
	Rosetta	**1,981**	**3,749**	1,330		**76,920**	**60,290**	6,493	
2P81	A11	5,762	8,250	1,297	**1,139E-05**	27,840	22,254	2,156	**1,402E-11**
	Rosetta	**4,489**	**6,628**	1,114		**38,070**	**32,747**	2,568	
2PMR	A11	**1,397**	5,920	3,015	**2,731E-06**	48,680	42,632	3,473	**4,051E-06**
	Rosetta	1,439	**2,527**	1,819		**50,990**	**46,854**	2,483	
3V1A	A11	**0,744**	**1,318**	0,290	**7,033E-05**	55,210	**53,039**	1,250	**3,233E-02**
	Rosetta	0,998	3,141	1,877		**55,730**	49,237	5,484	
Resumo		42,10% (8/19)	36,84% (7/19)		-	21,05% (4/19)	26,31% (5/19)		-

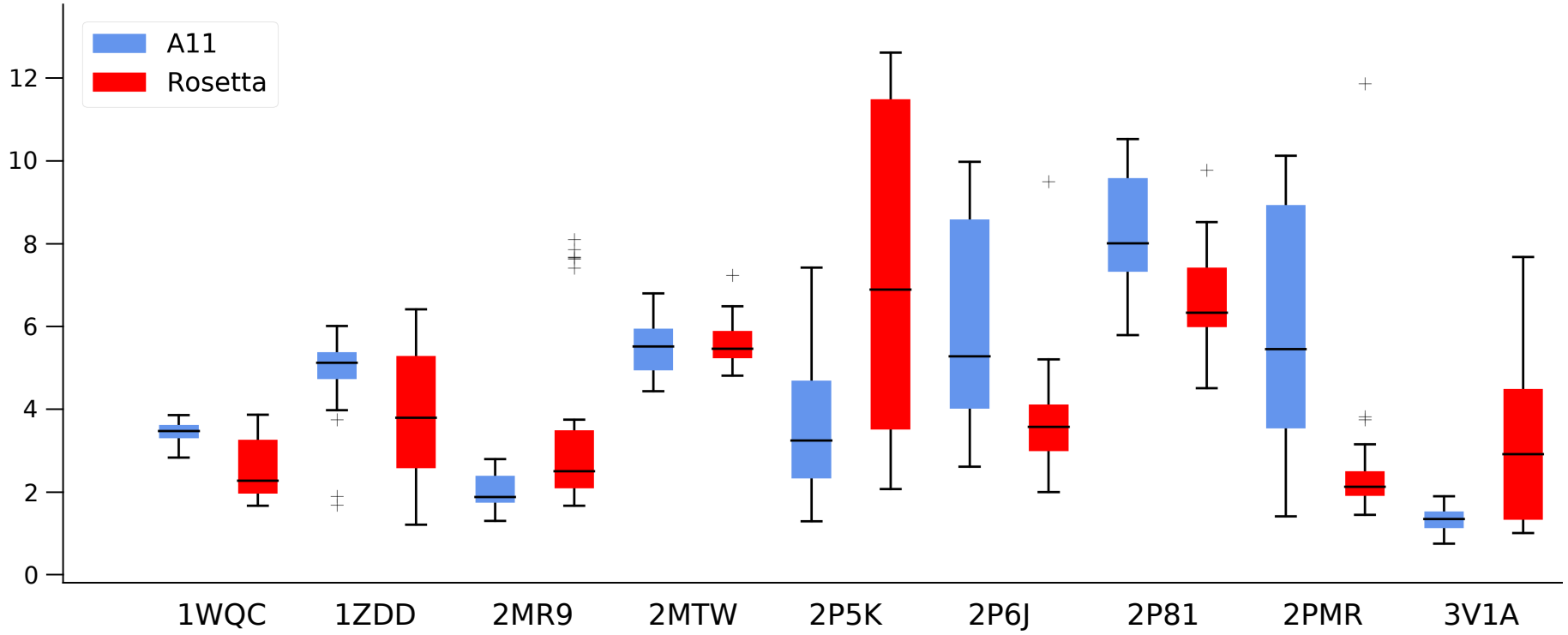
Fonte: Do autor.

Figura 5.4: Gráfico de caixa. Comparação entre A11 com o *Rosetta*. Eixo y apresenta os valores da métrica de similaridade RMSD, onde quanto menor seu valor, menor a diferença entre os modelos analisados.



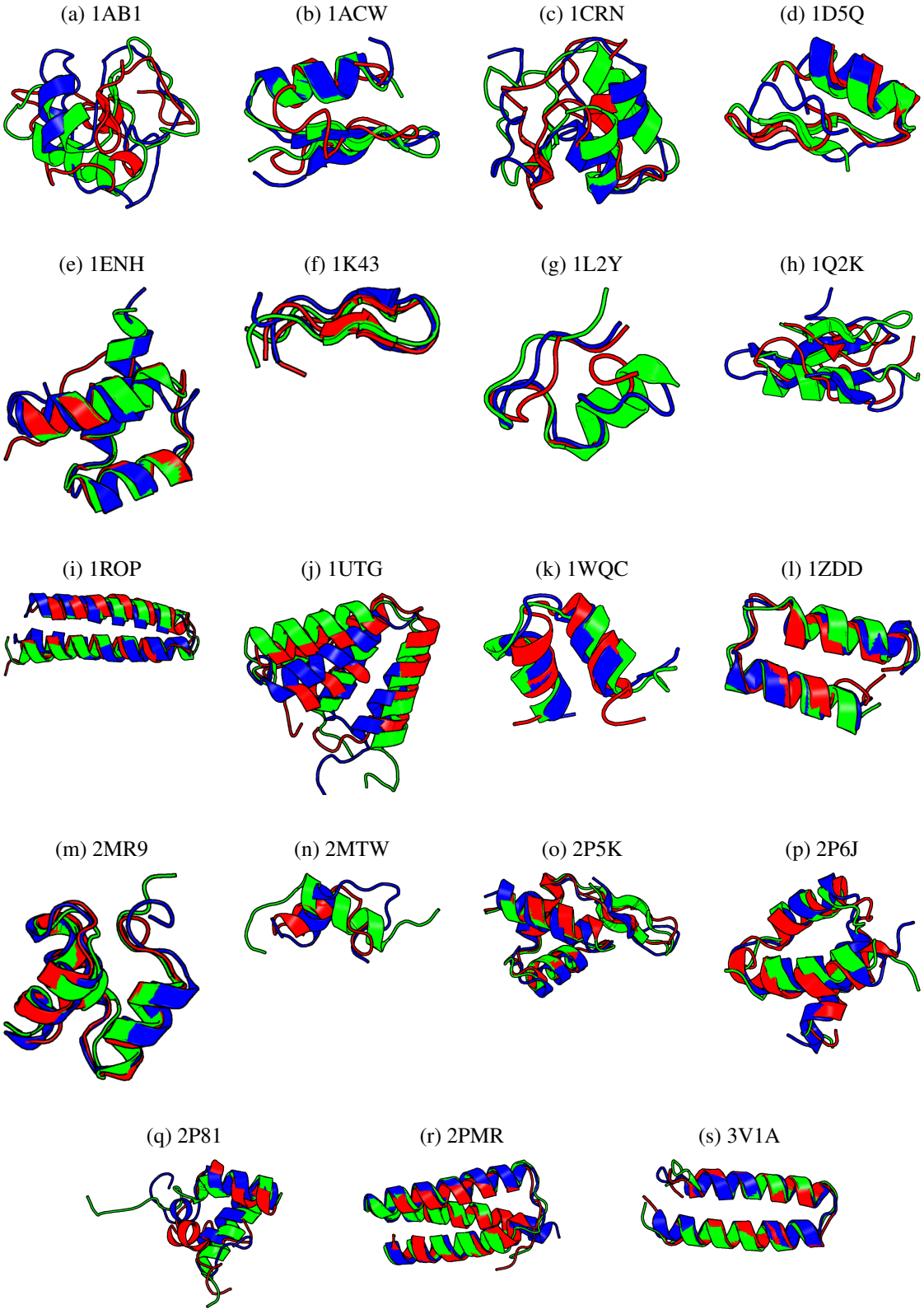
Fonte: Do Autor.

Figura 5.4: Continuação da Figura 5.4.



Fonte: Do Autor.

Figura 5.5: Representação gráfica das estruturas de menor RMSD - Experimental (verde), A11 (vermelho) e *Rosetta* (azul).



Fonte: Do autor. Preparadas com PYMOL (www.pymol.org)

5.3.2 Comparando com o M5

Nesta comparação, ambos os métodos utilizam a estrutura secundária atribuída. Isso tem como propósito eliminar a influência dos erros de predição de estrutura secundária no processo de otimização. Assim sendo, parte-se do princípio de que tal problema esteja resolvido (predição da estrutura secundária), informando ao método os padrões exatos. A Tabela 5.12 apresenta os resultados dos métodos A11 e M5, ambos em função de RMSD e GDT, onde destacaram-se em negrito os melhores resultados obtidos. Tais testes também foram submetidos à um teste de significância estatística, com a hipótese de amostras similares, sendo os valores em negrito aqueles aos quais tal hipótese foi rejeitada (utilizando um intervalo de confiança de 95%).

De acordo com os valores de RMSD, pode-se notar que o método proposto foi superior em grande parte dos casos testados. Em relação ao menor valor obtido, o A11 obteve melhores resultados em praticamente 70% dos casos. Se olharmos as médias, esse desempenho sobe para 80%. Com apenas três casos onde os resultados foram considerados estatisticamente similares (proteínas 1L2Y, 1Q2K e 1UTG), o método proposto obteve ambas as melhores médias e valores mínimos em mais de 60% dos casos. Apesar de a diferença entre os métodos ser de menos de 1,0Å em alguns casos (e.g., proteínas 1ACW, 1K43, 1ZDD), em muitos outros o método A11 alcança valores médios com diferença de mais de 3,0Å (e.g., 1AB1, 1ENH, 2P5K, 2P6J). Se considerarmos o limiar de 3,0Å, o método proposto obteve êxito em 89% dos casos em relação aos valores mínimos. Nos valores médios, a taxa foi de 42%. Já o método M5, nos mesmos critérios, obteve êxito de 73% e 10% nos valores mínimos e médios, respectivamente.

Ao observarmos os resultados em função do GDT, novamente o método proposto se apresenta superior. Apesar de alguns casos terem pouca diferença no valor médio (proteínas 1ENH, 1Q2K, 1UTG, 1ZDD, 2P81, 2PMR, 3V1A tiverem diferença média de menos de 3% entre ambos os métodos), houve situações com diferenças consideráveis (13% na 1AB1, 16% na 1CRN, 10% na 1D5Q, 24% na 2MR9, 10% na 2P6J). Nota-se que o método proposto foi superior em quase 70% dos casos em relação ao GDT médio. Já nos valores máximos, a situação se equilibra, tendo o método M5 alcançando melhores resultados em 53% dos casos.

Outro aspecto importante a se notar consiste na coesão dos resultados obtidos, que podem ser inferidos a partir dos gráficos de caixa apresentados pela Figura 5.6. Pode-se observar a baixa dispersão dos resultados obtidos pelo método proposto, ao passo que

M5 obteve resultados mais distribuídos. Tal comportamento pode se caracterizar devido à dificuldades do método M5 em lidar com alguns obstáculos durante o processo de otimização, levando o mesmo à uma maior variação dos resultados. Já o método A11, à exceção de alguns casos (e.g., proteínas 1ROP, 1UTG, 2P5K), obteve resultados com valores de média, mediana e mínimo semelhantes, provando a eficiência do mesmo em repetir seu desempenho apesar dos fatores estocásticos.

Entende-se que, neste cenário, o método proposto provou-se uma abordagem eficiente para o problema PSP, tanto pelos resultados superiores em relação ao método M5, quanto pela quantidade de casos de êxito ao considerarmos o limiar de 3,0Å. Apesar disso, também destaca-se o desempenho do método M5, onde nota-se seu potencial, principalmente em relação ao mínimos de RMSD abaixo do limiar proposto (aproximadamente 75% dos casos), tendo como principal obstáculo a manutenção destes resultados ao longo de múltiplas execuções. A Figura 5.7 apresenta graficamente as estruturas de menor RMSD dos métodos A11 (vermelho) e M5 (azul), em sobreposição ótima à estrutura experimental (verde).

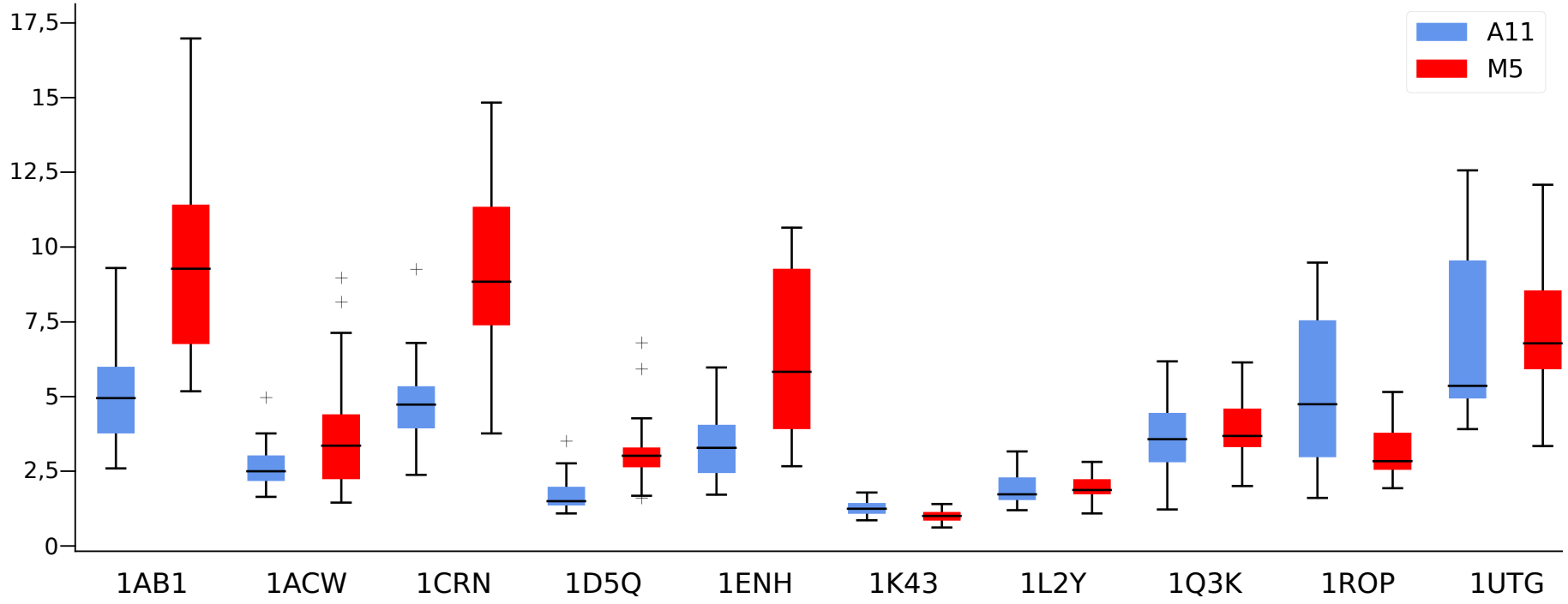
Destaca-se também o ganho de desempenho ao se utilizar a informação precisa da estrutura secundária. Tal diferença fica notória ao compararmos os valores de A11 apresentados pelas Tabelas 5.11 e 5.12, onde a primeira utiliza informação predita (passível de erro), enquanto a segunda utiliza a estrutura atribuída (informação 100% precisa). Pode-se observar que a utilização da informação correta da estrutura secundária culminou em melhoras em praticamente todos os casos testados, sendo que em muitos casos os valores de RMSD melhoram em mais de 3,0Å (e.g., proteína 1ACW teve a média de 6,4Å reduzida a 2,6Å, proteína 1Q2K teve a média de 8,6Å reduzida a 3,6Å). Os valores de GDT também subiram consideravelmente ao se utilizar a informação atribuída (e.g., proteína 1AB1 teve a média aumentada em 24%, proteína 1CRN teve a média aumentada em 16%). Apesar de já sabido, tal fato ilustra a importância de se predizer corretamente a estrutura secundária de proteínas, consequentemente gerando um grande impacto na predição da estrutura tridimensional.

Tabela 5.12: Resultados - A11 em comparação ao M5. Valores obtidos após 30 execuções de cada método para cada proteína. Melhores resultados destacados em negrito. Valor-*p* representa o valor obtido pelo teste estatístico *Wilcoxon-Mann-Whitney* utilizando um intervalo de confiança de 95%, onde os valores em negrito representam casos de rejeição da hipótese de similaridade.

PDB ID	**Métodos**	**RMSD (Å)**				**GDT (%)**			
		Mín.	**Médio**	**(Desv. p.)**	**Valor-*p***	**Máx.**	**Médio**	**(Desv. p.)**	**Valor-*p***
1AB1	A11	**2,598**	**5,022**	1,583	**2,546E-08**	**74,460**	**58,987**	9,055	**5,072E-07**
	M5	5,180	9,430	3,060		67,930	45,110	8,360	
1ACW	A11	1,649	**2,653**	0,686	**1,255E-02**	81,030	**68,620**	5,882	**1,177E-02**
	M5	**1,450**	3,800	1,910		**82,760**	63,480	9,200	
1CRN	A11	**2,382**	**4,796**	1,322	**3,259E-09**	**81,520**	**60,943**	8,586	**5,435E-09**
	M5	3,770	9,240	2,650		60,330	44,080	7,070	
1D5Q	A11	**1,095**	**1,711**	0,541	**7,147E-09**	**90,740**	**81,358**	6,392	**4,599E-07**
	M5	1,620	3,160	1,060		80,560	71,540	5,440	
1ENH	A11	**1,719**	**3,396**	1,092	**1,158E-06**	**46,760**	**41,682**	2,420	**8,034E-03**
	M5	2,670	6,490	2,640		**46,760**	39,380	3,640	
1K43	A11	0,869	1,270	0,243	**7,033E-05**	83,930	74,047	3,739	**2,547E-05**
	M5	**0,630**	**1,000**	0,210		**85,710**	**78,930**	4,030	
1L2Y	A11	1,207	2,003	0,622	2,796E-01	81,250	72,750	5,849	**1,618E-04**
	M5	**1,100**	**1,940**	0,410		**85,000**	**78,330**	4,140	
1Q2K	A11	**1,233**	**3,558**	1,203	1,855E-01	**83,060**	**64,274**	7,364	4,382E-01
	M5	2,010	3,830	0,940		79,840	63,520	5,380	
1ROP	A11	**1,619**	5,202	2,507	**1,591E-03**	**77,680**	58,749	11,784	**3,252E-03**
	M5	1,940	**3,160**	0,900		76,790	**67,280**	5,760	
1UTG	A11	3,915	**6,987**	2,744	2,367E-01	59,640	**49,203**	7,873	1,573E-01
	M5	**3,340**	7,150	2,230		**63,210**	46,820	8,430	
1WQC	A11	**1,647**	**2,344**	0,297	**6,644E-11**	**76,920**	**70,737**	2,870	**9,580E-11**
	M5	2,470	3,970	0,720		69,230	61,090	4,200	
1ZDD	A11	**1,083**	**2,603**	1,479	**3,274E-04**	44,850	41,912	1,259	**2,864E-04**
	M5	1,890	3,590	1,370		**48,530**	**43,630**	2,150	
2MR9	A11	**1,408**	**2,187**	0,543	**6,028E-11**	**84,660**	**73,352**	5,788	**2,432E-11**
	M5	2,620	5,920	1,440		66,480	49,470	6,050	
2MTW	A11	2,640	4,278	0,573	**4,959E-11**	71,250	59,250	4,397	**2,347E-11**
	M5	**2,100**	**2,590**	0,300		**80,000**	**73,580**	3,020	
2P5K	A11	**1,632**	**3,938**	2,426	**1,541E-08**	**53,170**	**47,937**	4,163	**1,482E-10**
	M5	4,260	9,590	3,660		45,630	34,950	4,440	
2P6J	A11	**2,656**	**4,065**	1,342	**5,333E-08**	**69,710**	**58,525**	5,870	**1,128E-07**
	M5	2,660	7,550	2,360		64,420	48,960	4,710	
2P81	A11	3,776	**5,601**	1,038	**1,026E-03**	36,930	34,223	1,710	**9,434E-05**
	M5	**3,750**	6,450	1,120		**37,500**	**35,810**	1,360	
2PMR	A11	**1,562**	**3,438**	1,837	**1,158E-06**	49,010	**43,553**	3,048	**4,222E-03**
	M5	2,530	6,370	2,600		**50,990**	41,620	3,600	
3V1A	A11	**0,816**	**1,425**	0,588	**9,284E-10**	55,210	**52,658**	2,157	**3,182E-02**
	M5	1,890	3,300	1,350		**60,940**	51,800	3,160	
Resumo		68,42% (13/19)	78,95% (15/19)		-	47,37% (9/19)	68,42% (13/19)		-

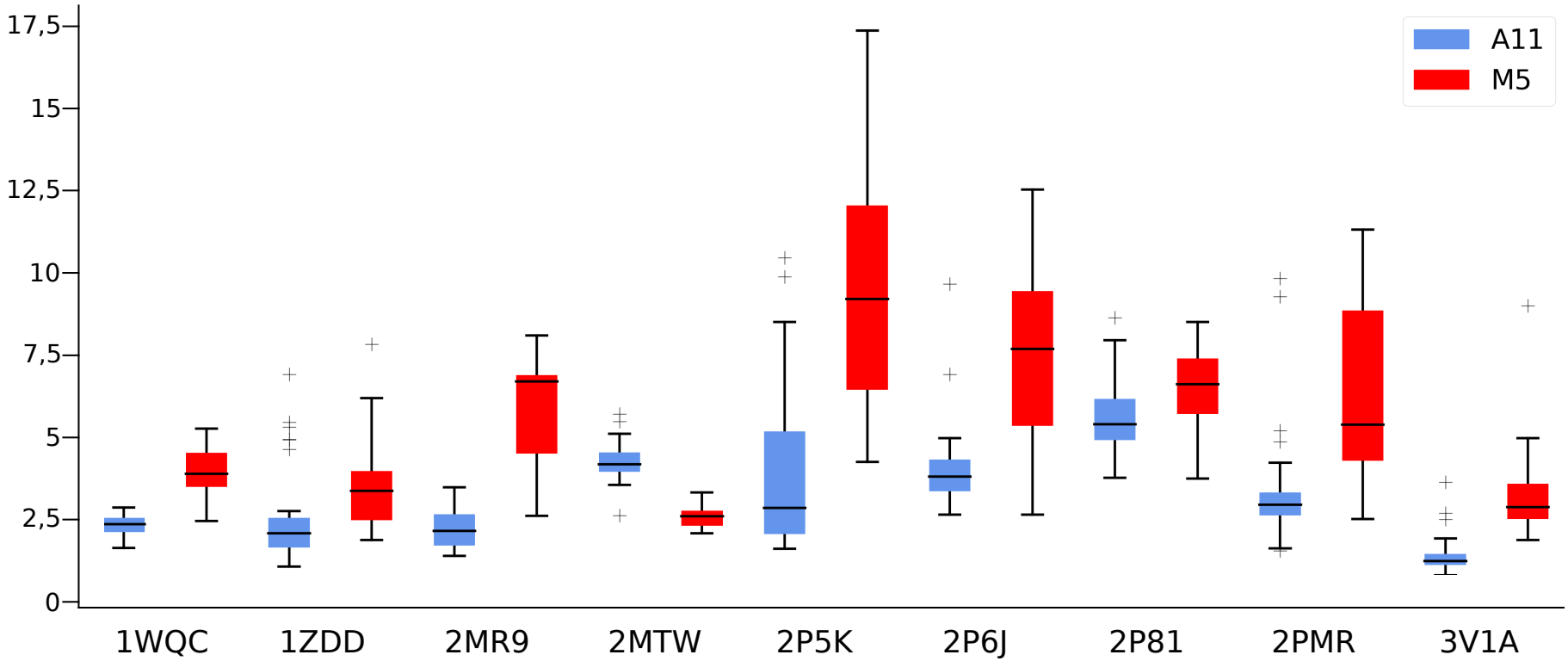
Fonte: Do autor.

Figura 5.6: Gráfico de caixa. Comparação entre A11 com o M5. Eixo y apresenta os valores da métrica de similaridade RMSD, onde quanto menor seu valor, menor a diferença entre os modelos analisados.



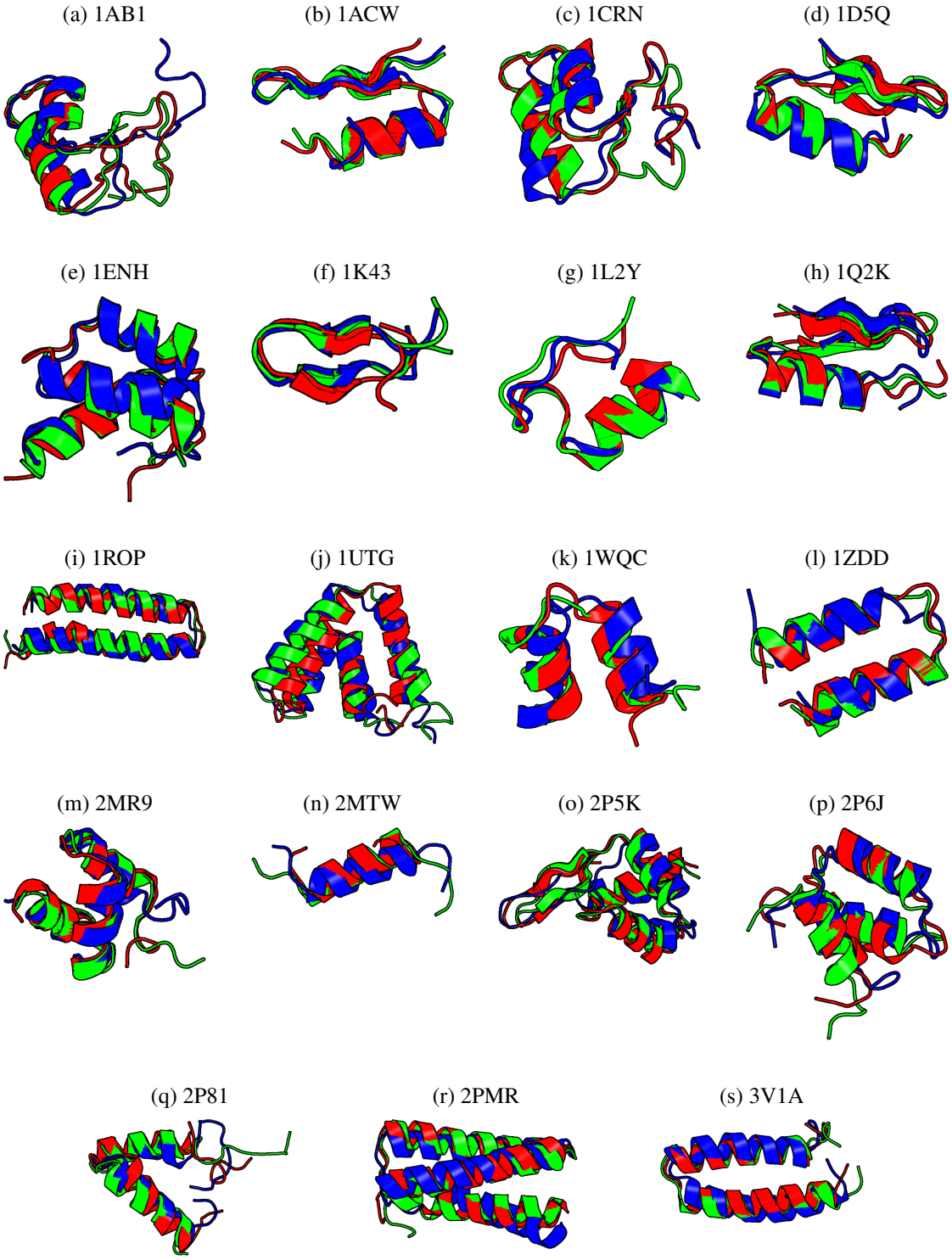
Fonte: Do Autor.

Figura 5.6: Continuação da Figura 5.6.



Fonte: Do Autor.

Figura 5.7: Representação gráfica das estruturas de menor RMSD - Experimental (verde), A11 (vermelho) e M5 (azul).



Fonte: Do autor. Preparadas com PYMOL (www.pymol.org)

5.4 Resumo do Capítulo

Neste capítulo apresentaram-se os resultados das três etapas de testes realizadas, sendo as duas primeiras relacionadas à construção da abordagem proposta, seguida da terceira onde realizou-se uma comparação final de seu desempenho. Na etapa I, os resultados indicaram um melhor desempenho da política de migração proposta, juntamente com o BRKGA. Na etapa seguinte, os operadores de inicialização e mutação que de alguma forma tomavam mais decisões baseados na métrica de similaridade foram aqueles de melhor desempenho, além do operador de recombinação misto, que aliou ambos os demais modelos propostos, alcançando valores equilibrados em *fitness* e RMSD. Por fim, na comparação com o método *Rosetta*, apesar deste se mostrar superior em mais da metade dos casos, o método proposto alcançou resultados competitivos em relação à qualidade estrutural. Já na comparação com o método M5, o método proposto foi amplamente superior, atingindo melhores médias de RMSD e GDT em 80% e 70% dos casos, respectivamente. Destacou-se também a influência da qualidade da informação da estrutura secundária, resultando em ganhos significativos em todos os critérios avaliados.

6 CONCLUSÕES

A estrutura tridimensional de uma proteína é informação crucial para o estudo de sua função biológica, além de conhecimento chave a outras tarefas, como o desenvolvimento de fármacos. Determinar a estrutura tridimensional de uma proteína é uma tarefa custosa e demorada, fazendo-se necessário o desenvolvimento de meios alternativos para a obtenção da mesma. Esta dissertação teve por objetivo o desenvolvimento de uma metaheurística baseada em AGD com controle de diversidade para o problema PSP.

Devido à complexidade do problema, existem certos fatores (imprecisão das funções de energia) que podem enviesar o processo de otimização do problema. Devido a isso, inicialmente desenvolveu-se uma política de migração aplicando-a em funções de testes, evitando assim a influência indesejada destes fatores. Como comparação, a política proposta foi analisada em conjunto com outras cinco, sendo quatro delas também baseadas em diversidade além da política canônica. Cada uma das políticas foi testada em dois casos: (*i*) utilizando o AG canônico e; (*ii*) utilizando o BRKGA.

Os resultados mostram que a abordagem desenvolvida nesta pesquisa foi superior às demais, principalmente nas funções mais complexas. Em conjunto com os gráficos de convergência, podemos notar que nossa abordagem proposta apresenta uma maior robustez, se fazendo suficiente para escapar de ótimos locais, tirando vantagem da diversidade gerada durante o processo de migração. Esses fatores, em conjunto com sua baixa complexidade, fazem da política proposta uma abordagem adequada para problemas de otimização complexos. Dentre as principais características que levaram nossa abordagem a tal desempenho, destacamos sua capacidade de explorar a diversidade da população, combinando indivíduos no mesmo intervalo de *fitness*. Entendemos que apenas migrar indivíduos diferentes entre si não seja o suficiente para lidar com certos tipos de problemas. Como indício disto, temos o número de *demes* que foi ajustado para cada abordagem testada. As políticas utilizadas com o AG canônico foram melhores com um número maior de *demes*. Com o BRKGA, o número de *demes* diminuiu em funções mais simples, entretanto em casos de maior complexidade novamente o ideal foi um número maior de *demes*. Tal comportamento não ocorreu quando combinou-se BRKGA com a política proposta, estabilizando-se no número mínimo de *demes* testado. Isso indica que os mecanismos da nossa abordagem foram eficientes, enquanto as demais políticas necessitaram dividir a população global em um maior número de *demes* de modo a evitar a convergência prematura.

Uma vez que a política proposta em conjunto com o *BRKGA* se mostraram a melhor opção, uma série de operadores genéticos foram elaborados e testados visando agregar conhecimento específico do problema PSP. Os resultados mostraram que a utilização da métrica de similaridade entre fragmentos é fator determinante na qualidade das soluções geradas. Na inicialização, o modelo II realiza apenas a primeira inserção de modo totalmente aleatório. Após isso, os demais fragmentos são inseridos baseados na similaridade ao adjacente. Tal fator também se faz presente no operador de mutação, que altera apenas regiões irregulares, preenchendo estes espaços com inserções baseadas na similaridade. Estes modelos obtiveram melhor desempenho tanto na otimização (expresso pelo valor de *fitness*) quanto em relação à qualidade estrutural (medido por meio do RMSD). No operador de recombinação, novamente a utilização de informação do problema se mostrou eficiente. Neste quesito, uma combinação entre dois outros modelos se mostrou equilibrada o suficiente para atingir bons resultados em *fitness* e RMSD. Fica notório que, apesar de meta-heurísticas não necessitarem de conhecimento específico para serem utilizadas, agregar tais informações ao processo de tomada de decisão pode resultar em um ganho significativo de desempenho.

Na etapa final de testes, a comparação feita em relação às duas abordagens voltadas ao PSP prova que o método desenvolvido apesar de também solucionar tal problema, pode ser considerado competitivo, alcançando resultados de grande qualidade em diversos casos. Fica claro que a informação da estrutura secundária é um fator imprescindível para o processo de predição da estrutura tridimensional, garantindo grandes saltos de qualidade quando informados corretamente. Apesar disso, mesmo nos testes utilizando a informação da estrutura secundária predita, o método proposto foi capaz de se igualar e até mesmo superar o *Rosetta*, uma das principais técnicas de predição da atualidade.

Dessa forma, julga-se que o objetivo desta dissertação tenha sido alcançado, mostrando que mecanismos de prevenção de convergência prematura baseados em controle de diversidade, em conjunto com conhecimento específico do problema podem ser fatores determinantes para a obtenção de soluções de qualidade em um método de otimização. Como principais contribuições temos: (*i*) o desenvolvimento de uma política de migração capaz de manter e explorar a diversidade e; (*ii*) o desenvolvimento de operadores que incorporem conhecimento específico do problema de maneira eficiente.

Como trabalhos futuros entende-se que seja importante avaliar o impacto de demais opções de meta-heurísticas aplicadas de maneira personalizada, buscando sempre a melhor adequação ao problema PSP. A inclusão de mecanismos de busca local surge

como uma opção de refinamento das soluções, uma vez que os indivíduos gerados pelo operador de inicialização baseia-se em fragmentos de tamanho 3 e 9, impossibilitando que certas regiões possam ser alteradas de modo local (e.g., apenas um resíduo por vez). Uma alterativa seria a investigação de novas bases de conhecimento, buscando até mesmo a combinação de múltiplas fontes de conhecimento experimental. Apesar dos operadores propostos se provarem eficientes, entende-se também que a inclusão de informação do problema é um fator de enorme potencial, havendo ainda inúmeras possibilidades de incrementos ao método proposto.

7 PUBLICAÇÕES E PRODUÇÃO TÉNICA

Este capítulo apresentará os trabalhos desenvolvidos durante o Mestrado, envolvendo as áreas de otimização, meta-heurísticas distribuídas e predição de estruturas tridimensional de proteínas.

7.1 Trabalhos Completos Publicados em Anais de Eventos

- **ALIXANDRE, B. F. F.**; DORN, M.. D-BRKGA: A Distributed Biased Random-Key Genetic Algorithm. In: IEEE Congress on Evolutionary Computation, 2017, Donostia. Proceedings of the IEEE Congress on Evolutionary Computation (IEEE CEC 2017), 2017. p. 1398 - [Qualis A2].

7.2 Artigos em Preparação

- **ALIXANDRE, B. F. F.**; DORN, M.. A Fitness-Based Migration Policy to Explore Diversity in Island Model. Artigo pronto, em processo de submissão. Pretende-se enviá-lo ao periódico *International Transactions in Operational Research* - [Qualis A2]. Este artigo busca apresentar a política de migração proposta nesta dissertação, de modo a destacar sua eficiência em relação à outras políticas igualmente baseadas em diversidade.
- **ALIXANDRE, B. F. F.**; DORN, M.. (Título à definir). Desenvolvimento prático e experimentação concluídas, em processo de escrita. Pretende-se enviá-lo ao Congresso *INTERNATIONAL CONFERENCE ON PARALLEL PROBLEM SOLVING FROM NATURE - (PPSN)* - [Qualis A2]. Este artigo visa demostrar a aplicação da técnica proposta no artigo "*A Fitness-Based Migration Policy to Explore Diversity in Island Model*"no problema PSP, juntamente com maneiras eficientes de de combinar conhecimento específico do problema ao processo de otimização.

REFERÊNCIAS

ABUAL-RUB, M. S. et al. A hybrid harmony search algorithm for ab initio protein tertiary structure prediction. **Network Modeling Analysis in Health Informatics and Bioinformatics**, Springer, v. 1, n. 3, p. 69–85, 2012.

ALBA, E.; TROYA, J. M. A survey of parallel distributed genetic algorithms. **Complexity**, v. 4, n. 4, p. 31–52, 1999.

ALIXANDRE, B. F. d. F.; DORN, M. D-brkga: a distributed biased random-key genetic algorithm. In: EVOLUTIONARY COMPUTATION (CEC), 2017 IEEE CONGRESS ON. **Proceedings...** [S.l.]: IEEE, 2017. p. 1398–1405.

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

ARAUJO, L.; MERELO, J. J. Diversity through multiculturality: Assessing migrant choice policies in an island model. **IEEE Transactions on Evolutionary Computation**, IEEE, v. 15, n. 4, p. 456–469, 2011.

AWAD M. Z. ALI, P. N. S. J. J. L. N. H.; QU, B. Y. Problem definitions and evaluation criteria for the cec 2017 special session and competition on single objective bound constrained real-parameter numerical optimization. **Nanyang Technological University, Singapore, Tech. Rep.**, 2016.

BANITALEBI, A.; AZIZ, M. I. A.; AZIZ, Z. A. A self-adaptive binary differential evolution algorithm for large scale binary optimization problems. **Information Sciences**, Elsevier, v. 367, p. 487–511, 2016.

BEAN, J. C. Genetic algorithms and random keys for sequencing and optimization. **ORSA journal on computing**, INFORMS, v. 6, n. 2, p. 154–160, 1994.

BELDING, T. C. The distributed genetic algorithm revisited. **arXiv preprint adap-org/9504007**, 1995.

BERMAN, H. M. et al. The protein data bank, 1999–. In: **International Tables for Crystallography Volume F: Crystallography of biological macromolecules**. [S.l.]: Springer, 2006. p. 675–684.

BIRATTARI, M. et al. Classification of metaheuristics and design of experiments for the analysis of components tech. rep. aida-01-05. 2001.

BORGUESAN, B.; INOSTROZA-PONTA, M.; DORN, M. Nias-server: Neighbors influence of amino acids and secondary structures in proteins. **Journal of Computational Biology**, Mary Ann Liebert, Inc. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA, v. 24, n. 3, p. 255–265, 2017.

BORGUESAN, B. et al. Apl: An angle probability list to improve knowledge-based metaheuristics for the three-dimensional protein structure prediction. **Computational biology and chemistry**, Elsevier, v. 59, p. 142–157, 2015.

BORGUESAN, B. et al. Apl: An angle probability list to improve knowledge-based metaheuristics for the three-dimensional protein structure prediction. **Comput. Biol. Chem.**, Elsevier, v. 59, p. 142–157, 2015.

BOUSSAÏD, I.; LEPAGNOT, J.; SIARRY, P. A survey on optimization metaheuristics. **Information Sciences**, Elsevier, v. 237, p. 82–117, 2013.

BOWIE, J. U.; LUTHY, R.; EISENBERG, D. A method to identify protein sequences that fold into a known three-dimensional structure. **Science**, American Association for the Advancement of Science, v. 253, n. 5016, p. 164–170, 1991.

BRANDEN, C.; TOOZE, J. **Introduction to protein structure**. 2. ed. New York, USA: Garland Science, 1999. 410 p.

CANTÚ-PAZ, E. A survey of parallel genetic algorithms. **Calculateurs paralleles, reseaux et systems repartis**, v. 10, n. 2, p. 141–171, 1998.

CARUGO, O. How root-mean-square distance (rmsd) values depend on the resolution of protein structures that are compared. **Journal of applied crystallography**, International Union of Crystallography, v. 36, n. 1, p. 125–128, 2003.

CAVANAGH, J. et al. **Protein NMR spectroscopy: principles and practice**. 1. ed. New York, USA: Academic Press, 1995.

CHAUDHURY, S.; LYSKOV, S.; GRAY, J. J. Pyrosetta: a script-based interface for implementing molecular modeling algorithms using rosetta. **Bioinformatics**, Oxford University Press, v. 26, n. 5, p. 689–691, 2010.

COHOON, J. P. et al. Punctuated equilibria: a parallel genetic algorithm. In: INTERNATIONAL CONFERENCE ON GENETIC ALGORITHMS AND THEIR APPLICATIONS. **Proceedings...** [S.l.]: Hillsdale, NJ: L. Erlhaum Associates, 1987., 1987. p. –.

COHOON, J. P. et al. Distributed genetic algorithms for the floorplan design problem. **IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems**, IEEE, v. 10, n. 4, p. 483–492, 1991.

CONWAY, P. et al. Relaxation of backbone bond geometry improves protein energy landscape modeling. **Protein Science**, Wiley Online Library, v. 23, n. 1, p. 47–55, 2014.

CORRÊA, L. et al. A memetic algorithm for 3-d protein structure prediction problem. **IEEE/ACM transactions on computational biology and bioinformatics**, IEEE, 2016.

CORRÊA, L. de L. et al. Three-dimensional protein structure prediction based on memetic algorithms. **Computers & Operations Research**, Elsevier, v. 91, p. 160–177, 2018.

CORRÊA, L. de L.; DORN, M. Multi-agent systems in three-dimensional protein structure prediction. **Multi-Agent-Based Simulations Applied to Biological and Environmental Systems**, IGI Global, p. 241, 2016.

CORRÊA, L. de L.; INOSTROZA-PONTA, M.; DORN, M. An evolutionary multi-agent algorithm to explore the high degree of selectivity in three-dimensional protein structures. In: EVOLUTIONARY COMPUTATION (CEC), 2017 IEEE CONGRESS ON. **Proceedings...** [S.l.]: IEEE, 2017. p. 1111–1118.

CRAINIC, T. G.; TOULOUSE, M. Parallel strategies for meta-heuristics. In: **Handbook of metaheuristics**. Boston, MA: Springer US, 2003. p. 475–513. ISBN 978-0-306-48056-0. Available from Internet: <https://doi.org/10.1007/0-306-48056-5_17">.

CREIGHTON, T. E. Protein folding. **Biochemical journal**, Portland Press Ltd, v. 270, n. 1, p. 1, 1990.

ČREPINŠEK, M.; LIU, S.-H.; MERNIK, M. Exploration and exploitation in evolutionary algorithms: a survey. **ACM Computing Surveys (CSUR)**, ACM, v. 45, n. 3, p. 35, 2013.

CRESCENZI, P. et al. On the complexity of protein folding. **J. Comput. Biol.**, v. 5, n. 3, p. 423–465, 1998.

DEB, K. et al. A fast and elitist multiobjective genetic algorithm: Nsga-ii. **IEEE transactions on evolutionary computation**, IEEE, v. 6, n. 2, p. 182–197, 2002.

DENZINGER, J.; KIDNEY, J. Improving migration by diversity. In: EVOLUTIONARY COMPUTATION, 2003. CEC'03. THE 2003 CONGRESS ON. **Proceedings...** [S.l.]: IEEE, 2003. v. 1, p. 700–707.

DORIGO, M.; MANIEZZO, V.; COLORNI, A. The ant system: Optimization by a colony of cooperating agents. **IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS-PART B**, v. 26, n. 1, p. 29–41, 1996.

DORN, M. et al. A knowledge-based genetic algorithm to predict three-dimensional structures of polypeptides. In: EVOLUTIONARY COMPUTATION (CEC), 2013 IEEE CONGRESS ON. **Proceedings...** [S.l.]: IEEE, 2013. p. 1233–1240.

DORN, M. et al. A knowledge-based genetic algorithm to predict three-dimensional structures of polypeptides. In: CONGRESS ON EVOLUTIONARY COMPUTATION (CEC). **Proceedings...** [S.l.]: IEEE, 2013. p. 1233–1240.

DORN, M. et al. Three-dimensional protein structure prediction: methods and computational strategies. **Comput. Biol. Chem.**, Elsevier, v. 53, p. 251–276, 2014.

DORN, M. et al. Three-dimensional protein structure prediction: methods and computational strategies. **Comput. Biol. Chem.**, Elsevier, v. 53, p. 251–276, 2014.

ELOFSSON, A.; GRAND, S. M. L.; EISENBERG, D. Local moves: An efficient algorithm for simulation of protein folding. **Proteins: Struct., Funct., Bioinf.**, Wiley Online Library, v. 23, n. 1, p. 73–82, 1995.

ELSAYED, S. M.; SARKER, R. A.; ESSAM, D. L. A new genetic algorithm for solving optimization problems. **Engineering Applications of Artificial Intelligence**, Elsevier, v. 27, p. 57–69, 2014.

ENGH, R. A.; HUBER, R. Accurate bond and angle parameters for x-ray protein structure refinement. **Acta Crystallographica Section A: Foundations of Crystallography**, International Union of Crystallography, v. 47, n. 4, p. 392–400, 1991.

FASMAN, G. D. **Prediction of protein structure and the principles of protein conformation**. [S.l.]: Springer Science & Business Media, 1989.

FONSECA, R.; PALUSZEWSKI, M.; WINTER, P. Protein structure prediction using bee colony optimization metaheuristic. **J. Math. Model. Algo.**, Springer, v. 9, n. 2, p. 181–194, 2010.

GENDREAU, M.; POTVIN, J.-Y. Metaheuristics in combinatorial optimization. **Annals of Operations Research**, v. 140, n. 1, p. 189–213, Nov 2005. ISSN 1572-9338. Available from Internet: <https://doi.org/10.1007/s10479-005-3971-7>.

GIBAS, C.; JAMBECK, P. **Developing bioinformatics computer skills**. 1. ed. Sebastopol, USA: O'Reilly Media, Inc., 2001. 448 p.

GLOVER, F. Future paths for integer programming and links to artificial intelligence. **Computers & operations research**, Elsevier, v. 13, n. 5, p. 533–549, 1986.

GOLBERG, D. E. Genetic algorithms in search, optimization, and machine learning. **Addion wesley**, v. 1989, p. 102, 1989.

GONÇALVES, J. F.; RESENDE, M. G. Biased random-key genetic algorithms for combinatorial optimization. **Journal of Heuristics**, Springer, v. 17, n. 5, p. 487–525, 2011.

GRONT, D. et al. Generalized fragment picking in rosetta: design, protocols and applications. **PloS one**, Public Library of Science, v. 6, n. 8, p. e23294, 2011.

HEINIG, M.; FRISHMAN, D. Stride: a web server for secondary structure assignment from known atomic coordinates of proteins. **Nucleic acids research**, Oxford University Press, v. 32, n. suppl_2, p. W500–W502, 2004.

HOLLAND, J. H. **Adaptation in Natural and Artificial Systems**. [S.l.]: University of Michigan Press, 1975.

HOLLAND, J. H. **Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence.** [S.l.]: U Michigan Press, 1975.

HOVMOLLER, T.; OHLSON, T. Conformation of amino acids in protein. **Acta Crystallogr.**, v. 58, n. 5, p. 768–776, 2002.

HUANG, T.-Y.; CHEN, Y.-Y. Diversity-based selection pooling scheme in evolution strategies. In: SYMPOSIUM ON APPLIED COMPUTING. **Proceedings...** [S.l.]: ACM, 2001. p. 351–355.

JONES, D. T. Protein secondary structure prediction based on position-specific scoring matrices. **Journal of molecular biology**, Elsevier, v. 292, n. 2, p. 195–202, 1999.

KABSCH, W.; SANDER, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. **Biopolymers**, Wiley Online Library, v. 22, n. 12, p. 2577–2637, 1983.

KARABOGA, D.; BASTURK, B. A powerful and efficient algorithm for numerical function optimization: artificial bee colony (abc) algorithm. **J. Global Optim.**, Springer, v. 39, n. 3, p. 459–471, 2007.

KARPLUS, K. et al. What is the value added by human intervention in protein structure prediction? **Proteins: Structure, Function, and Bioinformatics**, Wiley Online Library, v. 45, n. S5, p. 86–91, 2001.

KENNEDY, J.; EBERHART, R. C. Particle swarm optimization. In: INTERNATIONAL CONFERENCE ON NEURAL NETWORKS. **Proceedings...** [S.l.]: IEEE, 1995. p. 1942–1948.

KIM, D. E. et al. Sampling bottlenecks in de novo protein structure prediction. **J. Mol. Biol.**, Elsevier, v. 393, n. 1, p. 249–260, 2009.

KIM, D. E.; CHIVIAN, D.; BAKER, D. Protein structure prediction and analysis using the robetta server. **Nucleic acids research**, Oxford University Press, v. 32, n. suppl_2, p. W526–W531, 2004.

KINCH, L. N. et al. Evaluation of free modeling targets in casp11 and roll. **Proteins: Structure, Function, and Bioinformatics**, Wiley Online Library, v. 84, n. S1, p. 51–66, 2016.

KIRKPATRICK, S.; GELATT, C. D.; VECCHI, M. P. Optimization by simulated annealing. **SCIENCE**, v. 220, n. 4598, p. 671–680, 1983.

KUFAREVA, I.; ABAGYAN, R. Methods of protein structure comparison. In: **Homology Modeling**. [S.l.]: Springer, 2011. p. 231–257.

LEHNINGER, A.; NELSON, D.; COX, M. **Principles of Biochemistry**. 4. ed. New York, USA: W.H. Freeman, 2005. 1100 p.

LESK, A. **Introduction to protein science: architecture, function, and genomics**. S1. [S.l.]: Oxford university press, 2010. – p.

LESK, A. **Introduction to bioinformatics**. 4. ed. Oxford, UK: Oxford University Press, 2013. 371 p. ISBN 9780199651566.

LEUNG, Y.-W.; WANG, Y. An orthogonal genetic algorithm with quantization for global numerical optimization. **IEEE Transactions on Evolutionary computation**, IEEE, v. 5, n. 1, p. 41–53, 2001.

LEVINTHAL, C. Are there pathways for protein folding? **J. Chim. Phys. Phys.-Chim. Biol.**, v. 65, n. 1, p. 44–45, 1968.

LILJAS, A. et al. **Textbook of structural biology**. [S.l.]: World Scientific, 2009.

LIWO, A. et al. Protein structure prediction by global optimization of a potential energy function. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 96, n. 10, p. 5482–5485, 1999.

LODISH, H. et al. **Molecular Cell Biology**. 5. ed. New York, USA: Scientific American Books, W.H. Freeman, 1990. 970 p.

LORASCHI, A. et al. Distributed genetic algorithms with an application to portfolio selection problems. In: ARTIFICIAL NEURAL NETS AND GENETIC ALGORITHMS. **Proceedings...** [S.l.]: Springer, 1995. p. 384–387.

LOZANO, M.; HERRERA, F.; CANO, J. R. Replacement strategies to preserve useful diversity in steady-state genetic algorithms. **Information Sciences**, Elsevier, v. 178, n. 23, p. 4421–4433, 2008.

LUKE, S. **Essentials of Metaheuristics**. 2. ed. [S.l.]: Lulu, 2013. Available for free at http://cs.gmu.edu/∼sean/book/metaheuristics/.

LUKE, S. **Essentials of metaheuristics**. [S.l.]: Lulu Com, 2013.

LUQUE, G.; ALBA, E. **Parallel genetic algorithms: Theory and real world applications**. [S.l.]: Springer, 2011.

MANIEZZO, V.; STÜTZLE, T.; VOSS, S. **Matheuristics: Hybridizing Metaheuristics and Mathematical Programming**. Springer US, 2009. (Annals of Information Systems). ISBN 9781441913067. Available from Internet: <https://books.google.com.br/books?id=3uN95LwxRlAC>.

MARTÍ-RENOM, M. A. et al. Comparative protein structure modeling of genes and genomes. **Annu. Rev. Biophys. Biomol. Struct.**, Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA, v. 29, n. 1, p. 291–325, 2000.

MAULIK, U.; BANDYOPADHYAY, S. Genetic algorithm-based clustering technique. **Pattern recognition**, Elsevier, v. 33, n. 9, p. 1455–1465, 2000.

MCREE, D. **Practical protein crystallography**. 1. ed. London, UK: Academic press, 1999.

MEILER, J. et al. Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks. **Molecular modeling annual**, Springer, v. 7, n. 9, p. 360–369, 2001.

MLADENOVIC, N. A variable neighborhood algorithm-a new metaheuristic for combinatorial optimization. In: OPTIMIZATION DAYS. **Abstract of papers presented at...** [S.l.], 1995. p. 112.

NELSON, D. L.; COX, M. M.; LEHNINGER, A. L. Principles of biochemistry. **WH Freeman and Company, New York, fourth edition edition**, v. 1, n. 1.1, p. 2, 2005.

NEUMAIER, A. Molecular modeling of proteins and mathematical prediction of protein structure. **SIAM review**, SIAM, v. 39, n. 3, p. 407–460, 1997.

NGO, J. T.; MARKS, J.; KARPLUS, M. Computational complexity, protein structure prediction, and the levinthal paradox. In: MERZ, K.; LEGRAND, S. M. (Ed.). **The protein folding problem and tertiary structure prediction**. [S.l.]: Springer, 1997. p. 435–508.

OLIVEIRA, M.; BORGUESAN, B.; DORN, M. Sade-spl: A self-adapting differential evolution algorithm with a loop structure pattern library for the psp problem. In: EVOLUTIONARY COMPUTATION (CEC), 2017 IEEE CONGRESS ON. **Proceedings...** [S.l.]: IEEE, 2017. p. 1095–1102.

OSGUTHORPE, D. J. Ab initio protein folding. **Curr. Opin. Struct. Biol.**, Elsevier, v. 10, n. 2, p. 146–152, 2000.

PAULING, L.; COREY, R. B. The pleated sheet, a new layer configuration of polypeptide chains. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 37, n. 5, p. 251–256, 1951.

PAULING, L.; COREY, R. B.; BRANSON, H. R. The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 37, n. 4, p. 205–211, 1951.

PETTEY, C. B.; LEUZE, M. R.; GREFENSTETTE, J. J. Parallel genetic algorithm. In: INTERNATIONAL CONFERENCE ON GENETIC ALGORITHMS AND THEIR APPLICATIONS. **Proceedings...** [S.l.]: Hillsdale, NJ: L. Erlhaum Associates., 1987. p. –.

POWER, D.; RYAN, C.; AZAD, R. M. A. Promoting diversity using migration strategies in distributed genetic algorithms. In: IEEE. **Evolutionary Computation, 2005. The 2005 IEEE Congress on**. [S.l.], 2005. v. 2, p. 1831–1838.

RANGWALA, H.; KARYPIS, G. Introduction to protein structure prediction. In: **Introduction to Protein Structure Prediction**. [S.l.]: John Wiley & Sons, Inc. New York, 2010. p. 1–13.

REID, R. **Peptide and protein drug analysis**. [S.l.]: CRC Press, 1999.

RICHARDSON, J. S. The anatomy and taxonomy of protein structure. **Advances in protein chemistry**, Elsevier, v. 34, p. 167–339, 1981.

ROHL, C. A. et al. Protein structure prediction using rosetta. **Methods in enzymology**, Elsevier, v. 383, p. 66–93, 2004.

SALEH, S.; OLSON, B.; SHEHU, A. A population-based evolutionary search approach to the multiple minima problem in de novo protein structure prediction. **BMC Struct. Biol.**, BioMed Central, v. 13, n. Suppl 1, p. S4, 2013.

SCHEEF, E. D.; FINK, J. L. Fundamentals of protein structure. In: ____. **Structural Bioinformatics**. 1. ed. New York: John Wiley and Sons Inc., 2005. v. 44, chp. 2, p. 15–39.

SCHEEFF, E. D.; FINK, J. L. Fundamentals of protein structure. In: **Structural Bioinformatics**. [S.l.]: Wiley-Blackwell, Chichester, UK, 2005. p. 15–39.

SHMYGELSKA, A.; LEVITT, M. Generalized ensemble methods for de novo structure prediction. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 106, n. 5, p. 1415–1420, 2009.

SIMONS, K. T. et al. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. **Journal of molecular biology**, Elsevier, v. 268, n. 1, p. 209–225, 1997.

SIMONS, K. T. et al. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. **Proteins: Structure, Function, and Bioinformatics**, Wiley Online Library, v. 34, n. 1, p. 82–95, 1999.

SONG, Y. et al. High-resolution comparative modeling with rosettacm. **Structure**, Elsevier, v. 21, n. 10, p. 1735–1742, 2013.

SRINIVAS, M.; PATNAIK, L. M. Genetic algorithms: A survey. **Computer**, IEEE, v. 27, n. 6, p. 17–26, 1994.

SRINIVASAN, R.; ROSE, G. D. Linus: a hierarchic procedure to predict the fold of a protein. **Proteins: Struct., Funct., Bioinf.**, Wiley Online Library, v. 22, n. 2, p. 81–99, 1995.

STARKWEATHER, T.; WHITLEY, D.; MATHIAS, K. Optimization using distributed genetic algorithms. In: INTERNATIONAL CONFERENCE ON PARALLEL PROBLEM SOLVING FROM NATURE. **Proceedings...** [S.l.]: Springer, 1990. p. 176–185.

STILL, W. C. et al. Semianalytical treatment of solvation for molecular mechanics and dynamics. **Journal of the American Chemical Society**, ACS Publications, v. 112, n. 16, p. 6127–6129, 1990.

TAKAHASHI, R. Verification of thermo-dynamical genetic algorithm to solve the function optimization problem through diversity measurement - diversity measurement and its application to selection strategies in genetic algorithms. In: EVOLUTIONARY COMPUTATION (CEC), 2016 IEEE CONGRESS ON. **Proceedings...** [S.l.]: IEEE, 2016. p. 168–177.

TALBI, E.-G. **Metaheuristics: from design to implementation**. [S.l.]: John Wiley & Sons, 2009.

TANESE, R. Distributed genetic algorithms. In: INTERNATIONAL CONFERENCE ON GENETIC ALGORITHMS. **Proceedings...** [S.l.]: Morgan Kaufmann Publishers Inc., 1989. p. 434–439.

TANTAR, A.-A.; MELAB, N.; TALBI, E.-G. A grid-based genetic algorithm combined with an adaptive simulated annealing for protein structure prediction. **Soft Computing**, Springer, v. 12, n. 12, p. 1185–1198, 2008.

TANTAR, A.-A. et al. A parallel hybrid genetic algorithm for protein structure prediction on the computational grid. **Future Generation Computer Systems**, Elsevier, v. 23, n. 3, p. 398–409, 2007.

TANTAR, A.-A. et al. A parallel hybrid genetic algorithm for protein structure prediction on the computational grid. **Future Generation Computer Systems**, Elsevier, v. 23, n. 3, p. 398–409, 2007.

TOFFOLO, A.; BENINI, E. Genetic diversity as an objective in multi-objective evolutionary algorithms. **Evolutionary computation**, MIT Press, v. 11, n. 2, p. 151–167, 2003.

TRAMONTANO, A.; LESK, A. M. **Protein structure prediction**. 1. ed. Weinheim, Germany: John Wiley and Sons Inc., 2006. 208 p.

UNGER, R. The genetic algorithm approach to protein structure prediction. **Applications of Evolutionary Computation in Chemistry**, Springer, p. 2697–2699, 2004.

URSEM, R. K. Diversity-guided evolutionary algorithms. In: INTERNATIONAL CONFERENCE ON PARALLEL PROBLEM SOLVING FROM NATURE. **Proceedings...** Springer Berlin Heidelberg, 2002. p. 462–471. ISBN 978-3-540-45712-1. Available from Internet: <https://doi.org/10.1007/3-540-45712-7_45>.

VALENCIA, A.; PAZOS, F. Computational methods for the prediction of protein interactions. **Current opinion in structural biology**, Elsevier, v. 12, n. 3, p. 368–373, 2002.

VERLI, H. O que é bioinformática? In: **Bioinformática: da Biologia à Flexibilidade Moleculares**. 1. ed. São Paulo, Brasil: Sociedade Brasileira de Bioquímica e Biologia Molecular-SBBq, 2014. chp. 1, p. 1–12.

VIKTORIN, A.; PLUHACEK, M.; SENKERIK, R. Multi-chaotic system induced success-history based adaptive differential evolution. In: INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND SOFT COMPUTING. **Proceedings...** [S.l.]: Springer, 2016. p. 517–527.

WHITLEY, D. A genetic algorithm tutorial. **Statistics and Computing**, Springer, v. 4, n. 2, p. 65–85, Jun 1994. ISSN 1573-1375.

WHITLEY, D.; STARKWEATHER, T. Genitor ii: A distributed genetic algorithm. **Journal of Experimental & Theoretical Artificial Intelligence**, Taylor & Francis, v. 2, n. 3, p. 189–214, 1990.

WHITLEY, L. D. et al. The genitor algorithm and selection pressure: Why rank-based allocation of reproductive trials is best. In: INTERNATIONAL CONFERENCE ON GENETIC ALGORITHMS. **Proceedings...** [S.l.], 1989. v. 89, p. 116–123.

WOOLEY, J. C.; YE, Y. A historical perspective and overview of protein structure prediction. In: **Computational methods for protein structure prediction and modeling**. [S.l.]: Springer, 2007. p. 1–43.

YANG, X.-S. **Nature-inspired metaheuristic algorithms**. [S.l.]: Luniver press, 2010.

ZHANG, Y.; SKOLNICK, J. Scoring function for automated assessment of protein structure template quality. **Proteins: Structure, Function, and Bioinformatics**, Wiley Online Library, v. 57, n. 4, p. 702–710, 2004.