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

LEONARDO DE LIMA CORRÊA

Uma Proposta de Algoritmo Memético Baseado em Conhecimento para o Problema de Predição de Estruturas 3-D 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
2017

CIP — CATALOGAÇÃO NA PUBLICAÇÃO

Corrêa, Leonardo de Lima

Uma Proposta de Algoritmo Memético Baseado em Conhecimento para o Problema de Predição de Estruturas 3-D de Proteínas / Leonardo de Lima Corrêa. – Porto Alegre: PPGC da UFRGS, 2017.

154 f.: il.

Dissertação (mestrado) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR–RS, 2017. Orientador: Márcio Dorn.

1. Otimização. 2. Meta-heurísticas. 3. Algoritmos evolutivos. 4. Algoritmo baseado em conhecimento. 5. Bioinformática estrutural. 6. Algoritmos meméticos. I. Dorn, Márcio. II. Título.

RESUMO

Algoritmos meméticos são meta-heurísticas evolutivas voltadas intrinsecamente à exploração e incorporação de conhecimentos relacionados ao problema em estudo. Nesta dissertação, foi proposto um algoritmo memético multi populacional baseado em conhecimento para lidar com o problema de predição de estruturas tridimensionais de proteínas voltado à modelagem de estruturas livres de similaridades conformacionais com estruturas de proteínas determinadas experimentalmente. O algoritmo em questão, foi estruturado em duas etapas principais de processamento: (*i*) amostragem e inicialização de soluções; e (*ii*) otimização dos modelos estruturais provenientes da etapa anterior. A etapa I objetiva a geração e classificação de diversas soluções, a partir da estratégia Lista de Probabilidades Angulares, buscando a definição de diferentes grupos estruturais e a criação de melhores estruturas a serem incorporadas à meta-heurística como soluções iniciais das multi populações. A segunda etapa consiste no processo de otimização das estruturas oriundas da etapa I, realizado por meio da aplicação do algoritmo memético de otimização, o qual é fundamentado na organização da população de indivíduos em uma estrutura em árvore, onde cada nodo pode ser interpretado como uma subpopulação independente, que ao longo do processo interage com outros nodos por meio de operações de busca global voltadas a características do problema, visando o compartilhamento de informações, a diversificação da população de indivíduos, e a exploração mais eficaz do espaço de busca multimodal do problema. O algoritmo engloba ainda uma implementação do algoritmo colônia artificial de abelhas, com o propósito de ser utilizado como uma técnica de busca local a ser aplicada em cada nodo da árvore. O algoritmo proposto foi testado em um conjunto de 24 sequências de aminoácidos, assim como comparado a dois métodos de referência na área de predição de estruturas tridimensionais de proteínas, Rosetta e QUARK. Os resultados obtidos mostraram a capacidade do método em predizer estruturas tridimensionais de proteínas com conformações similares a estruturas determinadas experimentalmente, em termos das métricas de avaliação estrutural *Root-Mean-Square Deviation* e *Global Distance Total Score Test*. Verificou-se que o algoritmo desenvolvido também foi capaz de atingir resultados comparáveis ao Rosetta e ao QUARK, sendo que em alguns casos, os superou. Corroborando assim, a eficácia do método.

Palavras-chave: Otimização. meta-heurísticas. algoritmos evolutivos. algoritmo baseado em conhecimento. bioinformática estrutural. algoritmos meméticos.

ABSTRACT

Memetic algorithms are evolutionary metaheuristics intrinsically concerned with the exploiting and incorporation of all available knowledge about the problem under study. In this dissertation, we present a knowledge-based memetic algorithm to tackle the three-dimensional protein structure prediction problem without the explicit use of template experimentally determined structures. The algorithm was divided into two main steps of processing: (*i*) sampling and initialization of the algorithm solutions; and (*ii*) optimization of the structural models from the previous stage. The first step aims to generate and classify several structural models for a determined target protein, by the use of the strategy Angle Probability List, aiming the definition of different structural groups and the creation of better structures to initialize the initial individuals of the memetic algorithm. The Angle Probability List takes advantage of structural knowledge stored in the Protein Data Bank in order to reduce the complexity of the conformational search space. The second step of the method consists in the optimization process of the structures generated in the first stage, through the applying of the proposed memetic algorithm, which uses a tree-structured population, where each node can be seen as an independent subpopulation that interacts with others, over global search operations, aiming at information sharing, population diversity, and better exploration of the multimodal search space of the problem. The method also encompasses ad-hoc global search operators, whose objective is to increase the exploration capacity of the method turning to the characteristics of the protein structure prediction problem, combined with the Artificial Bee Colony algorithm to be used as a local search technique applied to each node of the tree. The proposed algorithm was tested on a set of 24 amino acid sequences, as well as compared with two reference methods in the protein structure prediction area, Rosetta and QUARK. The results show the ability of the method to predict three-dimensional protein structures with similar foldings to the experimentally determined protein structures, regarding the structural metrics *Root-Mean-Square Deviation* and *Global Distance Total Score Test*. We also show that our method was able to reach comparable results to Rosetta and QUARK, and in some cases, it outperformed them, corroborating the effectiveness of our proposal.

Keywords: Optimization, metaheuristics, evolutionary algorithms, knowledge based algorithm, structural bioinformatics, memetic algorithms.

LISTA DE ABREVIATURAS E SIGLAS

3-D	Tridimensional
RMN	Ressonância Magnética Nuclear
ME	Microscopia eletrônica
PSP	*Three-dimensional protein structure prediction*
PDB	*Protein Data Bank*
CASP	*Critical Assessment of Protein Structure Prediction*
AM	Algoritmo memético
EP	Estrutura primária
ES	Estrutura secundária
ET	Estrutura terciária
IDP	*Intrinsically disordered proteins*
SCOP	*Structural Classification of Proteins*
RG	Raio de giro
SASA	*Solvent accessible surface area*
FM	*Free-modeling*
REMC	*Replica Exchange Monte Carlo*
AE	Algoritmo evolutivo
AG	Algoritmo genético
PSO	Algoritmo *Particle Swarm Optimization*
APL	Lista de Probabilidades Angulares
RMSD	*Root-Mean-Square Deviation*
ABC	Algoritmo *Artificial Bee Colony*
CV	Coeficiente de variação

LISTA DE SÍMBOLOS

ϕ Ângulo Phi

ψ Ângulo Psi

ω Ângulo Omega

χ Ângulo Chi

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Ao longo das últimas décadas, alguns campos de pesquisa relacionados à biologia necessitaram assumir um caráter mais tecnológico, visto os avanços computacionais alcançados, aliados à massiva geração de dados biológicos realizada a partir do sequenciamento de genomas inteiros de organismos. Este caráter, portanto, volta-se ao uso intensivo de dados, formulando partes da Bioinformática como áreas altamente computacionais, necessárias à exploração de métodos responsáveis pela extração e análise de informações oriundas de grandes bases de dados (BADER et al., 2005; LESK, 2013). A Bioinformática consiste no estudo de problemas biológicos através da criação e emprego de modelos matemáticos e técnicas computacionais (CHOU, 2004). Caracteriza-se, primeiramente, como uma importante área do grande campo das ciências biológicas, no entanto, devido à enorme complexidade envolvida nos mais diversos processos abordados, diferentes áreas do conhecimento despendem esforços para, em conjunto, compreendê-los de forma mais clara e objetiva (VERLI, 2014).

A Bioinformática pode ser dividida em duas grandes vertentes, Bioinformática baseada em sequência e Bioinformática baseada em estrutura (CHOU, 2004). A primeira linha refere-se ao sequenciamento e estudo de sequências de nucleotídeos e aminoácidos, resultantes principalmente de Projetos Genoma[1] (CONSORTIUM et al., 2015), no que concerne a aplicações de uma ampla gama de métodos analíticos, cujo objetivo é compreender de uma forma mais clara questões concernentes às suas características, funções e processos de evolução. As estratégias empregadas para tanto, envolvem, por exemplo: métodos de alinhamento de sequências, os quais visam identificar regiões de similaridade entre sequências, provenientes de possíveis relações funcionais ou evolutivas existentes entre elas; estratégias de busca em bases de dados, aplicadas, por exemplo, na tentativa de inferir a família de uma proteína a partir da sequência de aminoácidos; geração de redes metabólicas, as quais objetivam mapear o conjunto de reações e interações efetuadas por um organismo; entre outros métodos (LANDER et al., 2001; CHOU, 2004).

A segunda linha está relacionada a pesquisas com foco na estrutura tridimensional (3-D) de moléculas e macromoléculas, incluindo a predição de estruturas 3-D de proteínas (DORN et al., 2014), atracamento molecular (*docking*) (KITCHEN et al., 2004), modelagem molecular e estudos acerca das relações entre estrutura e função de proteínas (WHISSTOCK; LESK, 2003). As informações estruturais correspondentes a cada

[1] <www.genome.gov>

tipo de molécula (DNA, proteína, ligante, etc.), podem ser obtidas através da aplicação de variados métodos experimentais, tais como cristalografia de raios-X (MCREE, 1999), Ressonância Magnética Nuclear (RMN) (CAVANAGH et al., 2006) e microscopia eletrônica (ME) (UNWIN; HENDERSON, 1975).

Atualmente, dentre os diversos cenários que compreendem a Bioinformática Estrutural, muitos problemas ainda permanecem parcialmente sem solução, como é o caso dos problemas de atracamento molecular (YURIEV; HOLIEN; RAMSLAND, 2015) e predição de estruturas 3-D de proteínas (PSP, sigla em inglês) (DILL; MACCALLUM, 2012). As razões pelas quais tais problemas ainda configuram grandes desafios às suas áreas de estudo, devem-se ao elevado custo e considerável tempo demandado pelos métodos experimentais, alta complexidade computacional e também pela falta de domínio completo das regras que governam os processos bioquímicos e suas relações envolvidos nestes problemas.

Mais especificamente, a predição de estruturas 3-D de proteínas é considerada uma das mais importantes linhas de pesquisa da Bioinformática Estrutural e pode ser resumida como sendo os esforços para desenvolver métodos e estratégias computacionais para, a partir da sequência de aminoácidos, determinar a estrutura 3-D de um polipeptídeo (TRAMONTANO; LESK, 2006). As proteínas estão presentes em todos os seres vivos e exercem um enorme conjunto de funções no organismo, participando de praticamente todos os processos celulares. Conhecer a estrutura de uma proteína, permite estudar os processos biológicos a ela associados de maneira mais aprofundada, posto que a função desempenhada pela proteína está estritamente relacionada à sua conformação (LESK, 2013), e esta, por sua vez, fornece aos pesquisadores informações cruciais acerca do funcionamento da proteína na célula (BRANDEN; TOOZE, 1999; LASKOWSKI; WATSON; THORNTON, 2005). Os métodos de cristalografia de raios-X e RMN são os métodos experimentais mais comuns empregados neste processo. No entanto, estes métodos apresentam algumas desvantagens, sendo que as principais limitações consistem no elevado custo de utilização, sem garantia de sucesso, e no tempo demandado para a predição da estrutura de uma única proteína.

Nas últimas décadas, o problema PSP voltado à modelagem de estruturas livres de similaridades conformacionais com estruturas de proteínas determinadas experimentalmente (categoria *free-modeling*) (FM, sigla em inglês), tem desafiado biólogos, bioquímicos, físicos, cientistas da computação e matemáticos, permanecendo até hoje como um desafio no campo de pesquisa da Bioinformática Estrutural (BAXEVANIS; OUEL-

LETTE, 2004). O PSP pode ser classificado de acordo com a teoria da complexidade computacional (COOK, 1983) como um problema NP-difícil (UNGER; MOULT, 1993; CRESCENZI et al., 1998), devido ao extenso espaço de busca multimodal e a alta dimensionalidade de variáveis, apresentando aumento gradativo de dificuldade à medida que o número de resíduos de aminoácidos, integrantes das proteínas-alvo, aumenta (GUYEUX et al., 2014). A complexidade do problema deve-se à enorme explosão de possibilidades de formatos 3-D aceitáveis, onde uma longa cadeia de aminoácidos pode originar algumas conformações em torno de um estado nativo dentre as inúmeras conformações existentes (BAXEVANIS; OUELLETTE, 2004). Em decorrência da dificuldade em determinar as estruturas 3-D de proteínas através de métodos experimentais, originou-se uma enorme lacuna entre o volume de dados (sequências de aminoácidos não redundantes) gerados através de Projetos Genoma, os quais estão armazenados no banco de dados *NCBI Reference Sequence*[2] (*RefSeq*) (PRUITT; TATUSOVA; MAGLOTT, 2005), e o número de estruturas 3-D não redundantes conhecidas, determinadas experimentalmente e armazenadas no banco de dados *Protein Data Bank*[3] (PDB) (BERMAN et al., 2000). Atualmente, menos de 1% das sequências de proteínas conhecidas e não redundantes possuem representantes no PDB. Observa-se que as bases de dados *RefSeq* e PDB figuram hoje como as maiores e mais populares bases de dados para armazenagem de sequências de aminoácidos, genômicas de DNA e transcrições não redundantes, e de estruturas 3-D de proteínas, respectivamente.

Ao longo dos últimos anos diversos métodos e técnicas computacionais foram propostos na tentativa de enfrentar o problema. Os métodos existentes podem ser classificados em quatro classes distintas (FLOUDAS et al., 2006; DORN et al., 2014): (*i*) métodos de primeiros princípios que não utilizam informações estruturais de proteínas de bases de dados (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 (*fold recognition*) (BOWIE; LUTHY; EISENBERG, 1991); e (*iv*) métodos de modelagem comparativa (*comparative modeling*) (MARTÍ-RENOM et al., 2000).

O primeiro grupo engloba os métodos *ab initio*, caracterizados pela não utilização de informações oriundas de modelos estruturais armazenados em banco de dados de proteínas. Esta classe de métodos objetiva predizer novas conformações, a partir da sequência linear de aminoácidos, baseando-se apenas em conceitos da termodinâmica e simulações

[2]<www.ncbi.nlm.nih.gov/refseq>
[3]<www.rcsb.org>

computacionais de propriedades físico-químicas do processo de enovelamento de proteínas na natureza (OSGUTHORPE, 2000). Estes métodos são guiados pelo fato de que algumas conformações em torno de um estado nativo da proteína correspondem a regiões de mínimos globais de sua energia livre (ANFINSEN, 1973), e por isso são capazes de obter novos e desconhecidos enovelamentos. Contudo, na prática, a alta dimensionalidade e complexidade do espaço de busca conformacional tornam o problema praticamente intratável computacionalmente. Destaca-se que as proteínas não assumem apenas uma única conformação nativa, e sim um conjunto de conformações em torno de um estado nativo que corresponde a mínimos globais de energia, visto que estas encontram-se em movimento na natureza.

Por outro lado, os grupos *ii*, *iii* e *iv* são classificados como métodos baseados em conhecimento, fazendo uso de informações relacionadas aos 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. No entanto, são utilizados como forma de contornar a enorme complexidade do espaço de soluções do problema. Especificamente, o grupo *ii* representa uma classe híbrida de métodos, no qual utilizam informações de fragmentos de aminoácidos combinados a uma abordagem puramente *ab initio* (SRINIVASAN; ROSE, 1995). Nestes métodos, busca-se extrair e utilizar características mais gerais de estruturas de proteínas conhecidas, com o intuito de construir modelos (soluções) iniciais para a proteína em estudo, onde posteriormente são refinadas por métodos de otimização baseados em conceitos de primeiros princípios. Este grupo, não visa comparar a sequência inteira da proteína-alvo com estruturas conhecidas, mas apenas pequenos fragmentos da sequência de aminoácidos, na tentativa de obter informações relevantes que contribuam no processo de predição, sem prender-se totalmente às bases de dados (ROHL et al., 2004). Os algoritmos apresentados nesta dissertação pertencem a esta classe de métodos.

A aplicação e eficiência de métodos parcialmente empíricos para a predição de estruturas de proteínas é dependente da qualidade e exploração dos bancos de dados experimentais adotados, bem como da função de energia utilizada no processo, sendo esta responsável por representar as estruturas em torno de um estado nativo da proteína na forma de mínimos globais (ANFINSEN, 1973), e de uma eficiente técnica de busca capaz de lidar com milhares de conformações em uma superfície energética extremamente rugosa (multimodal) (ROHL et al., 2004; LEE; WU; ZHANG, 2009). De acordo com

as últimas edições do *Critical Assessment of Protein Structure Prediction* [4] (CASP), os melhores resultados, em relação à categoria FM, estão sendo obtidos pelos métodos de primeiros princípios baseados em conhecimento (TAI et al., 2014; MOULT et al., 2016; KINCH et al., 2016a).

Considerando que o problema PSP relacionado à FM pertence à categoria de complexidade NP-difícil, sabe-se que a aplicação de métodos exatos (determinísticos) capazes de atingir soluções ótimas, torna-se, neste caso, computacionalmente inviável, devido ao tempo de execução não-polinomial apresentado por estes. Contudo, a utilização de métodos aproximados ou heurísticas são capazes de obter soluções aproximadas para o problema, em tempo de execução aceitável (TALBI, 2009). No entanto, não há garantias de que a solução ótima será encontrada.

Dessa forma, em decorrência da enorme complexidade apresentada pelo PSP e da impossibilidade de utilização de métodos exatos, uma vasta gama de algoritmos de otimização estocástica e meta-heurísticas vêm sendo aplicados com o objetivo de obter soluções aceitáveis e aproximadas para o problema, assim como a incorporação de conhecimento prévio de estruturas 3-D de proteínas determinadas experimentalmente, que visam apoiar estas técnicas por meio da redução do tamanho e da complexidade do espaço de busca conformacional (DORN et al., 2014).

1.1 Motivação

Apesar dos avanços dos métodos computacionais para lidar com o problema PSP voltado à categoria FM, este ainda permanece como um problema em aberto em Bioinformática Estrutural. O desenvolvimento de novas estratégias, a adaptação e investigação de novos métodos em conjunto com abordagens computacionais ditas estado da arte, são claramente uma necessidade, especialmente pelo fato de que atualmente não existem métodos computacionais capazes de obter a solução ótima para o problema (BRADLEY; MISURA; BAKER, 2005; DORN et al., 2014).

Sabe-se que o sucesso da predição de estruturas 3-D de proteínas requer uma função de energia acurada, que reflita de forma satisfatória o estado nativo das proteínas, bem como um método de busca eficiente na exploração do espaço conformacional e manutenção da diversidade dos modelos gerados ao longo das simulações. Combinando estes fatores a estratégias de incorporação de conhecimento acerca de estruturas de proteínas

[4] <www.predictioncenter.org>

previamente conhecidas, e corroborando com o problema da multimodalidade da função de avaliação, complexidade e alta dimensionalidade de variáveis do espaço de soluções do problema, acredita-se que através do estudo e desenvolvimento de meta-heurísticas voltadas aos problemas de otimização em larga escala e a adaptação das mesmas para lidar com as questões endereçadas pelo PSP, seja possível extrair o potencial máximo dos métodos de busca, enquanto obtêm-se melhores resultados para o problema. Acrescenta-se ainda, que em problemas desta natureza de complexidade, por vezes, as técnicas canônicas de otimização, focadas apenas em estratégias de busca global, nem sempre se comportam da forma esperada, visto a dificuldade em localizar, refinar e manter os diversos ótimos locais e globais contidos na superfície da função de avaliação (ISLAM; CHETTY, 2009).

Neste sentido, propõe-se o desenvolvimento de um algoritmo memético (AM) (MOSCATO, 1989) multi populacional, que incorpore conceitos de meta-heurísticas evolutivas (BACK, 1996) e técnicas de busca local, aliado ao conhecimento de estruturas 3-D de proteínas armazenadas no PDB, com o objetivo de explorar de uma forma mais eficiente o espaço de busca conformacional, mantendo a diversidade de soluções, para melhor lidar com o problema PSP. Os AMs simulam o comportamento e interações de populações de indivíduos, baseando-se no conceito de "mímica"ou replicação de ideias (DAWKINS, 1976). Este conceito originou-se a partir da evolução cultural e pode ser explicado como um componente de transmissão cultural, onde ideias complexas são divididas entre agentes de determinada população que as propagam e as modificam. Interações entre membros de uma mesma população ou entre diferentes populações são simuladas através de operadores de busca global e refinamentos locais que conduzem à evolução e melhoramentos constantes de seus indivíduos. Ainda, infere-se que ideias são os resultados provenientes de operadores de busca, e como em um ambiente cultural, boas ideias tendem a sobreviver, enquanto ideias ruins desaparecem ao longo das gerações, resultando em um conjunto final de soluções aceitáveis para o problema (NERI; COTTA, 2012). Em razão da enorme complexidade apresentada pelo PSP, a escolha inicial da meta-heurística a ser aplicada foi motivada a partir da definição básica dos AMs, a qual permite uma grande flexibilização no emprego de heurísticas de otimização global e local, facilitando a exploração do espaço de busca, o refinamento dos mínimos locais encontrados e a diversificação da população (MOSCATO; COTTA, 2010). Ainda, analisando através de uma perspectiva biológica, a aplicação de AMs para o problema, favorece a exploração do espaço de soluções através da utilização de estratégias de busca global, objetivando encontrar diferentes modelos estruturais, enquanto realiza pequenos ajustes nas estruturas encontradas, por

meio de mutações e pequenas alterações locais, como forma de melhorar estes modelos, corrigindo eventuais erros conformacionais.

O trabalho fornecerá um estudo sobre alguns dos métodos mais relevantes aplicados à Bioinformática Estrutural e ao problema PSP. Define-se como foco do trabalho, a análise e estudo acerca das meta-heurísticas evolutivas, bem como a utilização de informações experimentais oriundas do PDB, implementação de um novo método baseado em conhecimento, desenvolvimento de estratégias para auxiliar na multimodalidade do problema através da manutenção da diversidade populacional do algoritmo, além da comparação com as técnicas mais relevantes já propostas na literatura, ressaltando que o problema será tratado segundo a definição da classe de métodos de primeiros princípios baseados em conhecimento, voltado à categoria FM. As principais contribuições deste trabalho são o desenvolvimento e avaliação de uma técnica de otimização robusta para enfrentar o problema, buscando contornar os principais desafios já apontados.

Por fim, destaca-se que o contexto deste trabalho insere-se no propósito de catalizar o desenvolvimento de modelos acurados de simulação computacional para predição de estruturas 3-D de proteínas que, posteriormente, facilitem a inferência de funções desempenhadas por macromoléculas e suas relações com as estruturas 3-D assumidas.

1.2 Objetivos e metas

O objetivo geral desta pesquisa trata do estudo e desenvolvimento de uma meta-heurística evolutiva, que incorpore conhecimento sob diferentes formas, visando a exploração e diversidade do modelo, aplicada ao problema multimodal de predição de estruturas 3-D de proteínas. As metas a serem alcançadas com o desenvolvimento desta dissertação e necessárias ao cumprimento do objetivo geral são:

1. Estudar as principais características do problema PSP, relacionadas à categoria FM, visando a sua modelagem como problema de otimização, bem como suas restrições, limitações e desafios;
2. Pesquisar acerca das meta-heurísticas evolutivas e técnicas de busca mais relevantes atualmente, além dos métodos que descrevem o estado da arte para a predição de estruturas 3-D de proteínas;
3. Com base no levantamento bibliográfico realizado, propor um AM robusto para o problema PSP, tendo em vista a obtenção de soluções aproximadas para o mesmo,

devido à impossibilidade de aplicação de métodos exatos neste contexto;

4. Implementar e validar a abordagem de otimização proposta para enfrentar o problema;
5. Adaptar a meta-heurística aos conceitos de multimodalidade e complexidade concernentes ao PSP e à função de energia, visando verificar e contornar algumas ineficiências existentes;
6. Avaliar questões de performance do método proposto em relação aos aspectos computacionais e significância biológica dos resultados;
7. Validar o algoritmo desenvolvido por meio da realização de diversos testes em diferentes conjuntos de proteínas, e comparação com trabalhos de relevância na área, conforme os relatórios do CASP (MOULT et al., 2016; KINCH et al., 2016a).

Em conclusão, espera-se com este trabalho, obter o êxito de gerar um modelo computacional de otimização que incorpore de maneira inteligente informações experimentais de estruturas de proteínas, robusto e acurado, que possibilite encontrar a estrutura 3-D de proteínas a partir de sequências primárias de aminoácidos.

1.3 Estrutura da dissertação

Os próximos capítulos desta dissertação serão organizados do seguinte modo:

- Capítulo 2: Fundamentação Biológica: Neste capítulo serão abordados conceitos relacionados à área de predição de estruturas de proteínas, como a composição química das proteínas, níveis de abstração e classes estruturais, modelos computacionais utilizados na representação de estruturas de proteínas, modelagem do problema em termos computacionais e matemáticos, e as funções de avaliação empregadas neste problema. O objetivo deste capítulo é apresentar e discutir os principais conceitos de cunho biológico que envolvem esta pesquisa, visando o embasamento teórico do problema PSP, bem como de questões relativas a este;
- Capítulo 3: Trabalhos Relacionados: Este capítulo apresentará uma visão geral dos trabalhos relacionados à área de otimização e meta-heurísticas evolutivas aplicadas ao problema, bem como tenciona descrever os métodos considerados estado da arte na predição de estruturas de proteínas. O objetivo do capítulo consiste na demonstração da relação entre os trabalhos relacionados apresentados e o trabalho proposto nesta dissertação;

- Capítulo 4: Materiais e Métodos: Neste capítulo serão apresentados os algoritmos e estratégias empregados na elaboração desta dissertação, bem como a estruturação do método proposto. Este capítulo tem por objetivo descrever a forma como o desenvolvimento do trabalho foi conduzido, e a metodologia utilizada para isto;
- Capítulo 5: Análise e Resultados - Etapa I: Neste capítulo serão apresentados os experimentos realizados e os resultados obtidos concernentes à etapa I de amostragem e inicialização de indivíduos, proposta como forma de prover soluções diversificadas e de qualidade à meta-heurística de otimização desenvolvida. O objetivo deste capítulo é embasar as decisões tomadas na estruturação algorítmica da etapa I do método de otimização;
- Capítulo 6: Análise e Resultados - Etapa II: Neste capítulo serão apresentados os experimentos realizados e os resultados obtidos relativos à execução da abordagem proposta, aplicada na otimização de um conjunto de testes de proteínas-alvo. O objetivo do capítulo é avaliar a performance do algoritmo desenvolvido em relação a aspectos computacionais e significância biológica dos resultados;
- Capítulo 7: Conclusões: Este capítulo apresentará as considerações finais relativas ao trabalho desenvolvido. O objetivo deste último capítulo é discutir as conclusões formuladas através dos resultados obtidos, identificar os objetivos e metas atingidos, apontar as principais dificuldades encontradas, e delinear recomendações para futuros trabalhos relacionados ao método proposto e ao problema abordado.

2 FUNDAMENTAÇÃO BIOLÓGICA

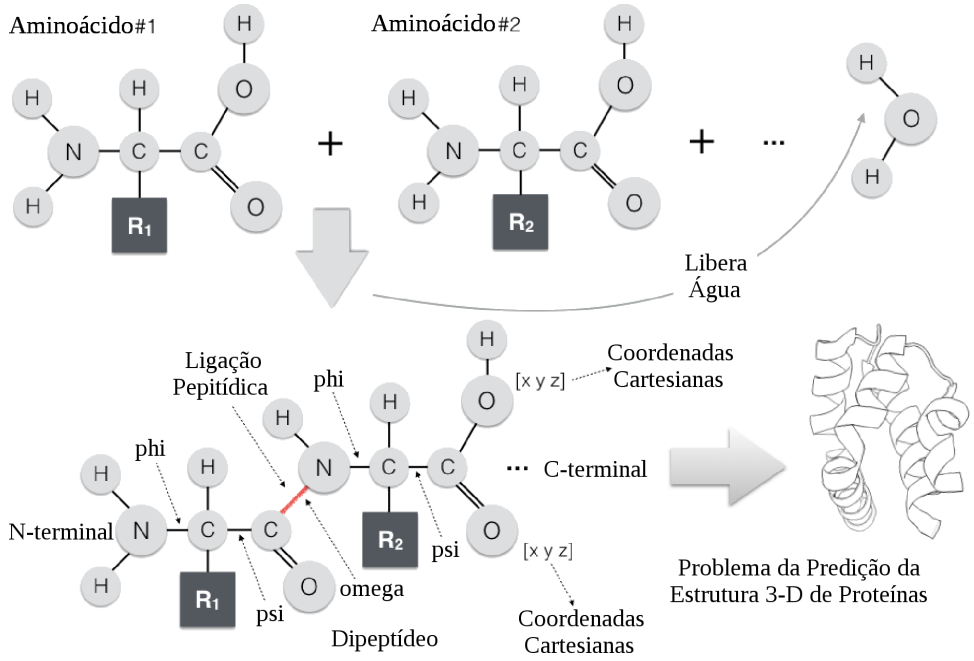
Neste capítulo serão apresentados conceitos de fundamentação biológica necessários ao entendimento da área de predição de estruturas de proteínas, que serão divididos em seções para facilitar o entendimento. O objetivo do capítulo é discutir os principais conceitos que envolvem esta pesquisa, visando o embasamento teórico do problema PSP, bem como de questões relativas a este.

2.1 Composição química das proteínas

Partindo de uma perspectiva estrutural, proteínas ou polipeptídeos podem ser vistos como cadeias lineares ordenadas de aminoácidos. Um resíduo de aminoácido (aa) é uma pequena molécula constituída por um grupo de átomos amina (NH_2), um grupo carboxila (COOH), e um átomo de hidrogênio (H) ligado a um carbono alfa central (C_α), conforme ilustrado na Figura 2.1. Esta estrutura determina a cadeia principal (*main-chain*) ou esqueleto peptídico (*backbone*) comum a todos os aminoácidos. Além disso, cada aa possui também um grupo orgânico R (cadeia lateral ou *side-chain*) ligado ao C_α da cadeia principal. O grupo R é responsável por distinguir um aa de outro, e conferir as propriedades físico-químicas específicas de cada resíduo. Na natureza, existem 20 diferentes tipos de aminoácidos codificados no genoma, onde cada um possui características próprias. As cadeias laterais dos aminoácidos podem diferir em tamanho, carga elétrica e polaridade. Ainda, dependendo da polaridade da cadeia lateral, o aa pode assumir um caráter mais hidrofóbico ou hidrofílico (LODISH et al., 2007).

Peptídeos são moléculas compostas por dois ou mais resíduos de aminoácidos ligados através de encadeamentos químicos, denominados ligações peptídicas (Fig. 2.1). A ligação peptídica ($C-N$) é formada quando o grupo carboxila de um aa reage com o grupo amina de outro aa, e libera assim uma molécula de água (H_2O). Dois ou mais resíduos de aminoácidos encadeados são referidos como um peptídeo, sendo que peptídeos maiores são chamados de polipeptídeos ou proteínas (CREIGHTON, 1990; LESK, 2013). Cada proteína é definida por uma sequência linear única de aminoácidos, responsável por determinar a sua conformação. Esta conformação ou enovelamento, concede à proteína propriedades bioquímicas específicas, que ditam o seu papel no organismo (LESK, 2010).

Figura 2.1: Representação química de dois resíduos de aminoácidos e de um modelo esquemático da formação de um peptídeo. O grupo carboxila (COOH) de um aa (1) reage com o grupo amina (NH_2) de outro aa (2), liberando assim uma molécula de água (H_2O) e originando uma ligação peptídica (C−N). N representa átomos de nitrogênio; C denota átomos de carbono; O representa partículas de oxigênio, e H são átomos de hidrogênio



Fonte: Adaptado de Corrêa e Dorn (2017).

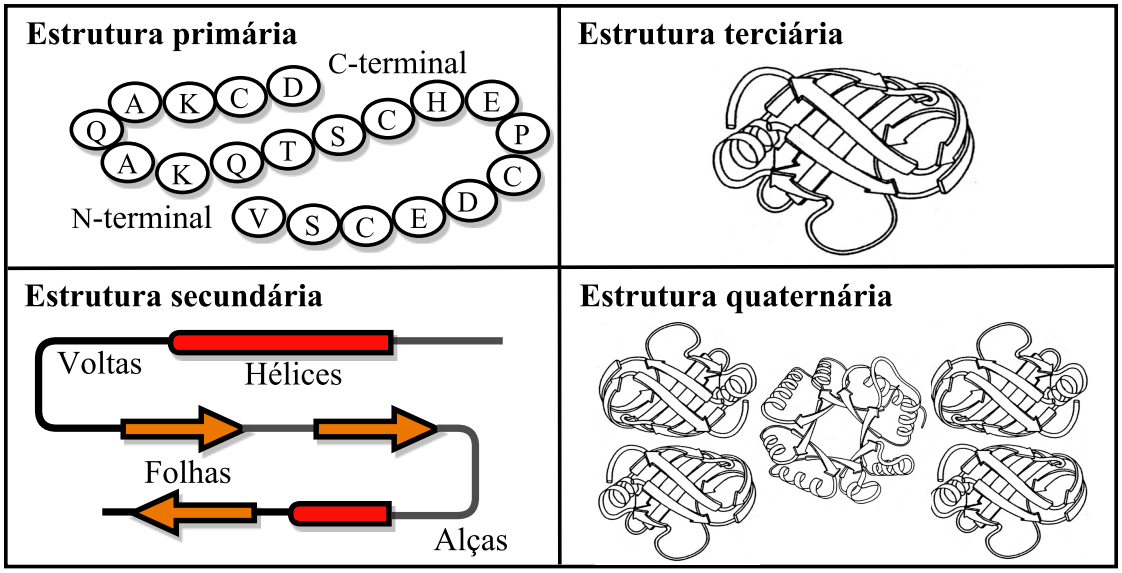
2.2 Níveis de abstração estrutural das proteínas

As proteínas podem ser divididas em quatro níveis de abstração estrutural, com o intuito de facilitar a descrição e a compreensão estrutural acerca das mesmas (BRANDEN; TOOZE, 1999; LODISH et al., 2007): (*i*) estrutura primária (EP); (*ii*) estrutura secundária (ES); (*iii*) estrutura terciária (ET); e (*iv*) estrutura quaternária. No entanto, esta hierarquia não visa descrever precisamente as leis da física que regem a formação estrutural das proteínas, mas tornar o estudo acerca das conformações adotadas um pouco mais claro (SCHEEF; FINK, 2009). A Figura 2.2 ilustra os quatro níveis de abstração estrutural das proteínas.

A EP descreve somente a sequência de resíduos de aminoácidos em uma ordem linear (BRANDEN; TOOZE, 1999). Cada aa se liga a outro através de uma ligação peptídica, onde o início da EP corresponde à região N-terminal da proteína, e a extremidade é denominada região C-terminal.

A ES é definida por arranjos estáveis de aminoácidos, determinados, principalmente, pela presença de padrões de ligações de hidrogênio, formados por meio das inte-

Figura 2.2: Representação dos quatro níveis de abstração estrutural das proteínas



Fonte: Do autor (2017).

rações entre átomos de H e átomos de O ou N constituintes das cadeias da proteína (LESK, 2013). Devido à elevada intensidade de força presente em interações moleculares desta natureza, elas são as responsáveis por garantir a estabilidade conformacional das estruturas secundárias (ESs) dispostas no espaço 3-D. As ESs podem ser classificadas em hélices (PAULING; COREY; BRANSON, 1951), folhas β (PAULING; COREY, 1951), voltas (*coils*) e alças (*turns*). As estruturas dos tipos hélices e folhas são conformações mais estáveis e por isto, chamadas de estruturas regulares. As hélices são estabilizadas por meio de ligações de hidrogênio entre o átomo de N de uma ligação peptídica ou o átomo de O do grupo carboxila do terceiro (3^{10}-hélice), quarto (α-hélice) ou quinto (π-hélice) aa da região N-terminal. Folhas β consistem em cadeias de aminoácidos estendidas combinadas à outras cadeias vizinhas que se estendem lado a lado, de forma paralela ou antiparalela. Os grupos amina e carboxila das ligações peptídicas se aproximam em um mesmo plano, de modo a permitir que ligações de hidrogênio ocorram entre cadeias polipeptídicas adjacentes. Voltas e alças representam conformações estruturais menos estáveis (irregulares). Originam-se em regiões onde a proteína altera a sua conformação, ou seja, acontecem antes ou após uma ES regular (LESK, 2013). Estas regiões são conhecidas como regiões irregulares das estruturas, visto que apresentam alto grau de flexibilidade devido à maior exposição ao solvente. Por este motivo são consideradas conformações de mais difícil predição.

Na sequência da hierarquia, a ET de uma proteína está relacionada à sua topologia

3-D, que por sua vez, é definida pela forma como as ESs estão arranjadas e conectadas, além do posicionamento assumido por elas no espaço 3-D, sendo que este enovelamento também é chamado de estrutura nativa ou funcional da proteína. A conformação é constituída e estabilizada através da variação de fatores termodinâmicos, como interações atômicas intramoleculares, ligações de hidrogênio, interações hidrofóbicas e eletrostáticas, forças de *van der* Waals de atração e repulsão, entre outros (GIBAS; JAMBECK, 2001; RICHARDSON, 1981).

No nível mais alto da abstração estrutural, encontra-se a estrutura quaternária de proteínas. Existem alguns polipeptídeos que apresentam mais de uma cadeia polipeptídica, onde cada grupo estrutural (estrutura 3-D) detêm características próprias quanto à EP, ES e ET. Estas diferentes cadeias presentes na proteína são chamadas de subunidades. Com isso, a estrutura quaternária de uma proteína consiste na combinação e disposição 3-D de duas ou mais cadeias polipeptídicas (ETs) que formam a proteína. Esta estrutura é mantida pelas mesmas forças que determinam as ESs e ETs (LODISH et al., 2007).

O entendimento sobre o enovelamento das proteínas, permite que a investigação de processos biológicos seja realizada de maneira mais direta, com resoluções e detalhamentos maiores, apesar da complexidade envolvida neste processo. Segundo o paradigma "sequência-proteína-estrutura", determinadas proteínas apenas assumem as suas funções biológicas quando enoveladas em uma única e estável conformação (ANFINSEN, 1973). Contudo, sabe-se que nem todas as funções desempenhadas pelas proteínas estão diretamente associadas a um estado nativo e estável (TOMPA, 2002; DUNKER et al., 2008b). Em alguns caso, as proteínas devem manter-se desordenadas para conseguir desempenhar as suas funções de maneira correta (GUNASEKARAN et al., 2003). Estas proteínas são chamadas de proteínas intrinsecamente desordenadas (IDP, sigla em inglês) (DUNKER et al., 2001) e englobam cerca de 30% das sequências de proteínas conhecidas. Apesar da existência de IDPs, um aspecto importante para explicar a função de uma determinada proteína, envolve a análise de interações moleculares complexas presentes na sua conformação. Estas interações podem ser intramoleculares (ligações iônicas, covalentes, metálicas) ou intermoleculares (ligações de hidrogênio e outras ligações não covalentes, como forças eletrostáticas e de *van der* Waals). Assim, conforme já referido, o conhecimento da estrutura 3-D de polipeptídeos fornece aos pesquisadores informações importantes a respeito da função desempenhada pela proteína no organismo (BRANDEN; TOOZE, 1999; LASKOWSKI; WATSON; THORNTON, 2005).

2.3 Classes estruturais de proteínas

Atualmente, sabe-se que a maior parte das proteínas apresentam semelhanças estruturais com outras proteínas, sendo que em muitos casos, estas também compartilham das mesmas origens evolutivas (FOX; BRENNER; CHANDONIA, 2015). Com isso, a classificação de proteínas em diferentes classes estruturais, considerando os arranjos de ESs e as conformações adotadas por estas, é capaz de prover descrições detalhadas acerca das relações entre as proteínas com estrutura 3-D conhecida. Estas classes configuram variadas interações intermoleculares, as quais originam diferentes arranjos estruturais de ESs e topologias 3-D.

O banco de dados *Structural Classification of Proteins*[1] (SCOP) (MURZIN et al., 1995; CONTE et al., 2000) consiste em uma iniciativa voltada à classificação de conformações estruturais de proteínas, considerando as similaridades de estruturas e de sequências de aminoácidos. O SCOP foi desenvolvido com o propósito de fornecer detalhadas e compreensivas categorizações relativas a estruturas de proteínas existentes no PDB. Estes recursos provêm amplas revisões acerca das diferentes conformações de proteínas, além de informações sobre proteínas similares estruturalmente à qualquer outra proteína classificada (FOX; BRENNER; CHANDONIA, 2015). De acordo com o SCOP, as proteínas podem ser classificadas em diferentes níveis hierárquicos, sendo que o nível mais alto da hierarquia compõe a divisão por classes, visando categorizá-las de acordo com os diferentes tipos de ESs e arranjamentos apresentados. Os outros níveis de agrupamento representam categorias mais específicas, a partir da classificação em classes.

Dessa forma, os diferentes níveis possíveis de classificação considerados pelo SCOP consistem (MURZIN et al., 1995; CONTE et al., 2000):

1. Classes: O agrupamento de proteínas em classes representa o nível mais alto da hierarquia de classificação. Neste nível, as proteínas são agrupadas de acordo com as principais conformações existentes, considerando os tipos de ESs assumidos. Cada proteína pode ser classificada em uma das classes abaixo:
    - Classe de hélices (*All-*α): Engloba estruturas essencialmente formadas de α-hélices;
    - Classe de folhas (*All-*β): Compreende estruturas essencialmente formadas de folhas β;

[1] <scop.mrc-lmb.cam.ac.uk/scop/>

- Classe híbrida-1 (α/β): Compreende estruturas formadas por α-hélices e folhas β;
- Classe híbrida-2 (α+β): Engloba estruturas formadas por α-hélices e folhas β, porém estas apresentam-se bem segregadas umas das outras, ou seja, existe a definição clara da região composta por α-hélices e da região formada de folhas β;
- Multi conformações (*multi-domain*): Esta última classe compreende proteínas que apresentam conformações que se encaixam em mais de uma classe apresentada acima.

2. Tipos de conformações: As classes descritas acima englobam estruturas que apresentam diferentes tipos de conformações, com a mesma composição de ES. Neste nível, as proteínas pertencentes a cada classe são subdivididas de acordo com as diferentes conformações e interações topológicas assumidas no espaço 3-D;
3. Super famílias: Dentre as proteínas que apresentam conformações similares, estas podem ser categorizadas através do indício de que possuem origens evolutivas em comum. Cada super família compreende proteínas que apresentam baixa similaridade de sequência de aminoácidos, mas, em muitos casos, possuem propriedades funcionais as quais sugerem que elas procedem da mesma origem evolutiva.
4. Famílias: Este nível representa a categorização mais específica para as proteínas. Estas são agrupadas em famílias, a partir das super famílias, seguindo um de dois critérios que visam estabelecer as origens evolutivas que elas apresentam em comum. O primeiro critério classifica em famílias as proteínas que apresentam similaridade de sequência significativa, e o segundo compõe grupos onde as proteínas apresentam estruturas 3-D e funções extremamente similares.

Sendo assim, a partir da descrição apresentada pelo projeto de classificação de proteínas SCOP, definiu-se para este trabalho, o próprio esquema de agrupamento de proteínas em classes. Esta divisão tem por objeto obter informações mais específicas acerca das proteínas que serão abordadas ao longo da dissertação, sendo que estas foram classificadas de acordo com a composição e arranjos de ESs relativos aos resíduos de aminoácidos das proteínas (CHOU; ZHANG, 1995). Observa-se que será adotada uma divisão de estruturas de proteínas mais simplificada do que a proposta pelo SCOP, a qual tende ser mais generalista quanto à classificação das proteínas, porém mais fácil de ser executada. Dessa forma, as proteínas que serão abordadas foram classificadas em quatro

classes distintas, que compreendem:

- Classe de regiões irregulares: Compreende estruturas que apresentem mais de 80% de voltas ou alças na constituição da ES;
- Classe de folhas: Engloba proteínas que apresentem o predomínio de mais de 60% de folhas β em suas ESs;
- Classe de hélices: Abrange estruturas de proteínas que possuem mais 60% de hélices na constituição da ES;
- Classe híbrida: Compreende estruturas que não se encaixam em nenhuma das classes anteriores, ou seja, apresentam uma combinação dos três tipos de ESs na constituição da mesma.

Destaca-se que nesta classificação, foi adicionada uma classe de regiões irregulares, visando englobar proteínas compostas, em sua maioria, de voltas ou alças, conforme indicado no trabalho de Chou e Zhang (1995). A partir das classes híbrida-1 e híbrida-2 apresentadas pelo SCOP, definiu-se uma única classe híbrida, sendo que esta engloba as duas classes anteriores. Os valores de predominância de ES definidos, seguem as linhas de classificação delineadas pelo SCOP e no trabalho de Chou e Zhang (1995).

2.4 Modelos de representação computacional da estrutura de proteínas

Proteínas podem assumir diversas conformações em um espaço 3-D, sendo que a estrutura de uma determinada proteína, definida pela sequência de aminoácidos, se enovela espontaneamente na natureza durante ou após a biossíntese. A relação entre a sequência de aminoácidos de uma proteína e a sua estrutura 3-D, foi demonstrada, pela primeira vez, através de experimentos realizados por Anfinsen et al. (1961, 1973), caracterizando-se como um processo complexo e dependente de muitos fatores, tais como elementos de solvatação, concentração de sais, temperatura, entre outros. Sendo assim, a representação computacional de estruturas 3-D de polipeptídeos é uma tarefa desafiadora, devido à dificuldade em representar a estrutura atômica e as interações moleculares de uma proteína, assim como simular todos os fatores responsáveis à sua estabilização, conforme ocorre na natureza ou até mesmo na determinação estrutural realizada por métodos experimentais.

A configuração computacional de um modelo de proteína está associado ao nível de detalhamento utilizado para descrever a sua estrutura 3-D, visto que quanto maior o nível de detalhes empregado no modelo, maior será a capacidade de descrever a proteína

tal como na natureza, e consequentemente maior será a complexidade computacional envolvida (MIRNY; SHAKHNOVICH, 2001; CORRÊA; DORN, 2017). Representações mais detalhadas visam englobar todos os átomos da proteína (*all-atom*) e ainda descrever as moléculas de solvente necessárias ao processo de enovelamento, enquanto outras, mais simplificadas, buscam abstrair alguns conceitos, na tentativa de reduzir a complexidade computacional. Contudo, quanto mais simplificada for a representação computacional utilizada, maior será a perda de informações e menor será a representatividade da mesma quando comparada a proteínas reais (MIRNY; SHAKHNOVICH, 2001).

A representação geométrica de estruturas é um dos componentes mais importantes utilizados em métodos computacionais de predição de estruturas 3-D de proteínas, sendo responsável pela redução ou aumento da complexidade do espaço de busca conformacional. Com isso, a utilização de modelos demasiadamente detalhados, que buscam representar todos os átomos constituintes da proteína e demais moléculas envolvidas no processo de enovelamento, acaba fazendo com que a representação torne-se computacionalmente custosa, como visto em simulações de dinâmica molecular, onde todos os átomos da proteína e moléculas auxiliares são considerados na simulação computacional. Assim, representações mais simplificadas são frequentemente preferíveis quando aplicadas ao problema PSP (CHIVIAN et al., 2003; CORRÊA; DORN, 2017). Comumente, duas representações computacionais destacam-se na literatura. O primeiro modelo representa a estrutura 3-D da proteína através da posição Cartesiana (x, y, z) dos átomos de cada aa. Neste caso, uma cadeia polipeptídica pode ser descrita como um conjunto A de átomos (a) dispostos no espaço 3-D, onde $\{a|a \in \mathbb{R}^3\}$. O segundo modelo representa a estrutura polipeptídica através do conjunto de ângulos de torção (diedros), constituídos pelo encadeamento de aminoácidos e ligações peptídicas. Esta representação baseia-se no fato de que os comprimentos de ligação apresentam-se quase sempre constantes em uma cadeia polipeptídica, o que possibilita a reconstrução do modelo de modo a representar todos os átomos da proteína (NEUMAIER, 1997). A principal vantagem do uso de ângulos de torção, quando comparado ao modelo Cartesiano, é a capacidade de reduzir o grau de liberdade do modelo estrutural, reduzindo conformações e complexidade. No entanto, a principal desvantagem desta representação é que uma pequena alteração em um ângulo diedro pode causar alterações drásticas no restante da estrutura 3-D da proteína, sendo que em representações Cartesianas, alterações pontuais em átomos pouco afetam a estrutura global do modelo. Também existem outros modelos de representação mais simplificados ainda, como os modelos reticulados (*lattice models*) (KOLINSKI; SKOLNICK, 2004)

e o modelo *off-lattice AB toy* (STILLINGER; HEAD-GORDON; HIRSHFELD, 1993). Nota-se que neste trabalho, a representação computacional a ser utilizada nos processos de otimização será baseada nos ângulos de torção da cadeias polipeptídicas das proteínas, de modo a reduzir a complexidade da representação *all-atom*, mantendo um certo grau de precisão dos modelos quando comparados a estruturas reais de proteínas.

2.5 Definições do problema

Esta seção tem por objetivo apresentar a estereoquímica presente entre as ligações de aminoácidos de uma proteína, bem como descrever o conjunto de ângulos diedros formados através destas. A partir desta coleção de ângulos diedros, pretende-se modelar computacionalmente o problema PSP.

2.5.1 Estereoquímica

As características específicas das ligações peptídicas entre aminoácidos geram implicações significativas à conformação adotada pelas proteínas. A ligação peptídica ($\mathtt{C}-\mathtt{N}$), responsável pela formação do ângulo diedro Omega (ω), possui uma ligação parcial dupla e tende a ser planar, com dois estados permitidos: *trans*, $\omega = 180^\circ$ (usualmente) e *cis*, $\omega = 0^\circ$ (raramente), apresentando pouca ou nenhuma rotação da molécula em torno desta ligação (BRANDEN; TOOZE, 1999). Rotações livres são permitidas em torno das ligações $\mathtt{N}-\mathtt{C}_\alpha$ e $\mathtt{C}_\alpha-\mathtt{C}$ (LODISH et al., 2007). Estas ligações representam os ângulos diedros Phi (ϕ) e Psi (ψ), respectivamente. Estes ângulos podem ser rotacionados livremente entre o intervalo de -180° a +180°, e são considerados os principais responsáveis pela conformação assumida pelo esqueleto peptídico da proteína. No entanto, a liberdade de rotação em torno dos ângulos ϕ e ψ é limitada por impedimentos estereoquímicos entre as cadeias principal e lateral dos resíduos de aminoácidos do polipeptídeo (BRANDEN; TOOZE, 1999; SCHEEF; FINK, 2009). Como consequência, a conformação de uma determinada proteína torna-se dependente das propriedades químicas dos aminoácidos.

Semelhante a cadeia principal da proteína, a cadeia lateral também possui ângulos diedros, denominados ângulos Chi (χ). A conformação da cadeia lateral dos aminoácidos contribui para a estabilização e empacotamento da proteína (LEVITT et al., 1997). O número de ângulos χ da cadeia lateral depende do tipo de aa, variando de 0 a 4 ângulos

com liberdades de rotação de -180° a +180°.

Dessa forma, o conjunto de ângulos ϕ, ψ e ω de todos os resíduos constituintes de uma proteína define a conformação do esqueleto peptídico, sendo que a combinação de ângulos χ de cada aa configura a cadeia lateral da proteína (HOVMÖLLER; ZHOU; OHLSON, 2002).

2.5.2 Modelagem computacional do problema

De acordo com a estereoquímica de aminoácidos, sabe-se que a partir do conjunto composto por todos os ângulos diedros que definem a conformação da proteína, é possível reconstruir o modelo computacional de forma a representar todos os átomos constituintes da estrutura, devido ao fato de que os comprimentos de ligação apresentam-se quase sempre constantes nas cadeias polipeptídicas. Com isso, a representação computacional através dos ângulos de torção da estrutura torna-se viável.

Assim, a estrutura 3-D de uma proteína P com n resíduos de aminoácidos, pode ser definida computacionalmente somente através da atribuição dos ângulos diedros da cadeia principal e lateral aos resíduos que englobam a proteína (Eq. 2.1), visto que os comprimentos de ligação entre os átomos são pouco variáveis.

$$P = (aa_1, aa_2, ..., aa_{n-1}, aa_n) \tag{2.1}$$

$$aa_i = (\phi_i, \psi_i, \omega_i, \chi_{i_{(0...4)}}) \tag{2.2}$$

Para a representação da cadeia principal de P, o modelo com n resíduos de aminoácidos possui $3 \times n$ graus de liberdade ou variáveis a serem otimizadas, com a ressalva de que ângulos ω possuem pouca variação. Acrescentando a isto os ângulos diedros da cadeia lateral, obtêm-se a cardinalidade do conjunto de ângulos de P (dimensionalidade de variáveis) através da Equação 2.3. Neste cálculo, desconsidera-se o ângulo ϕ da região N-terminal do início da cadeia principal e o ângulo ψ da região C-terminal da extremidade da cadeia principal do polipeptídeo, pois são inexistentes.

$$|P| = n \times 3 \times (\sum_{1}^{n} |\chi_i|) - 2 \tag{2.3}$$

Com isso, o problema PSP pode ser descrito matematicamente como um problema de otimização (LEUNG; WANG, 2001). Considerando $f(x)$ como função objetivo utili-

zada na avaliação de soluções, tal que $f(x)$ deve ser minimizada em relação ao intervalo de números reais definido por $l \leq x \leq u$, onde $x = P = \{aa_1, aa_2, ..., aa_{n-1}, aa_n\}$ (Eq. 2.1) representa um vetor de variáveis no espaço $\mathbb{R}^{|P|}$, visto que cada variável de x é também um vetor de variáveis que engloba os valores angulares de um determinado aa (Eq. 2.2). l e u definem o espaço de soluções permitido à cada variável, neste caso $l = -180$ e $u = +180$, sendo o intervalo $[l, u]$ comum a todas as variáveis de x.

2.6 Função de avaliação

A predição de estruturas de proteínas envolve a geração de diversas soluções estruturais (modelos preditos), objetivando eleger o modelo mais próximo da estrutura nativa da proteína (FARAGGI; KLOCZKOWSKI, 2014). Com isso, funções de energia são utilizadas no processo de predição de estruturas 3-D de proteínas para estimar a condição de enovelamento de um determinado modelo, considerando a disposição de seu valor de energia sobre a superfície energética representada pela função de avaliação. São empregadas como funções de minimização durante o processo de otimização e devem ter a capacidade de diferenciar configurações de proteínas fisicamente mais ou menos estáveis, visto que, teoricamente, conformações em torno de um estado nativo de uma proteína devem refletir regiões de mínimos globais de sua energia livre (ANFINSEN, 1973), como referido anteriormente. Diz-se "teoricamente", pois sabe-se que as funções de energia utilizadas nos processos de modelagem molecular apresentam grandes dificuldades em representar as reais estruturas da proteína na natureza, devido à enorme complexidade envolvida nestes processos (KIM et al., 2009).

As funções de energia utilizadas no problema PSP, compreendem a categoria de funções de avaliação multimodais, caracterizadas pelo espaço de busca altamente rugoso, fato que origina diversos vales e picos vistos como ótimos locais e globais da função de avaliação (HANDL; LOVELL; KNOWLES, 2008). Uma mesma função objetivo multimodal pode apresentar, para o mesmo dado de entrada, duas ou mais soluções distintas localizadas em diferentes regiões do espaço de busca, porém com valor de custo similar (GLIBOVETS; GULAYEVA, 2013).

No que se refere à predição de estruturas de proteínas, um mesmo valor de energia pode representar diferentes conformações estruturais para a mesma proteína-alvo, impossibilitando a diferenciação de qualidade entre estruturas oriundas de dois mínimos locais. Além de que, frequentemente, pontos ótimos encontrados ao longo do processo tendem a

não refletir as conformações em torno do estado nativo do polipeptídeo (KIM et al., 2009). Ainda, ressalta-se a existência de IDPs, as quais são desprovidas de estados mais estáveis ou desordenadas sob determinadas condições fisiológicas (DUNKER et al., 2008a), onde o estado nativo e funcional não pode ser representado através de um único mínimo global.

Normalmente, as funções de energia voltadas à predição de estruturas 3-D de proteínas, representam funções potenciais derivadas de estruturas de proteínas previamente conhecidas (*knowledge-based energy function*), sendo empiricamente derivadas de estruturas determinadas experimentalmente a partir do PDB (HAO; SCHERAGAT, 1999; CHIVIAN et al., 2003; LAZARIDIS; KARPLUS, 2000). Algumas funções também incorporam termos da mecânica molecular, que visam modelar as forças responsáveis por determinar as conformações de proteínas através de formatos funcionais parametrizados fisicamente, considerando dados de pequenas moléculas ou baseados em cálculos da mecânica quântica (JORGENSEN; TIRADO-RIVES, 2005). Protótipos mais genéricos de funções de energia potencial podem ser expressos através do somatório linear, ponderado ou não, de alguns termos de energia representantes das forças que determinam as conformações macromoleculares (HANDL; LOVELL; KNOWLES, 2008; MACKERREL, 2010), conforme mostra a Equação 2.4.

$$E_f = E_l + E_{nl} \tag{2.4}$$

Onde, E_f representa a função de energia final, podendo ser divida em duas outras equações, que denotam os termos de energia concernentes aos átomos ligados (Eq. 2.5) e forças atuantes em átomos não ligados (Eq. 2.6).

$$E_l = E_{cl} + E_{al} + E_{atp} + E_{at} \tag{2.5}$$

E_l denota a combinação dos principais termos de energia ligados, englobando comprimentos de ligações, ângulos de ligações, ângulos de torção proibidos e valores de ângulos de torção, respectivamente.

$$E_{nl} = E_{vdw} + E_{elet} + E_{lh} + E_{solv} \tag{2.6}$$

E_{nl} representa a combinação dos principais termos de energia para forças não ligadas, compreendendo as forças de atração e repulsão de *van der* Waals, interações eletrostáticas, ligações de hidrogênio e componentes implícitos de solvatação, respectivamente.

Nesta dissertação, embora a otimização de soluções estruturais para as proteínas-alvo seja realizada por meio de alterações nos ângulos de torção que descrevem as cadeias polipeptídicas de uma proteína, no que concerne à avaliação de energia destes modelos, será utilizada a representação Cartesiana dos mesmos, através da função de energia *all-atom* do Rosetta[2] (ROHL et al., 2004; KAUFMANN et al., 2010).

2.6.1 Função de energia do Rosetta

A função de energia do Rosetta, implementada no PyRosetta[3], foi utilizada nesta dissertação como função objetivo responsável pela avaliação da qualidade das soluções estruturais preditas. PyRosetta (CHAUDHURY; LYSKOV; GRAY, 2010) consiste em um conjunto de bibliotecas, desenvolvido na linguagem de programação Python[4], derivado do kit de modelagem molecular Rosetta (ROHL et al., 2004; KAUFMANN et al., 2010). Portanto, é correto afirmar que a função de energia do PyRosetta equivale a função de energia implementada no Rosetta[5].

De acordo com Combos et al. (2013), a função de energia do Rosetta, tratando-se de uma função potencial baseada em conhecimento, foi desenvolvida por meio de uma análise empírica das estruturas de proteínas determinadas experimentalmente e contidas no PDB. O uso desta abordagem inclui informações biológicas, tais como raio de giro (RG), densidade de empacotamento, distância entre ligações de hidrogênio, interações de resíduos par a par, entre outras. Estas informações são convertidas em valores de energia através do uso de estatísticas *Bayesianas*. Ressalta-se que conforme as últimas avaliações do CASP, os algoritmos que utilizaram funções potenciais fornecidas pelo Rosetta, figuraram entre os melhores métodos avaliados nos experimentos (HUANG et al., 2014; TAI et al., 2014).

O Rosetta fornece duas representações distintas para a descrição de átomos: modelo de centroide (*low-resolution*) e modelo *all-atom* (LEAVER-FAY et al., 2011). A diferença entre ambas está na representação da cadeia lateral, sendo que o primeiro modelo representa uma descrição reduzida, através de uma abordagem de mais alto nível, da estrutura 3-D da proteína, onde cada cadeia lateral de aa é representada por um centroide localizado no centro de massa da cadeia. A segunda opção proporciona um detalhamento

[2] <www.rosettacommons.org>
[3] <www.pyrosetta.org>
[4] <www.python.org>
[5] <www.goo.gl/z5xJ7V>

atômico maior, onde todos os átomos da cadeia lateral, incluindo átomos de hidrogênio, são representados (ROHL et al., 2004). A função de energia *all-atom* apresenta resoluções mais precisas quando da comparação com estruturas reais de proteínas e descrições mais detalhadas acerca das conformações modeladas (ROHL et al., 2004). Nota-se que neste trabalho será utilizado o modelo de representação *all-atom* para avaliações de energia.

Comumente, uma função de energia potencial, conforme referido anteriormente, incorpora duas categorias de termos (HANDL; LOVELL; KNOWLES, 2008; MACKERREL, 2010): ligados e não-ligados. A função do Rosetta considera mais de 18 termos de energia, sendo que grande parte são derivados de potenciais baseados em conhecimento. A função conta com termos fundamentados na física *newtoniana*, como o potencial *6-12 de Lennard-Jones*, dividido em termos de atração e repulsão necessários à descrição das interações de *van der* Waals (KUHLMAN; BAKER, 2000), e a energia de solvatação de Lazaridis-Karplus (LAZARIDIS; KARPLUS, 2000). O Rosetta também combina termos de interações eletrostáticas interatômicas baseadas em conhecimento obtidas através de potenciais entre pares de aminoácidos, e energias de ligações de hidrogênio dependentes da orientação (KORTEMME; MOROZOV; BAKER, 2003). Adicionalmente à estes termos, a função do Rosetta ainda emprega termos que buscam estimar a energia livre de aminoácidos dependentes da conformação. Estes termos objetivam avaliar o posicionamento da cadeia lateral de cada aa, conforme a biblioteca de rotações do Dunbrack (*Dunbrack rotamer library*) (DUNBRACK; COHEN, 1997). Além de verificar a preferência conformacional dos ângulos ϕ e ψ através de um gráfico de Ramachandran para os aminoácidos que compõe a proteína. Os principais termos utilizados na função de energia do Rosetta estão resumidos na Tabela 2.1.

Tabela 2.1: Termos da função de energia *all-atom* do Rosetta

Componente	**Referência**
Interações 6-12 de Lennard-Jones	(KUHLMAN; BAKER, 2000)
Energia de solvatação de Lazaridis-Karplus	(LAZARIDIS; KARPLUS, 2000)
Interações eletrostáticas interatômicas	(KUHLMAN; BAKER, 2000)
Potencial de ligações de hidrogênio	(KORTEMME; MOROZOV; BAKER, 2003)
Potencial de ligações dissulfeto	(SEVIER; KAISER, 2002)
Preferências conformacionais das cadeias laterais	(DUNBRACK; COHEN, 1997)
Preferências conformacionais dos aminoácidos	(ROHL et al., 2004)
Potencial derivado do fechamento do anel da prolina	(THOMASSON; APPLEQUIST, 1990)

Fonte: Adaptado de Corrêa et al. (2016).

O valor final de energia da função potencial do PyRosetta ($E_{pyrosetta}$), é dado pela soma de todas as ponderações realizadas sobre os termos de energia considerados no

cálculo. O peso para cada termo foi atribuído com base na função de energia *Talaris2014*, sendo atualmente a função padrão do Rosetta utilizada para avaliar modelos estruturais *all-atom*. Esta função também foi utilizada em Leaver-Fay et al. (2013) e O'Meara et al. (2015).

2.6.2 Função de avaliação final

Além dos termos da função de energia do Rosetta apresentados na seção anterior, foi incluído na função de energia final o termo de avaliação da área total da superfície acessível ao solvente (SASA, sigla em inglês) (CONNOLLY, 1983; RICHMOND, 1984) com raio atômico de 1,4Å, também fornecido pelo PyRosetta, com o intuito de ajudar no empacotamento das estruturas 3-D, visto as dificuldades encontradas na *Talaris2014*, no que se refere a esta tarefa. O SASA visa medir o grau de exposição ao solvente de uma determinada estrutura de proteína, estimando a energia livre das interações entre soluto e solvente. Com isso, conformações de proteínas menos estáveis tendem a apresentar valores de SASA mais altos, pois existem mais resíduos em contato com o solvente, sendo que valores de SASA mais baixos indicam que menos aminoácidos da proteína estão expostos ao solvente, fazendo com que ela mostre-se mais compacta (ROSE et al., 1985). Acredita-se que este fator agregado à função de energia, tende a guiar o processo a conformações mais empacotadas.

Ainda, objetivando favorecer a formação correta de ES foi incorporado à função de avaliação final um termo de estrutura secundária (Eq. 2.7). O procedimento consiste em dar um reforço positivo, adicionando uma constante negativa $(-const)$ ao somatório de todos os aminoácidos da proteína P, quando a ES (zp_i) correspondente ao i-ésimo aminoácido (aa_i) for igual à ES (ze_i), equivalente ao mesmo resíduo, fornecida como entrada para o algoritmo. Por outro lado, a técnica dá um reforço negativo ao somatório, adicionando uma constante positiva $(+const)$, quando a ES dos resíduos de aminoácidos correspondentes não forem iguais. Todos os aminoácidos da proteína são comparados durante a avaliação da solução. O algoritmo DSSP (KABSCH; SANDER, 1983) foi utilizado para atribuir as ESs ao longo da simulação.

$$ES_{termo} = \sum_{aa \in P}^{i+1} V_{aa,zp,ze}(aa_i, zp_i, ze_i) \tag{2.7}$$

$$V_{aa,zp,ze}(aa, zp, ze) = \begin{cases} \textit{-const}, & zp = ze \\ \textit{+const}, & zp \neq ze \end{cases} \tag{2.8}$$

Por fim, os termos descritos acima são então adicionados ao resultado da função de energia do PyRosetta, constituindo a função de avaliação final (E_{final}) adotada neste trabalho (Eq. 2.9). Esta função de avaliação foi previamente proposta no trabalho de Corrêa et al. (2016), e será referida neste trabalho como função de energia composta.

$$E_{final} = E_{pyrosetta} + SASA_{termo} + ES_{termo} \tag{2.9}$$

2.7 Resumo do capítulo

Este capítulo apresentou fundamentações biológicas necessárias ao entendimento do problema PSP, bem como de questões relativas a este.

Os principais conceitos abordados foram: (*i*) composição química das proteínas; (*ii*) níveis de abstração estrutural; (*iii*) classes estruturais das proteínas, onde foi descrita a forma como as proteínas foram classificadas neste trabalho; (*iv*) modelos computacionais utilizados na representação de estruturas de proteínas, onde foi definido que neste trabalho a representação computacional de estruturas será baseada no conjunto de ângulos diedros dos aminoácidos; (*v*) definições do problema quanto à estereoquímica das proteínas, modelagem em termos computacionais, e formalização matemática do PSP como um problema de otimização; e (*vi*) funções de energia empregadas no problema, onde definiu-se a função de avaliação final que será adotada nesta dissertação.

O próximo capítulo apresentará uma visão geral dos trabalhos relacionados à área de otimização e meta-heurísticas aplicadas ao problema, bem como objetiva descrever os métodos considerados estado da arte na área de predição de estruturas de proteínas.

3 TRABALHOS RELACIONADOS

Segundo Kryshtafovych et al. (2014), a modelagem de estruturas de proteínas e a biologia estrutural experimental complementam-se na criação e disponibilização de informações estruturais aos pesquisadores, endereçando importantes problemas relacionados às ciências biológicas. Especialmente, devido aos avanços presenciados nas últimas décadas, a modelagem baseada em conhecimento consolidou-se como um campo de pesquisa maduro, tornando-se hoje um dos principais recursos relacionados à pesquisa neste contexto. No entanto, apesar dos progressos significativos realizados, o problema PSP ainda está longe de ser solucionado, fazendo-se necessário mais investigações no sentido de melhorar o processo de predição, especialmente na linha de modelagem *ab-initio*. Tais direcionamentos foram delineados a partir de análises de resultados oriundos de experimentos do CASP, cujos objetivos são determinar o estado da arte na predição de estruturas de proteínas e apontar os avanços mais notáveis já realizados (MOULT et al., 2016).

3.1 CASP e métodos estado da arte

O CASP consiste da realização de uma série de experimentos, abertos à comunidade científica, que visam avaliar as predições de estruturas de proteínas realizadas a partir da sequência de resíduos de aminoácidos. Os experimentos de predição conduzidos pelo CASP são organizados sob a responsabilidade do Centro de Predição de Estruturas de Proteínas[1] (*Protein Structure Prediction Center*).

O principal objetivo do CASP se resume em auxiliar no progresso de métodos de predição de estruturas de proteínas. O Centro de Predição foi criado com o intuito de fornecer meios para avaliar objetivamente estes métodos através do processo de predição cega (*blind prediction*), onde os participantes desconhecem a estrutura nativa das proteínas-alvo. Os experimentos do CASP tem como objetivo estabelecer o estado da arte na área de predição de estruturas de proteínas, identificando avanços alcançados e direcionando para pontos críticos que futuramente podem ser atacados de forma mais produtiva (KINCH et al., 2016a).

O CASP acontece a cada 2 anos, desde 1994, onde os experimentos cobrem um período aproximado de 9 meses. Primeiramente, sequências de proteínas que estão prestes a serem determinadas experimentalmente (cristalografia de raios-X ou RMN) são solicita-

[1] <www.predictioncenter.org>

das à comunidade de predição experimental. Estas sequências são então disponibilizadas aos membros da comunidade de modelagem computacional de estruturas para que sejam modeladas estruturalmente. Os modelos preditos devem ser submetidos à avaliação antes da liberação dos dados experimentais. A avaliação de resultados é realizada através de uma bateria de testes automatizados e por avaliadores independentes vinculados ao Centro de Predição (MOULT et al., 2014). Ao final dos experimentos, acontece uma conferência onde todas as análises de resultados são divulgadas e discutidas com a comunidade.

Atualmente, as proteínas-alvo são divididas conforme duas categorias principais de dificuldade: (*i*) modelagem baseada em modelos (*template-based modeling*) e (*ii*) categoria FM. A primeira categoria engloba sequências de aminoácidos que possuam similaridades evolutivas detectáveis às estruturas experimentais, tornando possível a identificação de um ou mais modelos estruturais e facilitando o processo de predição, visto que muitas vezes estas estruturas são modeladas relativamente bem a partir dos dados experimentais empregados. O segundo grupo de alvos representa sequências de aminoácidos que não apresentam similaridades estruturais com proteínas determinadas experimentalmente. A dificuldade neste caso está na modelagem de alvos através de métodos *ab-initio* que podem ou não incorporar conhecimento de estruturas experimentais, conforme referido anteriormente.

Com a publicação de artigos científicos, o CASP busca descrever como se dá a estruturação e condução dos experimentos, medidas de avaliação utilizadas, relatórios técnicos desenvolvidos pelas equipes de avaliadores destacando o estado da arte em diferentes categorias de predição, descrição dos métodos com melhores resultados e o progresso em vários aspectos do processo de predição.

Sendo assim, de acordo com os resultados relacionados às últimas edições (TAI et al., 2014; KINCH et al., 2016a), no que se refere a técnicas automáticas de predição FM, sem intervenção manual (*servers*), os métodos QUARK (XU; ZHANG, 2012) e BAKER-ROSETTASERVER (KIM; CHIVIAN; BAKER, 2004) podem ser apontados como sendo os métodos mais relevantes para a área, devido à sequência de melhores resultados alcançados.

3.1.1 Método Rosetta

BAKER-ROSETTASERVER[2] (KIM; CHIVIAN; BAKER, 2004) representa um importante servidor web responsável por executar os protocolos de predição de estruturas de proteínas do Rosetta[3] (ROHL et al., 2004; BRADLEY; MISURA; BAKER, 2005), tanto para predições de estruturas *ab initio*, quanto para modelagens comparativas (SONG et al., 2013).

O Rosetta é um método baseado em fragmentos (*fragment assembly*) (SIMONS et al., 1997), que utiliza pequenos segmentos estruturais (3 e 9 resíduos de aminoácidos) de proteínas conhecidas, extraídos do banco de dados PDB. As configurações iniciais da proteína-alvo são criadas a partir da combinação de diferentes fragmentos. Este método é dividido em múltiplos estágios de otimização, onde diferentes representações estruturais e funções de avaliação são empregadas (LEAVER-FAY et al., 2011). O método parte de uma otimização de baixa precisão estrutural (*low-resolution*), e aumenta gradativamente o nível de precisão até finalizar o processo com uma técnica mais precisa de refinamento *all-atom*, onde as melhores estruturas encontradas topologicamente distintas são consideradas. Nota-se que a complexidade das funções de energia acompanha o aumento do nível de precisão da representação computacional dos modelos estruturais durante o processo.

Através de técnicas de agrupamento e amostragens de milhares de indivíduos, o método busca localizar diferentes conformações distribuídas sobre a superfície do espaço de busca conformacional. Os diferentes grupos estruturais são otimizados por diversas execuções de simulações de Monte Carlo, conhecidas como *Replica Exchange Monte Carlo* (REMC), através de processos de troca de parâmetros das simulações e fragmentos estruturais.

3.1.2 Método QUARK

O servidor web QUARK[4] (XU; ZHANG, 2012) implementa um algoritmo de primeiros princípios baseado em conhecimento, desenvolvido para predizer as estruturas 3-D de proteínas a partir das sequências de aminoácidos. Este método também baseia-se em fragmentos de aminoácidos (SIMONS et al., 1997), o qual utiliza pequenos segmentos de

[2] <www.robetta.bakerlab.org>
[3] <www.rosettacommons.org>
[4] <www.zhanglab.ccmb.med.umich.edu/QUARK/>

tamanhos variados de estruturas de proteínas conhecidas, compreendendo desde 1 até 20 resíduos de aminoácidos, para construir as soluções iniciais da proteína-alvo. Os modelos computacionais são criados a partir do arranjo de fragmentos de diferentes tamanhos.

Diferentemente do Rosetta, o QUARK não é divido em estágios, mas em algumas etapas de execução. A otimização computacional das soluções é realizada por meio de simulações de REMC sob a orientação de uma função de energia *all-atom* baseada em conhecimento. Após um primeiro processo de otimização dos modelos iniciais, os diferentes grupos estruturais (*decoys*) resultantes da simulação são agrupados de acordo com as suas similaridades estruturais e, seguidamente, novas simulações de REMC são realizadas a partir dos centroides de cada grupo formado, principalmente, com o objetivo de remover choques estereoquímicos e refinar a topologia global dos centroides. Os modelos resultantes são agrupados novamente, onde as soluções que apresentam valores de energia mais baixos são selecionadas como resultados finais do processo de modelagem.

Por padrão, o QUARK retorna 10 modelos finais que representam o conjunto de possíveis soluções estruturais para determinada sequência de aminoácidos. Segundo os autores, nenhuma informação de modelos experimentais é utilizada no processo de predição, tornando o servidor adequado para proteínas que não possuem estruturas homólogas no PDB (categoria FM).

3.2 Meta-heurísticas

Dependendo da complexidade imposta pelo problema que está sendo estudado, este pode ser solucionado através de métodos exatos (determinísticos) ou de soluções aproximadas. Métodos exatos são capazes de atingir soluções ótimas, contudo, quando aplicados a problemas pertencentes à classe de complexidade computacional NP-difícil (COOK, 1983), algoritmos exatos apresentam tempo de execução não-polinomial, tornando-os inviáveis à aplicação. Por outro lado, métodos aproximados ou heurísticas conseguem obter soluções de boa qualidade (aproximadas da solução ideal), em tempo de execução aceitável, quando aplicados à problemas reais. No entanto, não garantem que a solução ótima global será encontrada (TALBI, 2009). Problemas reais caracterizam problemas que em sua grande maioria não são capazes de ser solucionados de maneira ótima, por nenhum método determinístico dentro de um tempo de execução razoável (BOUSSAÏD; LEPAGNOT; SIARRY, 2013).

Sabendo que muitos dos problemas de otimização existentes, independente do

domínio de aplicação, não podem ser resolvidos de maneira ótima e em tempo hábil devido à enorme complexidade com que os espaços de busca se apresentam, abordagens baseadas em meta-heurísticas estão sendo largamente utilizadas (DRÉO et al., 2006; BOUSSAÏD; LEPAGNOT; SIARRY, 2013; LUKE, 2013). Meta-heurísticas são técnicas que se caracterizam por serem praticamente independentes do problema, sendo definidas como heurísticas de alto nível que podem ser empregadas em uma vasta gama de problemas com nenhuma ou pequenas modificações em seus parâmetros. O prefixo grego "meta" é utilizado justamente para contrastar com heurísticas dependentes do domínio de aplicação (BOUSSAÏD; LEPAGNOT; SIARRY, 2013). Segundo Boussaïd et al. (2013), praticamente todas as meta-heurísticas compartilham de algumas características: (*i*) são inspiradas por fenômenos da natureza, baseados em princípios da biologia, física ou etologia; (*ii*) incorporam estruturas algorítmicas estocásticas, as quais exploram o conceito de aleatoriedade; (*iii*) não dependem do gradiente ou da matriz hessiana das funções objetivo; e (*iv*) apresentam diversos parâmetros que necessitam ser ajustados ao problema de aplicação.

Entretanto, especialmente em problemas de Bioinformática Estrutural, a simples aplicação de métodos canônicos, nem sempre é o suficiente para atingir bons resultados. Isto deve-se ao elevado número de variáveis (alta dimensionalidade) a serem otimizadas nestes problemas. Dessa forma, a incorporação de conhecimento prévio acerca do problema e a exploração de características específicas podem ser vistas como alternativas para aumentar a eficácia dos métodos, reduzindo a complexidade através da restrição do espaço de soluções. Exemplos comuns de meta-heurísticas utilizadas são os algoritmos evolutivos (AEs) (BACK, 1996), métodos baseados em inteligência de enxame (KENNEDY et al., 2001) e algoritmos que simulam processos físicos (DAS; CHAKRABARTI, 2005).

3.2.1 Meta-heurísticas multimodais

Sabe-se que inúmeros problemas das mais diversas áreas do conhecimento englobam complexas funções objetivo para determinar possíveis soluções para os mesmos (GLIBOVETS; GULAYEVA, 2013). As funções de energia utilizadas na otimização de estruturas 3-D de proteínas, por exemplo, encaixam-se na complexa categoria de funções de avaliação multimodais, conforme mencionado anteriormente. Otimizações multimodais procuram de algum modo contornar as dificuldades impostas pela multi-

modalidade das funções por meio de adaptações nos algoritmos de busca. O objetivo é encontrar variadas soluções ótimas ou subótimas e não apenas uma única solução para o problema (DAS et al., 2011). A descoberta de múltiplas soluções pode auxiliar no desempenho dos métodos de busca, visto que variados pontos do espaço são otimizados e facilmente podem ser alterados sem afetar a performance global do processo. A otimização multimodal também pode ser capaz de revelar propriedades e relações até então desconhecidas acerca da superfície da função objetivo.

Algoritmos de busca simples, voltados à otimização de uma única solução por execução, são tradicionalmente utilizados com o intuito de encontrar apenas um resultado ótimo da função de avaliação. Quando algoritmos deste tipo são utilizados em otimizações multimodais, é necessário que sejam aplicados repetidas vezes, esperando que a cada execução, uma solução diferente seja encontrada. Neste sentido, AEs apresentam vantagens sobre outras heurísticas de busca mais clássicas que não baseiam-se em populações de soluções. Idealmente, se determinado AE for capaz de manter a diversidade das soluções proveniente de uma boa exploração do espaço de busca, ao final da execução do algoritmo, é possível que se obtenha múltiplas soluções boas ao invés de apenas uma (DAS et al., 2011). Porém, em relação aos espaços de busca maiores, devido à deriva genética inerente à evolução das populações (BELDA et al., 2007), os AEs também tendem a convergir naturalmente para um único ótimo global, fazendo com que toda a população se restrinja a ele e despenda todos os esforços para otimizá-lo. Com isso, a descoberta e manutenção de múltiplas soluções ao longo da execução do algoritmo, configuram os principais desafios no uso de meta-heurísticas evolutivas aplicadas à otimização multimodal.

As estratégias mais comuns utilizadas em otimizações multimodais são fundamentadas na ideia de formação de nichos (*niching*) (MAHFOUD, 1995; GLIBOVETS; GULAYEVA, 2013), e se referem à tentativa de encontrar e preservar múltiplos nichos ou partes do espaço de busca em torno de múltiplas soluções, com o objetivo de prevenir a convergência a um único ponto ótimo. Diversos métodos de nichos foram propostos ao longo dos anos (QU; SUGANTHAN; DAS, 2013), sendo que uma das principais técnicas é a técnica de agrupamento de soluções (*crowding*) (THOMSEN, 2004), que visa limitar os recursos entre soluções similares (grupos) da população.

No que concerne à predição de estruturas 3-D de proteínas, como postulado anteriormente, um mesmo valor de energia pode representar diferentes conformações estruturais para a mesma proteína-alvo. Sabendo das dificuldades que as funções de energia

possuem no que tange a representação de ótimos globais como sendo as melhores soluções, é interessante que se descubra o maior número possível de conformações ótimas ou subótimas com o intuito de fornecer recursos suficientes para futuras análises de especialistas. Observa-se, por exemplo, nos métodos Rosetta e QUARK que o resultado final de um processo de predição não consiste apenas de um único modelo estrutural, mas de um conjunto de soluções energicamente favoráveis e muitas vezes topologicamente distintas resultante dos diversos processos de otimização e agrupamento realizados ao longo da simulação.

3.3 Meta-heurísticas aplicadas ao problema PSP

Em consequência da enorme complexidade apresentada por diversos problemas da Bioinformática Estrutural e da carência de métodos computacionais eficazes, diferentes tipos de meta-heurísticas, com destaque para os algoritmos genéticos (AGs) (HOLLAND, 1975; MITCHELL, 1998), vêm sendo aplicados na tentativa de obter soluções aceitáveis para estes, como pode ser observado nos trabalhos de Le Grand e Merz Jr (1993), Sun (1995), Jones et al. (1997), Krasnogor et al. (2002), Zhang et al. (2010), Maulik et al. (2011), Olson e Shehu (2013), e Li et al. (2015). AGs são métodos de busca baseados na abstração da teoria da evolução, proposta por Darwin, e na seleção natural de sistemas biológicos, os quais são representados através de operadores matemáticos, como cruzamento ou recombinação (*crossover*), mutação, aptidão (*fitness*) e seleção do indivíduo mais apto (YANG, 2010). Basicamente, estes algoritmos alteram a orientação da estrutura 3-D da macromolécula em estudo por meio de operações matemáticas, com o objetivo de minimizar uma função de energia visando aproximar os modelos computacionais (mínimos globais) da solução ótima (DESJARLAIS; CLARKE, 1998).

Conforme trabalhos relacionados à literatura da área de predição de estruturas de proteínas, diversas técnicas de otimização estão sendo propostas para lidar com o problema PSP (DORN et al., 2014). Por exemplo, Elofsson et al. (1995) propuseram um AG combinado a uma heurística responsável por realizar pequenos movimentos nos ângulos diedros da estrutura da proteína, com o objetivo de melhorar a exploração de mínimos locais. No trabalho de Dorn et al. (2011), foi desenvolvido um AG que utiliza uma população estruturada em castas aliado ao procedimento de busca de religação de caminhos (*Path Relinking*) (GLOVER, 1994). Em outro trabalho, Dorn et al. (2013) propuseram um método computacional baseado em conhecimento, cujo objetivo é reduzir o espaço de

busca utilizando um AG que procura considerar as ocorrências anteriores dos aminoácidos da proteína-alvo em proteínas experimentais já conhecidas. Devido à especificidade de cada função de energia, De Sancho e Rey (2008) combinaram duas funções de avaliação simplificadas (*low-resolution*) que representam diferentes tipos de interações entre moléculas para avaliar a energia global das proteínas-alvo. Os testes foram realizados por meio de um AE com o intuito de analisar a combinação das funções de avaliação e o potencial para serem empregadas na predição de estruturas de proteínas. No trabalho de Fonseca et al. (2010), uma variação do algoritmo de otimização de colônia de abelhas (*Bee Colony Optimization*) (PHAM et al., 2006; 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 et al. (2013) propuseram um AM composto por duas estratégias de busca evolutivas, baseadas em fragmentos estruturais de aminoácidos, para tratar o problema dos múltiplos mínimos locais presentes na função de energia. Para isso, os autores trabalharam a modelagem de duas funções de energia distintas, sendo uma versão modificada da função *Associative Memory Hamiltonian with Water* (AMW) (SHEHU; KAVRAKI; CLEMENTI, 2009) e a função de energia baseada em centroide do Rosetta (ROHL et al., 2004).

No que se refere à exploração de conhecimento acerca de preferências conformacionais de aminoácidos (Seção 4.1), no trabalho de Borguesan et al. (2015) é realizada uma demonstração de utilização da Lista de Probabilidades Angulares (APL, sigla em inglês), onde os autores mostram as contribuições da APL através da otimização de um conjunto de estruturas 3-D de proteínas realizada por meio de duas meta-heurísticas diferentes, que consistem em um AG e um algoritmo de otimização por enxame de partículas (PSO, sigla em inglês). Além disso, através da criação do servidor online NIAS[5] (BORGUESAN; INOSTROZA-PONTA; DORN, 2016), os autores disponibilizam à comunidade científica a criação de APLs, a fim de serem aproveitadas em métodos de modelagem de estruturas ou em quaisquer outros problemas que possam requerer o uso de preferências conformacionais de aminoácidos. Ainda, em um trabalho anterior de Inostroza-Ponta et al. (2015), desenvolvido pelos mesmos autores, foi demonstrada uma primeira tentativa de desenvolvimento de um AM que incorpora informações provenientes de uma variação da APL.

Além destes, diversas estratégias multiobjetivo também estão sendo propostas para

[5] <sbcb.inf.ufrgs.br/nias/>

lidar com o problema PSP. Geralmente, problemas mais complexos apresentam funções objetivo com diversos termos, sendo muitos deles conflitantes entre si, o que impossibilita a otimização simultânea dos mesmos de forma adequada (KONAK; COIT; SMITH, 2006). Dessa forma, a utilização de métodos multiobjetivo permite que termos conflitantes não compitam para encontrar as melhores soluções, e também possibilita a incorporação de novas propriedades acerca do problema, com o intuito de agregar informações aos termos de avaliação já considerados, e melhor guiar a otimização por meio de novas visões do espaço de busca (ZHOU et al., 2011). Cutello et al. (2006) desenvolveram um AE multiobjetivo na tentativa de contornar as deficiências das funções de energia. Os autores mostraram que os termos de energia mais comuns utilizados nos cálculos de interações moleculares, termos ligados e não-ligados, são conflitantes entre si. Com isso, estes dois termos foram tratados como dois objetivos distintos na otimização. No trabalho de Brasil et al. (2013), um método puramente *ab initio* que não utiliza nenhum tipo de conhecimento prévio acerca de estruturas determinadas experimentalmente foi proposto. Esta abordagem implementa um AE que conta com quatro objetivos distintos a ser minimizados. Rocha et al. (2016) propuseram um AG multiobjetivo, que utiliza como métrica de similaridade para seleção de indivíduos a estratégia de *phenotypic crowding*, onde duas soluções são selecionadas de acordo com as suas diferenças estruturais. Neste trabalho, os autores preocuparam-se com a manutenção da diversidade na fronteira de Pareto, incorporando a técnica de *crowding distance* do algoritmo *Non-dominated Sorting Genetic Algorithm* (NSGA-II) (DEB et al., 2002) como critério para inserção de novos indivíduos na população. Foram comparadas otimizações considerando apenas um único objetivo contra a mesma função desmembrada em dois e três objetivos.

Quanto à otimização multimodal aplicada ao problema PSP, Wong et al. (2010) apresentaram um AE que considera a similaridade estrutural das soluções nas operações de cruzamento e mutação. Islam e Chetty (2013) propuseram um AM onde durante a otimização é realizado um processo iterativo de agrupamento dos indivíduos baseado em suas respectivas conformações e similaridades. Como um dos últimos trabalhos publicados neste contexto, Garza-Fabre et al. (2016) propuseram um AM que associa como heurística de busca local o método Rosetta (Seção 3.1.1). Os autores desenvolveram operadores genéticos especializados por meio da incorporação de conhecimentos intrínsecos ao problema, visando ampliar a exploração de variadas conformações estruturais. Como alternativa às imprecisões das funções de energia e a rugosidade do espaço de busca, Garza-Fabre et al. utilizaram como técnica de seleção do AE, o procedimento de ranque-

amento estocástico (*stochastic ranking-based selection*), técnica utilizada em otimizações multimodais com o objetivo de minimizar a função de avaliação, porém mantendo a diversidade estrutural dos indivíduos da população. Além disso, o método incorpora uma versão modificada da inicialização de indivíduos, baseada em fragmentos, utilizada pelo Rosetta, na tentativa de manter um balanço apropriado entre exploração e refinamento do espaço de busca. O método utiliza as mesmas funções de energia e representações utilizados nos diferentes estágios do processo de otimização do Rosetta.

Contudo, apesar dos avanços obtidos na área de predição de estruturas de proteínas por meio do desenvolvimento de técnicas computacionais de otimização, principalmente pela exploração de conhecimento experimental e incorporação nos métodos de busca, sabe-se que as limitações de qualidade das bibliotecas de fragmentos estruturais e preferências conformacionais, ineficiências das funções de energia e a alta dimensionalidade do espaço conformacional, tornam o problema PSP, referente à categoria FM, extremamente complexo (KIM et al., 2009; GARZA-FABRE et al., 2016). Dessa forma, entende-se que as técnicas de busca, como alvo principal de melhoramentos, precisam incorporar de alguma maneira mecanismos mais robustos, capazes de gerar e manter estruturas energicamente aceitáveis e que ao mesmo tempo correspondam a distintas conformações, distribuídas em diferentes mínimos globais. A geração e preservação de um conjunto variado de soluções, frente a um problema multimodal (DAS et al., 2011), são fatores determinantes para se atingir resultados competitivos.

Diante disto, uma das meta-heurísticas mais proeminentes para resolver problemas complexos de otimização são os algoritmos meméticos. AMs, também referidos como algoritmos genéticos híbridos, podem ser definidos como meta-heurísticas híbridas que incorporam conceitos e operadores de métodos evolutivos e baseados em população com vistas a buscas globais, como os AGs, combinados a um método mais simples de busca local responsável pelo refinamento dos mínimos locais encontrados (MOSCATO, 1989). Dessa forma, os AMs baseiam-se na combinação de estruturas algorítmicas existentes, evitando assim a limitação de utilização de um único método para o problema, e proporcionando uma maior flexibilização no trato das complexidades envolvidas (KRASNOGOR; SMITH, 2005; MOSCATO; COTTA, 2010; NERI; COTTA; MOSCATO, 2012). Um dos maiores desafios na estruturação dos AMs consiste em definir como o espaço de soluções deve ser explorado. Para que estes algoritmos obtenham bons resultados, além de um desempenho satisfatório, é essencial que se encontre o balanceamento correto entre as técnicas de busca global e local incorporadas no algoritmo (MOSCATO; COTTA, 2010;

BOUSSAÏD; LEPAGNOT; SIARRY, 2013).

Nesta dissertação, pretende-se desenvolver um AM (MOSCATO, 1989) multi populacional para o problema PSP, que incorpore conceitos de AEs com vistas a explorações globais, como operadores de cruzamento modificados para melhor lidar com as propriedades intrínsecas do problema, à medida que utiliza estratégias de busca local voltadas ao refinamento do espaço de soluções multimodal, aliado ao conhecimento de estruturas 3-D de proteínas armazenadas no PDB, por meio da utilização de uma versão ampliada da estratégia de preferências conformacionais APL (BORGUESAN et al., 2015; BORGUESAN; INOSTROZA-PONTA; DORN, 2016). Define-se como principal objetivo para melhor lidar com as complexidades do PSP, atingir uma eficiente exploração do espaço de busca conformacional através da geração e manutenção da diversidade de soluções ao longo do processo de predição. Serão ainda incorporados à meta-heurística e ao processo de otimização, termos de energia adicionais agregados à função de energia do Rosetta (Seção 2.6.2) como forma de contornar algumas dificuldades apresentadas por esta.

Por fim, ressalta-se que este trabalho representa o seguimento de uma proposta de AM para a predição de estruturas 3-D de proteínas (CORRÊA et al., 2016), publicado pelo autor e colaboradores desta dissertação, apresentando modificações em algumas abordagens e inserções de novas estratégias objetivando melhores resultados. O artigo em questão será abordado na seção de Materiais e Métodos, juntamente com a metodologia de trabalho proposta.

3.4 Resumo do capítulo

Este capítulo apresentou uma visão geral dos trabalhos relacionados à área de otimização e meta-heurísticas, como as meta-heurísticas voltadas à otimização multimodal e aplicadas ao problema PSP. O capítulo também introduziu o CASP, e os experimentos de predição de estruturas de proteínas promovidos por este. A partir disto e conforme os resultados relacionados às últimas edições do CASP (TAI et al., 2014; KINCH et al., 2016a), os dois métodos mais relevantes para a área foram introduzidos, Rosetta e QUARK.

O próximo capítulo descreverá os algoritmos e estratégias utilizados na elaboração desta dissertação. Objetiva também apresentar a forma como o desenvolvimento do trabalho foi conduzido, estruturação dos métodos, e metodologia utilizada para isto.

4 MATERIAIS E MÉTODOS

Este capítulo tem por objetivo descrever os algoritmos e estratégias empregados na elaboração desta dissertação, bem como a metodologia e estruturação do método proposto. A partir do trabalho de Corrêa et al. (2016), foi desenvolvido um AM multi populacional (MOSCATO, 1989) para lidar com o problema PSP. O cerne desta meta-heurística consiste na organização da população de indivíduos em uma estrutura hierárquica em árvore, primeiramente proposta para o problema PSP por Inostroza-Ponta et al. (2015) e demonstrada em Corrêa et al. (2016). Em ambos os trabalhos, a estrutura de dados empregada configura uma árvore ternária, composta por treze nodos, a qual foi modificada nesta dissertação com o objetivo de prover à meta-heurística de otimização uma melhor exploração do espaço de busca conformacional. Tal modelo de organização populacional favorece o gerenciamento da performance do processo de exploração do algoritmo de busca sobre a superfície energética multimodal do problema, visto que cada nodo da árvore pode ser caracterizado como uma subpopulação independente, onde características individuais podem ser incorporadas e exploradas. O algoritmo também combina operadores de busca global, objetivando o aumento da capacidade de exploração do método voltando-se a características do PSP, aliados à uma implementação do algoritmo colônia artificial de abelhas (*Artificial Bee Colony*) (ABC, sigla em inglês) (KARABOGA; BASTURK, 2007), com o propósito de ser utilizado como uma técnica de busca local a ser aplicada em cada nodo da árvore.

Ainda, visando a redução da complexidade do espaço de busca por meio da restrição de possíveis conformações (valores angulares) adotadas pelos modelos estruturais, o algoritmo proposto incorpora o conhecimento prévio acerca de estruturas 3-D de proteínas determinadas experimentalmente e armazenadas no PDB. Para isso, foi utilizada a técnica APL, proposta primeiramente por Borguesan et al. (2015) e ampliada nos trabalhos de Corrêa et al. (2016) e Borguesan et al. (2016).

Por fim, o algoritmo de otimização concebe uma etapa anterior ao processo de otimização de estruturas de proteínas, denominada amostragem e inicialização de soluções, delineada à geração e classificação de diversos modelos estruturais para a proteína-alvo, a partir da APL, buscando a definição de diferentes grupos estruturais e a criação de melhores estruturas para serem incorporadas à meta-heurística como soluções iniciais das multi populações de otimização. Todas estas implementações serão detalhadas nas próximas seções.

4.1 Preferências conformacionais dos resíduos de aminoácidos

A estratégia de Lista de Probabilidades Angulares (CORRÊA et al., 2016; BORGUESAN; INOSTROZA-PONTA; DORN, 2016) busca determinar as preferências conformacionais dos resíduos de aminoácidos de uma determinada proteína-alvo, levando em consideração as ESs dos mesmos, através da análise de ocorrências em proteínas cuja estrutura foi determinada experimentalmente. No trabalho de Corrêa et al. (2016) foi demonstrado o uso da APL incorporada a um AM para o problema PSP, com o intuito de definir as preferências angulares de um aa considerando a influência de seus vizinhos, constituintes da mesma sequência de aminoácidos. Já no trabalho de Borguesan et al. (2016) foi proposto um servidor online, denominado NIAS[1] (*Neighbors Influence of Amino acids and Secondary structures*), com o objetivo de prover à comunidade uma ferramenta que possibilite a extração de informações acerca das preferências conformacionais dos aminoácidos através da geração de diferentes tipos de APLs.

A técnica APL foi desenvolvida a partir da análise de estruturas armazenadas no banco de dados PDB, e em ambos os trabalhos ela foi estruturada do mesmo modo. Segundo os trabalhos de Corrêa et al. (2016) e Borguesan et al. (2016), foi aplicado um conjunto de filtros para garantir a qualidade dos dados experimentais. Um conjunto de 11.130 estruturas de proteínas foi selecionado, onde todas foram determinadas experimentalmente através da técnica de cristalografia de raios-X. Para a construção da APL foram utilizadas apenas estruturas com resolução menor ou igual a 2,5Å, tendo sido depositadas no PDB até dezembro de 2014. A resolução tem por objetivo avaliar o nível de detalhe dos dados de difração de raios-X, o qual é considerado um bom indicador acerca da qualidade da estrutura experimental (HOVMÖLLER; ZHOU; OHLSON, 2002). Outro índice de avaliação de qualidade da estrutura considerado na filtragem de estruturas foi o R-observado (*R-observed*), empregado para avaliar a concordância entre o modelo cristalográfico e os dados experimentais de difração de raios-X (KLEYWEGT; BRÜNGER, 1996). Todas as estruturas com R-observado acima de 0,20 foram removidas do conjunto. Cada estrutura foi ainda analisada quanto ao seu grau de homologia em relação às outras proteínas do conjunto, sendo que apenas uma estrutura homóloga de cada grupo foi mantida para evitar posteriores redundâncias. As estruturas foram consideradas homólogas quando apresentaram identidade de sequências acima de 30%. A partir deste conjunto, os autores consideraram apenas os resíduos de aminoácidos que estavam

[1] <sbcb.inf.ufrgs.br/nias/>

posicionados corretamente em suas respectivas estruturas, utilizando como métricas de filtragem o fator-B (*B-factor*) com limiar menor ou igual a 30Å^2, e a ocupação (*occupancy*) com valor igual a 1. Ao final do processo de filtragem, resultaram 2.399.401 resíduos de aminoácidos a serem analisados posteriormente e utilizados na formação das APLs. O software STRIDE[2] (HEINIG; FRISHMAN, 2004) foi utilizado para designar a ES das estruturas experimentais. Atualmente, o STRIDE é um dos métodos mais populares utilizados para a atribuição de ESs de proteínas (DUPUIS; SADOC; MORNON, 2004). A Tabela 4.1, adaptada de Corrêa et al. (2016), resume os filtros empregados na construção da base de dados da APL.

Tabela 4.1: Conjunto de filtros aplicados na geração da base de dados da APL

Filtro	**Limiar**
Tamanho do conjunto	11.130 estruturas
Resolução	$\leq$ 2,5Å
R-observado	$\leq$ 0,20
Similaridade de sequências	$\leq$ 30%
Fator-B	$\leq$ 30Å^2
Ocupação	1
Total de aminoácidos extraídos	2.399.401

Fonte: Adaptado de Corrêa et al. (2016).

Sendo assim, para incorporar ao método de otimização as informações estruturais provenientes do conjunto de proteínas experimentais, foram utilizados diferentes tipos de APLs, propostos nos trabalhos de Corrêa et al. (2016) e Borguesan et al. (2016). Os autores construíram um banco de dados para representar as preferências conformacionais de aminoácidos, baseando-se em informações experimentais filtradas do PDB, conforme descrito acima. Para cada tipo de aa, foram computadas as ocorrências dos ângulos diedros ϕ e ψ nas estruturas experimentais do conjunto filtrado, e definido as preferências conformacionais dos aminoácidos considerando as suas respectivas ESs. A partir de análises realizadas sobre esta base de dados, os autores observaram que para uma dada ES, é possível perceber diferentes preferências conformacionais dependendo do aa, bem como que resíduos de aminoácidos iguais contendo diferentes ESs, podem apresentar preferências conformacionais particulares. Corrêa et al. (2016) e Borguesan et al. (2016) ampliaram a versão anterior da APL de Borguesan et al. (2015), na tentativa de melhor explorar as preferências conformacionais dos aminoácidos, encontrando informações mais especializadas.

[2]<webclu.bio.wzw.tum.de/stride/>

Originalmente, a APL considerava apenas as ocorrências angulares de um determinado aa e sua ES para a definição das preferências conformacionais deste aa, chamado de aminoácido de referência (aa_{ref}). Nesta nova versão da técnica, também é levado em consideração a influência que a vizinhança de aminoácidos exerce sobre o aa_{ref}. Com isso, além da APL-original ou **APL-1**, que considera apenas o aa_{ref} e sua ES, foram desenvolvidos três outros tipos de APLs: (*i*) **APL-2e** que considera a influência do aa da esquerda e sua respectiva ES; (*ii*) **APL-2d** que considera a influência do aa da direita e sua ES; e (*iii*) **APL-3** que considera a influência dos aminoácidos à esquerda e à direita e suas respectivas ESs sobre o aa_{ref}. Este conjunto de APLs será referenciado por **APL-vizinhança**. Ainda, no trabalho de Borguesan et al. (2016) foi proposta outra variação da APL. A **APL-centroide** objetiva considerar somente o aa_{ref} e sua ES, ignorando a vizinhança de aminoácidos, porém ainda considerando a ES dos mesmos, ou seja, para determinar as preferências conformacionais do aa_{ref}, a partir da APL-centroide, é utilizada apenas a sua própria ES e as ESs dos aminoácidos vizinhos sem considerá-los, aceitando assim qualquer tipo de aa que contenha a mesma ES. A APL-centroide pode ser dividida em três tipos, dependendo do número de ESs de aminoácidos considerado na geração das preferências conformacionais: (*i*) **APL-5** que considera a influência das ESs dos dois aminoácidos à esquerda e dos dois aminoácidos à direita sobre o aa_{ref} e sua ES; (*ii*) **APL-7** que considera a influência das ESs dos três aminoácidos à esquerda e dos três aminoácidos à direita; e (*iii*) **APL-9** que considera a influência das ESs dos quatro aminoácidos à esquerda e dos quatro aminoácidos à direita sobre o aa_{ref} e sua ES. Segundo os autores, esta variação da APL fez-se necessária devido à pouca quantidade de dados experimentais presentes em algumas combinações de aminoácidos e ESs da APL-3, fornecendo assim uma variedade maior de dados conformacionais aos usuários.

A Figura 4.1 ilustra os diferentes tipos de APLs através da sequência de aminoácidos "`CSTQKAQAK`" com ES "`HHHTTTEEE`", retirada de um segmento da proteína 1ACW. Nesta exemplificação, o aa escolhido como referência (aa_{ref}) foi o `K` (Lisina) com ES de volta (`T`), ambos representados na cor azul. Os aminoácidos vizinhos, influentes na determinação das preferências conformacionais do aa_{ref}, estão destacados na cor verde, de acordo com o tipo de APL. Nota-se que diferentemente de abordagens baseadas em fragmentos de aminoácidos (SIMONS et al., 1997), como apresentado anteriormente nos métodos Rosetta e QUARK, na APL cada combinação de aminoácidos é utilizada para atribuir os ângulos somente ao aa_{ref}, enquanto que nas abordagens baseadas em fragmentos os ângulos de todos os aminoácidos que englobam o fragmento são atribuídos.

Figura 4.1: Exemplificação dos diferentes tipos de APLs para a sequência de aminoácidos "CSTQKAQAK" com ES "HHHTTTEEE", retirada de um segmento da proteína 1ACW. As células de cor azul denotam o aa_{ref} e sua respectiva ES, e as células de cor verde, dependendo do tipo de APL, destacam os aminoácidos vizinhos considerados na definição das preferências conformacionais do aa_{ref}

EP	...	C	S	T	Q	K	A	Q	A	K	...
ES	...	H	H	H	T	T	T	E	E	E	...
APL-1	...	C	S	T	Q	K	A	Q	A	K	...
	...	H	H	H	T	T	T	E	E	E	...
APL-2e	...	C	S	T	Q	K	A	Q	A	K	...
	...	H	H	H	T	T	T	E	E	E	...
APL-2d	...	C	S	T	Q	K	A	Q	A	K	...
	...	H	H	H	T	T	T	E	E	E	...
APL-3	...	C	S	T	Q	K	A	Q	A	K	...
	...	H	H	H	T	T	T	E	E	E	...
APL-5	...	-	-	-	-	K	-	-	-	-	...
	...	H	H	H	T	T	T	E	E	E	...
APL-7	...	-	-	-	-	K	-	-	-	-	...
	...	H	H	H	T	T	T	E	E	E	...
APL-9	...	-	-	-	-	K	-	-	-	-	...
	...	H	H	H	T	T	T	E	E	E	...

Fonte: Do autor (2017).

Para transformar as informações oriundas do conjunto de proteínas experimentais filtrado do PDB em listas de preferências conformacionais, foram construídos histogramas ($H_{aa,z}$) de $[-180°, +180°] \times [-180°, +180°]$ células, organizadas em um espaço de estados discreto, onde $\{H_{aa,z}(i,j) = x | x \in \mathbb{N}\}$ e x denota o número de vezes (ocorrências) que um aa (ou combinação de aminoácidos) com ES z apresentou um par de ângulos de torção (ϕ e ψ) de valores i e j, respectivamente, no conjunto de estruturas experimentais. Observa-se que os ângulos de torção da base de dados de aminoácidos que deu origem à criação das APLs, tiveram seus valores arredondados para encaixarem-se nas células discretizadas dos histogramas. Cada $H_{aa,z}$ representa uma combinação diferente entre resíduo(s) de aminoácido(s) e estrutura(s) secundária(s), as quais são utilizadas na definição das preferências conformacionais do aa_{ref}, sendo que o número de aminoácidos

considerados em cada combinação varia conforme o tipo de APL.

Em relação às APLs que compõe o conjunto APL-vizinhança, foram geradas diferentes combinações de aminoácidos e ESs com tamanho de até três resíduos de aminoácidos (1-3 aa), considerando a vizinhança do aa_{ref} para combinações de comprimento maior do que 1 aa. Quanto às APLs do conjunto APL-centroide, foram estabelecidas combinações de diferentes tamanhos entre ESs para definir as preferências conformacionais do aa_{ref}, sendo que as combinações englobam tamanhos de ESs de 5 (APL-5), 7 (APL-7) e 9 (APL-9) aminoácidos.

Portanto, cada célula (i, j) do histograma $H_{aa,z}$ contém o número de vezes que um determinado aa (ou combinação de aminoácidos) possui um par de ângulos de torção (i$\leq \phi <$i+1, j$\leq \psi <$j+1) correspondente à ES z na base de dados experimental. Com o intuito de destacar as regiões conformacionais mais abundantes, para cada célula (i, j) de um dado histograma, é ainda somado o valor das oito células vizinhas, conforme a Equação 4.2. Onde r e s representam as posições (i, j) dos oito vizinhos de uma determinada célula na matriz do histograma $H_{aa,z}(r, s)$. Então, para cada histograma $H'_{aa,z}$ é calculado a Lista de Probabilidades Angulares (APL$_{aa,z}$) (Eq. 4.1) que representa a frequência normalizada de cada célula.

$$APL_{aa,z}(i,j) = \frac{H'_{aa,z}(i,j)}{\sum\limits_{x,y} H'_{aa,z}(x,y)} \tag{4.1}$$

$$H'_{aa,z}(i,j) = \sum_{r=i-1}^{i+1} \sum_{s=j-1}^{j+1} H_{aa,z}(r,s) \tag{4.2}$$

Nesta abordagem de geração de APLs, quatro histogramas da APL-vizinhança e três histogramas da APL-centroide são gerados para cada aa_{ref}. Para a APL-vizinhança, o primeiro histograma está relacionado à APL-1 e retorna a frequência relativa de ocorrências apenas do aa_{ref} e sua respectiva ES. O segundo arquivo retorna a frequência relativa de ocorrência do aa à direita da combinação (aa_{ref}) sob a influência do aa à esquerda (APL-2e), enquanto que o terceiro arquivo contém a frequência relativa de ocorrência do aa à esquerda da combinação (aa_{ref}) considerando a influência do aa à direita (APL-2d). O último arquivo retorna a frequência relativa de ocorrência para o aa central (aa_{ref}) da combinação considerando a influência dos resíduos à esquerda e à direita (APL-3). Tratando-se da APL-centroide, o primeiro arquivo retorna a frequência relativa de ocorrências para o aa_{ref} e sua respectiva ES sob a influência das ESs dos dois aminoácidos à esquerda e dos dois à direita (APL-5). O segundo histograma difere apenas no

tamanho da combinação de aminoácidos, considerando a influência das ESs dos três aminoácidos à esquerda e dos três aminoácidos à direita do aa_{ref} (APL-7). O último arquivo retorna a frequência relativa de ocorrências para o aa_{ref} considerando as ESs dos quatro aminoácidos à esquerda e dos quatro aminoácidos à direita (APL-9) da combinação. Para descrições mais detalhadas acerca deste esquema de preferências conformacionais de aminoácidos, bem como de suas estruturações e propriedades, é indicada a consulta aos trabalhos de Corrêa et al. (2016) e Borguesan et al. (2016).

Neste trabalho, todos os tipos de APLs descritos acima foram incorporados à etapa de inicialização de soluções do método de otimização, através da geração de diferentes combinações de aminoácidos (comprimento de 1-3 aa e 5-9 aa), na tentativa de alimentar o algoritmo com soluções de alta qualidade quando comparadas às oriundas de inicializações aleatórias, as quais compreendem todo o espaço de busca conformacional. Além disto, a APL-1 também foi utilizada na restrição do espaço de busca quando da aplicação de operadores de mutação, a fim de restringir posições angulares não existentes na APL. A forma como as APLs foram utilizadas será descrita nas próximas seções.

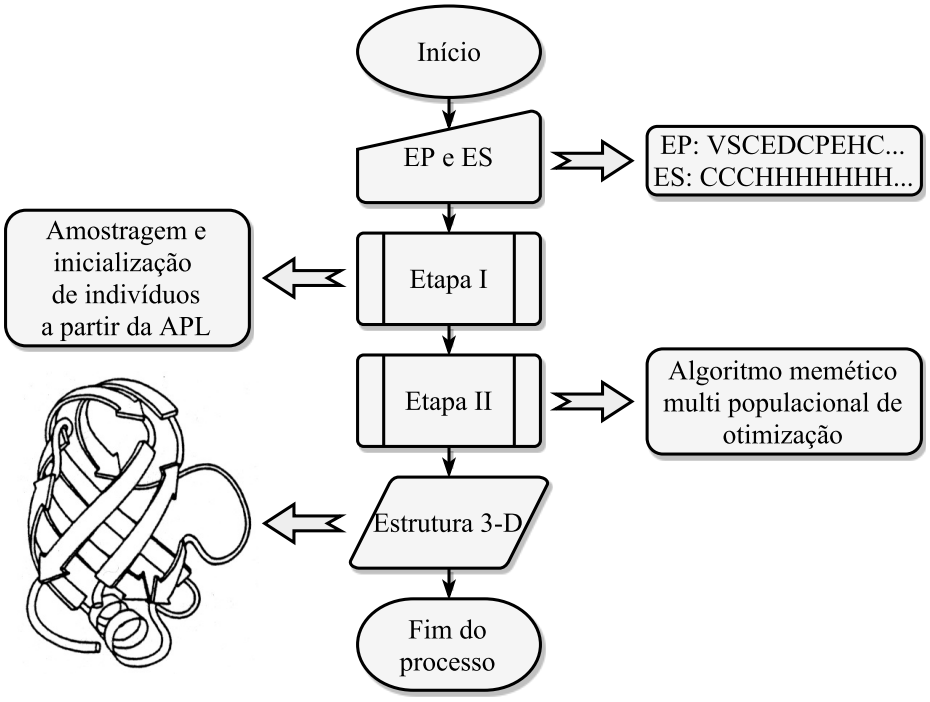
4.2 Método de otimização proposto

O método de otimização proposto neste trabalho para lidar com o problema PSP, pode ser dividido em duas etapas principais: (*i*) amostragem e inicialização de modelos estruturais para a proteína-alvo (Seção 4.3); e (*ii*) otimização das estruturas selecionadas a partir da etapa anterior (Seção 4.4). A Figura 4.2 ilustra o fluxograma de execução da abordagem proposta. O método recebe como parâmetros de entrada as estruturas primária e secundária da proteína-alvo e retorna a estrutura 3-D predita.

4.3 Etapa I: Amostragem e inicialização de modelos estruturais

A primeira etapa do método de otimização é responsável por realizar amostragens de indivíduos, recebendo como entrada de dados apenas a sequência linear de aminoácidos da proteína-alvo e sua respectiva ES. Estes indivíduos são inicializados a partir da técnica APL (Seção 4.1), utilizada para restringir o espaço conformacional de aminoácidos, considerando as probabilidades de ocorrências e preferências conformacionais de aminoácidos em estruturas previamente conhecidas. Nesta etapa, os indivíduos gerados

Figura 4.2: Fluxograma de execução, representado em alto nível, da abordagem de otimização proposta. O método recebe como parâmetros de entrada as estruturas primária e secundária da proteína-alvo e retorna a estrutura 3-D predita



Fonte: Do autor (2017).

passam por um processo de filtragem e agrupamento, visando contornar o problema da rugosidade do espaço de busca, com a finalidade de prover boas soluções iniciais ao AM empregado na etapa II de otimização de modelos estruturais.

A amostragem de indivíduos consiste na simples geração de soluções (modelos estruturais) para a proteína-alvo, considerando ou não as preferências conformacionais de aminoácidos. Neste caso, um indivíduo configura um vetor de ângulos diedros (variáveis de otimização), que descreve as cadeias polipeptídicas da proteína-alvo, caracterizando a representação computacional de modelos estruturais através do conjunto de ângulos de torção formados a partir da sequência linear de aminoácidos, descrita na Seção 2.4.

Compreendendo a complexidade proveniente da característica multimodal apresentada pela função de energia utilizada para guiar os processos de otimização de estruturas de proteínas, analisou-se o problema de que, geralmente, os métodos de busca inicializados com soluções ruins, quando aplicados a problemas complexos, tendem a convergir para resultados finais igualmente ruins. Isto deve-se às ineficiências dos métodos e a complexidade da função objetivo, visto que tais soluções possuem alta propensão a permanecer estagnadas em regiões desfavoráveis do espaço de busca durante o processo

de otimização ou prenderem-se a mínimos locais, fazendo com que haja desperdício de esforço computacional. Tratando-se de meta-heurísticas evolutivas baseadas em população, estes cenários conduzem à convergência prematura dos indivíduos a algum tipo de mínimo local, comprometendo todo o processo de otimização. Sabe-se que técnicas capazes de possibilitar explorações eficazes do espaço de soluções, as quais ofereçam ao algoritmo capacidade de gerar e manter indivíduos variados durante o processo, são fundamentais para contornar cenários deste tipo, contudo, estratégias que conduzam a boas inicializações de soluções também fazem-se necessárias na condição de auxiliar os métodos de busca e facilitar o processo de otimização.

Neste sentido, a etapa de amostragem de indivíduos iniciais foi idealizada como forma de explorar inicialmente o espaço de busca conformacional, tendo como objetivo localizar diferentes grupos estruturais para a proteína-alvo e melhorar a inicialização de soluções da meta-heurística. Ressalta-se que a etapa de inicialização é parte integrante de todo método de otimização, porém, na maioria das vezes, este processo se dá de forma aleatória, sem a consideração de propriedades específicas do problema. Ainda, a inicialização de soluções através da APL pode ser considerada uma forma de reduzir a complexidade do problema por meio da restrição de possibilidades conformacionais, sendo que a amostragem e classificação de variados indivíduos em grupos visam melhorar a exploração do espaço de busca conformacional e aumentar a diversidade do método. Devido a isto, estas duas estratégias foram empregadas de forma conjunta no procedimento.

Nesta etapa, primeiramente, são gerados 10.000 indivíduos com base nos dados de entrada da proteína-alvo, utilizando as preferências conformacionais dos aminoácidos a partir da combinação dos diferentes tipos de APLs. Este número de soluções geradas foi definido conforme as limitações de hardware de memória disponíveis e tempo de processamento, sendo que para amostragens maiores o consumo de memória é muito alto e o processo acaba tornando-se lento. As estruturas resultantes do processo de amostragem, são filtradas a partir de limiares relacionados a duas métricas de conformações estruturais, RG (LOBANOV; BOGATYREVA; GALZITSKAYA, 2008) e SASA (CONNOLLY, 1983; RICHMOND, 1984), que visam refletir o estado de empacotamento das estruturas. Os limiares de RG e SASA são estabelecidos de acordo com características específicas da proteína-alvo, as quais consideram o tamanho da sequência de aminoácidos e a sua classe estrutural, estabelecida conforme os arranjos apresentados na ES. A partir do tamanho de aminoácidos da proteína-alvo e de sua classe, os limiares máximos são definidos através da análise de estruturas de proteínas experimentais que sigam esse mesmo padrão

(tamanho e classe).

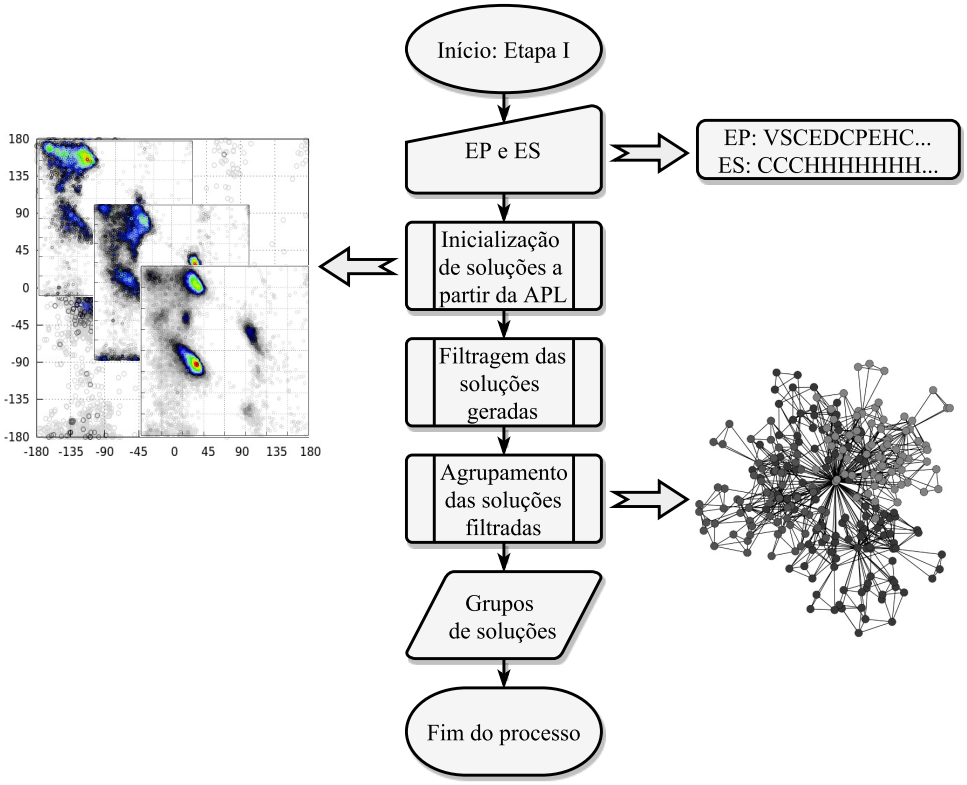
O RG de uma estrutura de proteína é definido pela média quadrática da distância entre os átomos da proteína e o seu centro de massa, e pode ser utilizado como indicador de empacotamento, visto que quanto menor for o RG, maior será a proximidade dos átomos com o centro da proteína (LOBANOV; BOGATYREVA; GALZITSKAYA, 2008). Do ponto de vista biológico, se uma estrutura de proteína estiver estável em seu estado nativo, provavelmente o RG se manterá estável. Porém, quando a proteína encontra-se fora de seu estado nativo (conformação menos estável), os valores de RG tendem a variar com frequência. O SASA é utilizado para medir o grau de exposição ao solvente de uma determinada estrutura de proteína, estimando a energia livre das interações entre soluto e solvente. Valores de SASA mais baixos indicam que menos resíduos da proteína estão expostos ao solvente, fazendo com que ela mostre-se mais compacta. Conformações de proteínas menos estáveis tendem a apresentar valores de SASA mais altos, pois existem mais resíduos em contato com o solvente, sendo que quando a proteína atinge conformações mais estáveis ocorrem perdas de área acessível ao solvente (ROSE et al., 1985). Nota-se que os cálculos de RG e SASA para os modelos estruturais foram realizados através das bibliotecas fornecidas pelo PyRosetta.

As estruturas resultantes do procedimento de filtragem são agrupadas de acordo com as suas similaridades estruturais através da métrica de similaridade *Root-Mean-Square Deviation* (RMSD) (ZHANG; SKOLNICK, 2004). O RMSD é utilizado para avaliar o grau de semelhança entre duas estruturas. Os grupos formados são ranqueados considerando o valor de RG médio do grupo, prezando pelo empacotamento das estruturas. Os grupos que apresentarem os menores valores são utilizados como indivíduos iniciais da meta-heurística. A Figura 4.3 ilustra o fluxograma de execução da etapa de amostragem de indivíduos. Cada processo desta etapa será detalhado nas próximas seções.

4.3.1 Inicialização de soluções

A amostragem de soluções é realizada por meio da inicialização de indivíduos, considerando as preferências conformacionais dos aminoácidos constituintes da proteína-alvo. Visando primeiramente a definição da melhor forma de inicializar as soluções a partir dos diferentes tipos de APLs disponíveis, avaliou-se a geração de soluções através de variadas combinações da APL, alternando a utilização na determinação dos ângulos

Figura 4.3: Fluxograma de execução, representado em alto nível, da etapa de amostragem e inicialização de indivíduos



Fonte: Do autor (2017).

diedros dos resíduos de aminoácidos.

Conforme descrito anteriormente, a APL por ser dividida em APL-vizinhança e APL-centroide, onde em cada uma destas divisões existem subtipos de APLs que consideram diferentes combinações de aminoácidos para definir as preferências conformacionais do aa_{ref}. A APL-vizinhança leva em consideração os aminoácidos vizinhos ao aa_{ref} juntamente com as suas respectivas ESs, formando combinações de 1, 2 ou 3 aminoácidos. Por outro lado, a APL-centroide considera apenas as ESs dos aminoácidos presentes na vizinhança do aa_{ref}, sem considerá-los, gerando combinações de 5, 7 ou 9 ESs de resíduos de aminoácidos.

Portanto, decidiu-se que cada aa da proteína-alvo será inicializado através de uma das duas APLs (APL-vizinhança ou APL-centroide), onde ambas possuem a mesma chance de serem escolhidas, ou seja, é dada a probabilidade de 50% para cada uma das duas. Sabendo ainda que, quanto maior for o tamanho da combinação entre aminoácidos na APL, mais específica e restrita será a lista de probabilidades angulares para um determinado aa_{ref}. Por este motivo, os subtipos das APLs são escolhidos seguindo a ordem de prioridade que vai de mais específicos (combinações maiores), com maiores chances

de serem escolhidos, a menos específicos (combinações menores). Dessa forma, se a APL-vizinhança for a escolhida para determinar os ângulos diedros do aa_{ref}, é dada a probabilidade de 70% à APL-3 ser utilizada, 20% à APL-2e e APL-2d, sendo que ambas possuem a mesma chance de serem selecionadas (50% para cada), e 10% à APL-1. O mesmo esquema de probabilidades é aplicado quando a APL-centroide é a escolhida, onde é dada a chance de 70% à APL-9 ser empregada, 20% à APL-7 e 10% à APL-5. Nota-se aqui que a APL-1, assim como a APL-5, englobam todos os dados das outras APLs mais específicas. Observa-se também que dependendo da posição do aa_{ref}, nem todos os tipos de APLs podem ser aplicados, por exemplo, os aminoácidos localizados nas extremidades da sequência de aminoácidos da proteína-alvo só podem ser inicializados através da APL-1 e da APL-2d, para o início da sequência, ou APL-1 e APL-2e, para a extremidade final, visto que os outros tipos de combinações não se encaixam.

O pseudocódigo do Algoritmo 1 descreve o procedimento que define qual das APLs será utilizada para atribuir os ângulos de torção a um determinado aa, conforme descrito acima. Assume-se que a posição do aa_{ref} na EP da proteína-alvo permite a utilização de qualquer um dos tipos de APLs disponíveis. A função $ConsultaAPL$ (Alg. 1, linhas 4-24) representa a escolha dos ângulos de torção para o aa_{ref} baseada na Lista de Probabilidades Angulares gerada para certa combinação, sendo necessárias as informações de EP e ES da proteína-alvo e a posição do aa_{ref} na sequência de aminoácidos. A função $NumeroAleatorio(0, 1)$ (Alg. 1, linhas 1-17) gera números aleatórios no intervalo contínuo de 0 a 1. O algoritmo retorna uma lista de ângulos diedros para o aa_{ref}.

O esquema de alternância de APLs foi definido a partir da análise de amostragens de 100.000 indivíduos, para um conjunto de 8 proteínas-alvo. Fez-se o processo de geração de indivíduos utilizando: (*i*) apenas a APL-1; (*ii*) apenas a APL-vizinhança, considerando as probabilidades descritas acima para os seus subtipos; (*iii*) apenas a APL-centroide, também considerando as probabilidades descritas acima para os seus subtipos; e (*iv*) APL-vizinhança e APL-centroide combinadas. Os indivíduos gerados foram comparados por meio das métricas RMSD em relação às estruturas experimentais, RG e SASA. Através das análises realizadas, não observou-se nenhuma superioridade entre os diferentes tipos de APLs quanto à qualidade de indivíduos gerados. Porém, o fato da combinação de APLs fornecer uma gama maior de dados experimentais, motivou a utilização desta em detrimento das outras. Nesta estratégia de geração de 10.000 indivíduos, este conjunto maior de informações torna-se interessante, visto que proporciona

Algoritmo 1: Definição da APL a ser utilizada na atribuição dos ângulos de torção a um determinado aminoácido

Entrada: EP, ES: estruturas primária e secundária, $pos_{aa_{ref}}$: posição do aa_{ref}

Saída: $angulos$: conjunto de ângulos diedros para o aa_{ref}

```
// Escolha entre APL-vizinhança e APL-centroide
```

se $0,5 \leq NumeroAleatorio(0,1)$ **então**

```
// APL-vizinhança
```

$probabilidade \leftarrow NumeroAleatorio(0,1)$

  se $probabilidade \leq 0,7$ **então**

    $angulos \leftarrow ConsultaAPL3(EP, EP, pos_{aa_{ref}})$

  senão

    se $probabilidade \leq 0,9$ **então**

      se $0,5 \leq NumeroAleatorio(0,1)$ **então**

        $angulos \leftarrow ConsultaAPL2e(EP, EP, pos_{aa_{ref}})$

      senão

        $angulos \leftarrow ConsultaAPL2d(EP, EP, pos_{aa_{ref}})$

      fim

    senão

      $angulos \leftarrow ConsultaAPL1(EP, EP, pos_{aa_{ref}})$

    fim

  fim

senão

```
// APL-centroide
```

$probabilidade \leftarrow NumeroAleatorio(0,1)$

  se $probabilidade \leq 0,7$ **então**

    $angulos \leftarrow ConsultaAPL9(EP, EP, pos_{aa_{ref}})$

  senão

    se $probabilidade \leq 0,9$ **então**

      $angulos \leftarrow ConsultaAPL7(EP, EP, pos_{aa_{ref}})$

    senão

      $angulos \leftarrow ConsultaAPL5(EP, EP, pos_{aa_{ref}})$

    fim

  fim

fim

retorna $angulos$

uma maior exploração das características contidas nos dados experimentais, resultando em melhores soluções. A combinação de APLs será referida como APL-combinada. Esta discussão será detalhada na seção de análises de resultados referentes à etapa I do método de otimização (Seção 5.1).

4.3.2 Filtragem de soluções

A filtragem de soluções foi inserida após a geração de indivíduos, objetivando a retirada de estruturas consideradas ruins, como forma de prevenir que estas sejam contabilizadas no processo de agrupamento de soluções e na definição dos grupos estruturais que serão utilizados na inicialização das populações da meta-heurística. Neste caso, estruturas ruins são caracterizadas pela falta de empacotamento, o que tende a representar que elas estão distantes de seus respectivos estados nativos.

Dessa forma, as estruturas resultantes do processo de amostragem de indivíduos, são filtradas a partir de limiares relacionados a duas métricas de avaliação de estruturas de proteínas, RG e SASA, utilizadas neste contexto como indicadores do nível de empacotamento dos modelos estruturais. Estes limiares foram definidos através da análise de 25.135 proteínas extraídas do PDB. Este conjunto de proteínas foi idealizado seguindo os mesmos elementos de filtragem utilizado na montagem da base de dados da APL, e estão resumidos na Tabela 4.2.

Tabela 4.2: Conjunto de filtros aplicados na geração da base de dados utilizada na definição dos limiares de RG e SASA

Filtro	**Limiar**
Tamanho do conjunto	25.135 estruturas
Resolução	$\leq$ 2,5Å
R-observado	$\leq$ 0,20
Similaridade de sequências	$\leq$ 30%
Fator-B	$\leq$ 30Å^2
Ocupação	1

Fonte: Adaptado de Corrêa et al. (2016).

Contudo, para a criação desta base de dados, todas as cadeias polipeptídicas ou subunidades (estrutura quaternária) de uma proteína presentes no seu arquivo PDB foram consideradas, sendo que na estruturação do banco de dados da APL, apenas a primeira cadeia polipeptídica de cada estrutura foi mantida (BORGUESAN; INOSTROZA-PONTA;

DORN, 2016). Observa-se que em alguns casos, as proteínas depositadas no PDB apresentam mais de uma cadeia polipeptídica, o que caracteriza a estrutura quaternária da proteína. Estas cadeias por vezes podem ser todas estruturalmente iguais ou apresentar diferenças. Para o propósito desta base dados, a manutenção de cadeias iguais não influência na definição dos limiares de RG e SASA, e por este motivo foram mantidas.

Para cada proteína presente na base de dados estruturada a partir do PDB, foi atribuída a sua ES utilizando o software STRIDE e calculado os valores de RG e SASA. A base de dados contou com 25.135 estruturas de proteínas de diversos tamanhos, variando de 5 resíduos de aminoácidos até 3.680. A variação dos valores de RG compreendeu o intervalo de 5,58Å a 72,03Å. Já os valores de SASA calculados compreenderam o intervalo de 708,1Å^2 a 92.532,6Å^2. A Tabela 4.3 resume as informações calculadas para as proteínas integrantes da base de dados criada.

Tabela 4.3: Resumo das informações calculadas a partir da base de dados de proteínas utilizada na definição dos limiares de RG e SASA

Informação	
Tamanho do conjunto	25.135 estruturas
Variação de tamanho	[5 aa, 3.680 aa]
Variação de RG	[5,58Å, 72,03Å]
Variação de SASA	[708,1Å^2, 92.532,6Å^2]

Fonte: Do autor (2017).

Conforme descrito na Seção 2.3, as proteínas podem ser classificadas de acordo com os arranjos de ESs assumidos. Na tentativa de obter informações mais específicas acerca das proteínas contidas no conjunto de proteínas idealizado, estas foram classificadas em classes de acordo com seus arranjos de ESs, resultando no cenário descrito na Tabela 4.4. As proteínas foram classificadas em quatro classes distintas, que compreendem: (*i*) classe de regiões menos estáveis, que engloba estruturas que apresentem mais de 80% de voltas ou alças na constituição da ES; (*ii*) classe de folhas, que compreende proteínas que apresentem o predomínio de mais de 60% de folhas β em suas ESs; (*iii*) classe de hélices, que abrangem estruturas de proteínas que possuem mais 60% de hélices na constituição da ES; e (*iv*) classe híbrida, que compreende estruturas que não se encaixam em nenhuma das classes anteriores, ou seja, apresentam uma combinação dos três tipos de ES na constituição da mesma. Observa-se nesta tabela, que a classe híbrida detêm o maior número de estruturas de proteínas.

A fim de relacionar o tamanho das proteínas determinadas experimentalmente com

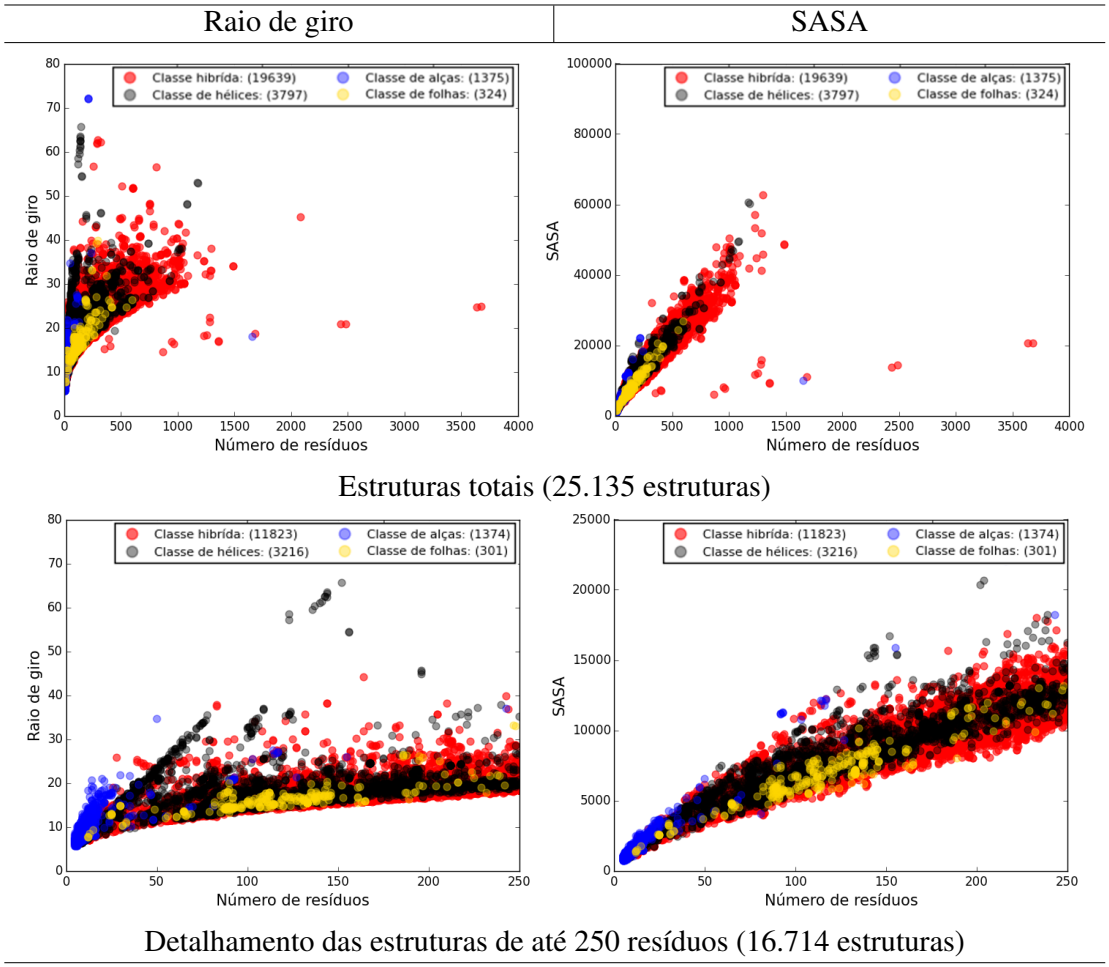
Tabela 4.4: Resumo das diferentes classes de proteínas geradas à partir da base de dados oriunda do PDB

Informação	
Classe de regiões menos estáveis	
Definição	Predominância de 80% de voltas ou alças
Tamanho do conjunto	1.375 estruturas
Variação de tamanho	[5 aa, 1.656 aa]
Variação de RG	[5,58 Å, 72,03 Å]
Variação de SASA	[708,1 Å^2, 22.044,6 Å^2]
Classe de folhas	
Definição	Predominância de 60% de folhas
Tamanho do conjunto	324 estruturas
Variação de tamanho	[12 aa, 598 aa]
Variação de RG	[7,6 Å, 39,8 Å]
Variação de SASA	[1.252,4 Å^2, 26.788,9 Å^2]
Classe de hélices	
Definição	Predominância de 60% de hélices
Tamanho do conjunto	3.797 estruturas
Variação de tamanho	[6 aa, 1.184 aa]
Variação de RG	[6,15 Å, 65,51 Å]
Variação de SASA	[858,0 Å^2, 60.509,3 Å^2]
Classe híbrida	
Definição	Estruturas híbridas
Tamanho do conjunto	19.639 estruturas
Variação de tamanho	[5 aa, 3.680 aa]
Variação de RG	[5,89 Å, 62,59 Å]
Variação de SASA	[772,1 Å^2, 92.532,6 Å^2]

Fonte: Do autor (2017).

as métricas calculadas de RG e SASA, as estruturas do conjunto experimental, classificadas em classes, foram dispostas em gráficos (Fig 4.4) que mostram a relação entre estes indicadores. Teoricamente, quanto maior o número de resíduos de aminoácidos de uma proteína, maior deverão ser os valores de RG e SASA, pois a estrutura 3-D da proteína é maior. Contudo, através da análise da Figura 4.4, percebe-se que existem proteínas que não seguem esta tendência. Isto pode ser atribuído aos diferentes arranjos de ESs e conformações presentes no conjunto, além das propriedades físico-químicas específicas dos aminoácidos que constituem as proteínas. A partir desta análise nota-se a importância de considerar tais componentes no processo de filtragem de soluções, visto que a classificação das proteínas em classes e a correlação do tamanho de sequência de aminoácidos com RG e SASA auxiliam no processo de exploração de características específicas da proteína-alvo quando relacionadas a estruturas de proteínas previamente conhecidas.

Figura 4.4: Distribuição das proteínas analisadas pertencentes às diferentes classes geradas a partir da base de dados oriunda do PDB, relacionando número de aminoácidos, RG e SASA



Fonte: do autor (2017).

Sendo assim, para uma determinada proteína-alvo, os limiares máximos de RG e SASA são definidos através da consulta à base de dados, relacionando o tamanho da sequência de aminoácidos e a sua classe. A partir do conjunto de proteínas experimentais retornadas por esta consulta, os limiares são atribuídos a partir dos maiores valores de RG e SASA atrelados às estruturas retornadas. Nota-se que as proteínas retornadas em resposta a consulta à base de dados possuem o mesmo número de resíduos de aminoácidos e a mesma classe da proteína-alvo. Por fim, tendo sido definidos os limiares de RG e SASA, as estruturas resultantes do processo de amostragem, descrito na seção anterior, são então filtradas com base nestes indicadores. As estruturas que ultrapassarem os limiares máximos definidos são descartadas do processo.

Ainda, com o intuito de analisar o comportamento do procedimento de filtragem de soluções, realizou-se processos de amostragem de 10.000 indivíduos, para um conjunto

de 8 proteínas-alvo, aplicando e não aplicando os limiares de RG e SASA para a exclusão de proteínas. Os indivíduos gerados foram comparados através da métrica RMSD em relação às estruturas determinadas experimentalmente. A partir da análise dos resultados, observou-se que para algumas proteínas-alvo há a redução significativa de soluções através do descarte, sendo que a maior parte abrange indivíduos com RMSD elevado, conseguindo manter um conjunto menor de soluções quando comparado ao processo de amostragem que não utilizou os limiares de exclusão. Os conjuntos de soluções oriundos dos processos de filtragem que utilizaram restrições de RG e SASA, englobam estruturas de RMSD relativamente mais baixos em comparação com as soluções descartadas, porém sabe-se que estas estruturas também estão contidas nos conjuntos resultantes dos processos que não consideraram os limiares. No entanto, estes conjuntos ainda consideram as soluções de RMSD mais altos que foram descartadas nos conjuntos com restrições de RG e SASA, sendo que estas também serão consideradas no processo de agrupamento e exercerão influência sobre a inicialização da meta-heurística. Esta discussão de resultados será detalhada na seção de análises de resultados referentes à etapa I do método de otimização (Seção 5.3).

4.3.3 Agrupamento de soluções

O agrupamento de dados é uma importante técnica baseada na classificação não-supervisionada de dados, podendo ser definida como sendo o processo de agrupar objetos de um mesmo conjunto de dados de entrada, levando em consideração seus padrões de similaridade ou dissimilaridade estabelecidos. O processo visa alocar, em um mesmo grupo, elementos com alto grau de similaridade e, em grupos diferentes, elementos com baixo grau de similaridade (JAIN; MURTY; FLYNN, 1999).

A organização de dados em grupos ajuda a mostrar, de uma forma mais clara, a estrutura interna do conjunto de dados, assim como seu comportamento, possibilitando que posteriormente estudos mais aprofundados possam ser feitos com base nos grupos formados. O agrupamento consiste em uma análise preliminar, visto que muitas vezes não se têm nenhum conhecimento sobre determinado conjunto de dados (XU; WUNSCH, 2005).

Neste sentido, o agrupamento de modelos estruturais tem por objetivo destacar os diferentes grupos conformacionais oriundos de um mesmo processo de criação. Este procedimento auxilia na descoberta de características específicas relacionadas à proteína-

alvo, por exemplo, durante o processo de otimização, o agrupamento de uma população de indivíduos pode revelar tendências conformacionais, mínimos locais de energia encontrados, e até mesmo indicar a convergência do algoritmo. Esta prática é bastante comum entre os métodos de referência na área, como pode ser observado no Rosetta (ROHL et al., 2004) e QUARK (XU; ZHANG, 2012).

Sendo assim, nesta etapa, as estruturas resultantes do processo de filtragem, formados a partir da amostragem de indivíduos, são agrupadas de acordo com as suas similaridades estruturais através da métrica RMSD, objetivando a identificação de diferentes padrões conformacionais para a proteína-alvo. Os grupos formados são ranqueados considerando o valor de RG médio do grupo, sendo que aqueles que apresentarem os menores valores são utilizados como indivíduos iniciais da meta-heurística, prezando pelo empacotamento das estruturas. Observa-se que os grupos gerados também poderiam ser ranqueados de acordo com o número de indivíduos em cada grupo, valor de energia médio ou valor de SASA médio. No entanto, optou-se pelos valores de RG médio dos grupos devido à uma análise de experimentos de agrupamento realizada utilizando estas métricas. O RG médio mostrou-se capaz de ordenar os grupos gerados, de forma que as melhores soluções tendessem a ficar entre os primeiros grupos, os quais dão origem as populações do AM. A discussão detalhada destes experimentos será desenvolvida na seção de análises de resultados referentes à etapa I do método de otimização (Seção 5.4).

O agrupamento dos modelos estruturais foi realizado através da técnica de clusterização hierárquica aglomerativa. A clusterização de dados hierárquica gera uma sequência de grupos aninhados, formando uma estrutura de hierarquia entre eles. O nível mais alto desta hierarquia consiste de um único grupo que compreende todas as outras partições formadas, e o nível mais baixo contém cada objeto individualmente alocado em um grupo. O agrupamento hierárquico pode ocorrer de forma aglomerativa ou divisiva (JAIN; MURTY; FLYNN, 1999). O agrupamento aglomerativo é considerado um processo *bottom-up*, ou seja, inicialmente cada objeto do conjunto de dados representa um grupo individual, e a cada passo do processo, os grupos mais similares ou próximos são agrupados, considerando alguma medida de proximidade, até que ao final do processo todos os grupos formados pertençam a um único grupo. Este é o processo mais utilizado nos algoritmos hierárquicos e por isto foi empregado nesta etapa.

A medida de proximidade utilizada para avaliar o grau de similaridade entre diferentes estruturas foi a métrica estrutural RMSD. O critério de avaliação da distância entre os grupos (*inter-cluster distance*) foi calculado a partir da estratégia de ligação completa

(*complete linkage*), a qual considera a maior distância entre dois objetos de grupos distintos para calcular a distância entre grupos. O limiar de corte (*cut-off*) adotado no processo respeita a formação de no mínimo 9 diferentes grupos, onde cada grupo deve conter 24 ou mais estruturas. Estes limiares respondem a estruturação do AM que será explicado nas próximos seções. Nota-se que os processos de agrupamento foram realizados através das bibliotecas de clusterização fornecidas pelo software SciPy[3][4].

Por fim, a etapa de amostragem e classificação de indivíduos, a qual culmina com o agrupamento dos mesmos em diferentes grupos estruturais, é empregada para alimentar as multi populações do AM, visando a diversidade das populações de indivíduos, na tentativa de contornar os problemas da multimodalidade da função objetivo, ao mesmo tempo que preocupa-se com a qualidade das soluções iniciais.

4.4 Etapa II: Otimização de estruturas

A segunda parte do método proposto consiste no processo de otimização das estruturas oriundas da etapa de amostragem e inicialização de modelos estruturais. Nesta etapa, foi desenvolvido um AM multi populacional para o problema PSP, que organiza a população de indivíduos em uma estrutura hierárquica em árvore, primeiramente proposta por Inostroza-Ponta et al. (2015) e utilizada no trabalho de Corrêa et al. (2016). Em ambos os trabalhos, tal estrutura foi parametrizada como uma árvore ternária composta por treze nodos, a qual foi mantida neste trabalho.

O AM foi idealizado como seguimento do algoritmo proposto em Corrêa et al. (2016), o qual incorporou uma etapa anterior de amostragem e classificação de soluções, conforme descrito na seção anterior, recebeu modificações relacionadas à operação e interações dos componentes da estrutura em árvore, e também foram integrados operadores genéticos voltados ao problema PSP, como operações de cruzamento considerando a ES da proteína-alvo e mutação baseada em regiões mais flexíveis da proteína.

Diferentemente do AM proposto por Corrêa et al. (2016), cada nodo da árvore ternária foi modificado de forma a caracterizar uma subpopulação independente, sendo que cada um recebe como entrada um grupo estrutural diferente oriundo da etapa anterior, e possui internamente uma meta-heurística de otimização de execução independente. Ao longo do processo, estes nodos interagem através de operações de cruzamento como

[3] <www.scipy.org>
[4] <www.goo.gl/FWK8CB>

forma de diversificar as populações e combinar conhecimentos, e a partir de diferentes conformações estruturais prover uma exploração mais eficaz do espaço de soluções. Esta estruturação multi populacional, a qual engloba otimizações independentes e interações entre diferentes populações, foi concebida visando contornar os problemas da multimodalidade presentes na função objetivo empregada no processo de otimização, por meio da geração e preservação da diversidade de soluções.

As subpopulações do AM representam grupos de soluções independentes que são otimizados através de uma meta-heurística própria, e por intermédio de interações entre as subpopulações tenta-se fazer com que soluções boas, localizadas em diferentes regiões do espaço de busca representadas na forma de nichos, emerjam. Na otimização de soluções empregou-se uma versão modificada do algoritmo ABC (KARABOGA; BASTURK, 2007), baseado no comportamento de forrageamento das abelhas, onde cada nodo da árvore incorpora uma execução independente deste algoritmo.

O ABC possui a capacidade tanto de exploração quanto de refinamento local. No entanto, modificou-se alguns mecanismos do algoritmo visando uma melhor adequação ao problema PSP. Foi adicionado internamente ao algoritmo operações de cruzamento, na tentativa de ampliar o seu potencial exploratório, bem como restrições nas operações de mutação considerando regiões desfavoráveis do espaço de busca conformacional, indicadas através de preferências conformacionais de aminoácidos extraídas da APL-1. Sendo assim, cada componente do AM proposto será detalhado nas próximas seções.

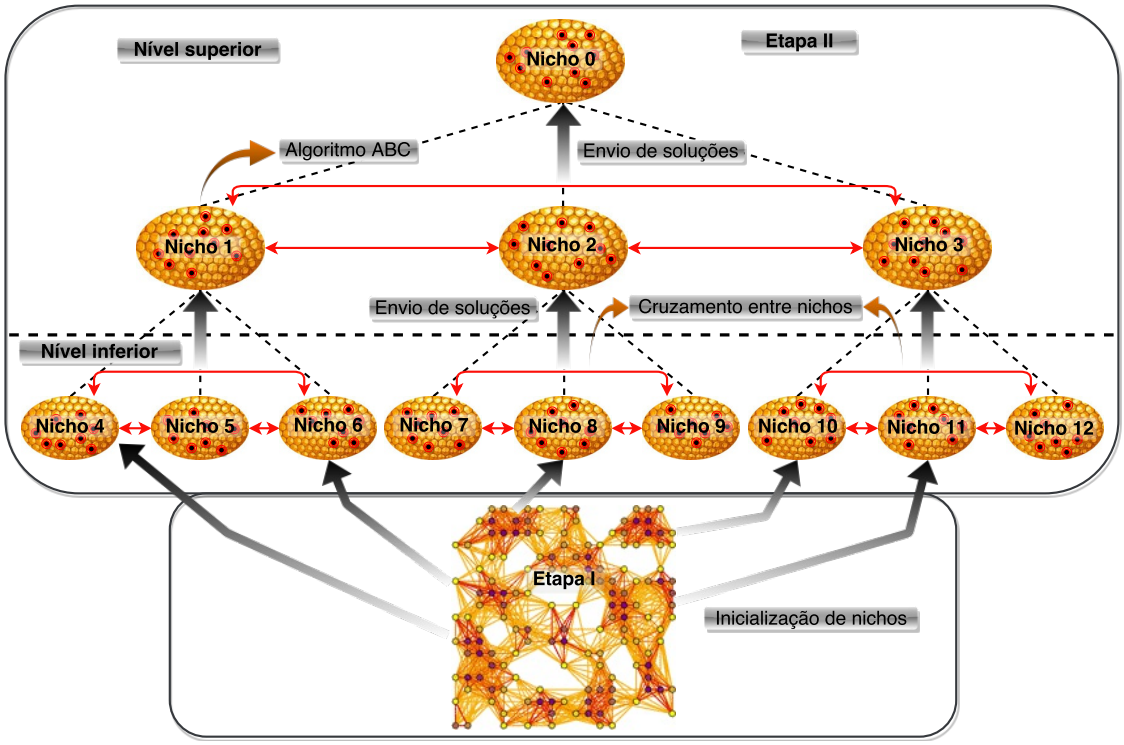
4.4.1 Estrutura algorítmica do método proposto

A população de indivíduos do AM organiza-se em uma estrutura em árvore ternária com treze nodos, que caracterizam treze subpopulações independentes. Esta estrutura é divida em dois níveis: (*i*) nível superior e (*ii*) nível inferior. O primeiro nível abrange quatro subpopulações e o segundo compreende as nove populações restantes. Tal divisão é utilizada para determinar como se dará o processo de inicialização de soluções em cada subpopulação, bem como ilustrar as interações entre as diferentes populações. Observa-se que o conjunto envolvendo, subpopulação, algoritmo de otimização e interações com outras populações será referido neste trabalho como nicho.

Conforme ilustra a Figura 4.5, cada subpopulação mantém um conjunto de soluções que são otimizadas por meio da execução independente do algoritmo ABC, que será explicado na Seção 4.4.5. Ainda, os nichos 1, 2 e 3 pertencentes ao nível superior da

árvore, recebem soluções oriundas de operações de cruzamento realizadas entre as subpopulações localizadas no nível inferior da estrutura. O nicho 0, raiz da árvore, recebe as soluções resultantes dos processos de cruzamento efetuados entre as subpopulações 1, 2 e 3. As linhas destacadas em vermelho na Figura 4.5 representam as operações de cruzamento entre as subpopulações. Este envio de soluções tem a intenção de combinar conhecimentos provenientes de diferentes grupos estruturais, diversificar a população de indivíduos e favorecer a exploração do espaço de busca. No entanto, cada nicho só pode interagir com os nichos vizinhos pertencentes a mesma subpopulação mãe, conforme determina a hierarquia da árvore, e com a sua subpopulação mãe propriamente dita. As etapas executadas pelo algoritmo serão resumidas abaixo e detalhadas nas próximas seções.

Figura 4.5: Organização hierárquica da população implementada no AM



Fonte: do autor (2017).

A inicialização das subpopulações do AM ocorre de duas maneiras, conforme a divisão em níveis da estrutura em árvore. Este procedimento está ilustrado no pseudocódigo do Algoritmo 2. Os indivíduos dos nichos pertencentes ao nível inferior da árvore (4-12), são inicializados a partir dos diferentes grupos estruturais oriundos da etapa anterior de amostragem de indivíduos (Alg. 2, linha 2), justificando assim o motivo do processo de agrupamento utilizar como um dos limiares de corte a formação de pelo menos 9 grupos

de estruturas. Já os indivíduos dos nichos vinculados ao nível superior da árvore (0-3), são inicializados através de operações de cruzamento realizadas entre os seus filhos, localizados um nível abaixo na hierarquia. Cada subpopulação possui um conjunto de N soluções. Dessa forma, os nichos do nível inferior ordenam de forma crescente as soluções contidas nos grupos recebidos considerando os seus valores de energia (Alg. 2, linha 3), e utilizam os N primeiros indivíduos como soluções iniciais, descartando as excedentes (Alg. 2, linha 4). Os nichos do nível superior recebem exatamente N estruturas provenientes das operações de cruzamento realizadas entre os seus filhos.

Observa-se que as operações de cruzamento entre nichos ocorrem também durante o processo de otimização do AM, diferindo apenas no número de soluções geradas e enviadas aos nichos do nível superior. Estas operações, bem como as interações envolvidas, serão detalhadas mais abaixo. O pseudocódigo do Algoritmo 3 ilustra o procedimento genérico de cruzamento entre nichos, representado pela função $CruzamentoEntreNichos$ (Alg. 2, linha 6), a qual recebe como parâmetro o número de soluções que deverão ser enviadas a cada nicho do nível superior.

Algoritmo 2: Descrição do processo de inicialização de nichos do AM

Entrada: N : número de indivíduos em cada subpopulação, $grupos[9]$: grupos estruturais oriundos da etapa I
Saída: nichos inicializados

```
// Inicialização dos nichos do nível inferior
```
1 **para cada** $nicho_i$, $i \leftarrow 4 : 12$ **faça**
2 $\quad nicho_i \leftarrow grupos[i]$
3 $\quad OrdenaPopulacao(nicho_i)$
4 $\quad DescartaIndividuosExcedentes(nicho_i)$
5 **fim**
```
// Inicialização dos nichos do nível superior
```
6 $CruzamentoEntreNichos(N)$

Após a inicialização das subpopulações do AM, as etapas enumeradas abaixo são executadas a cada geração, sendo que a condição de parada do algoritmo é determinada pelo número de cálculos de energia realizados durante a execução:

1. Cada nicho (0-12) realiza a execução do algoritmo ABC de forma independente com o propósito de otimizar o seu conjunto de soluções. O ABC é executado por g_{ABC} gerações em cada nicho (Seção 4.4.5);
2. Os nichos do nível inferior realizam operações de cruzamento com os seus vizinhos, vinculados ao mesmo pai, de modo a gerar 9 soluções resultantes (*offsprings*) para cada grupo de vizinhos. Neste caso, o primeiro grupo representa os nichos 4, 5 e 6,

o segundo grupo engloba os nichos 7, 8 e 9, e o terceiro grupo abrange os nichos 10, 11 e 12 (Alg. 3, linhas 1-18);

3. As soluções resultantes do cruzamento entre os nichos do nível inferior são enviadas aos seus respectivos pais na hierarquia. As soluções recebidas são integradas na população de cada pai (Alg. 3, linha 20), onde ocorre a ordenação dos indivíduos considerando os valores de energia (Alg. 3, linha 21), e o descarte das piores soluções excedentes, de forma a garantir N indivíduos em cada subpopulação (Alg. 3, linha 22);
4. Os nichos 1, 2 e 3 realizam operações de cruzamento entre eles, seguindo o mesmo processo descrito acima, de modo a gerar 9 soluções resultantes (*offsprings*). São realizadas 9 operações de cruzamento, visto que cada operação resulta em uma solução, sendo que 3 delas ocorrem entre os nichos 1 e 2, outras 3 ocorrem entre os nichos 1 e 3, e as outras 3 restantes entre os nichos 2 e 3 (Alg. 3, linhas 24-39). O mesmo procedimento equivale para cada grupo de vizinhos do nível inferior;
5. As soluções resultantes do cruzamento entre os nichos 1, 2 e 3 são enviadas ao nicho 0. As soluções recebidas por ele são integradas na sua população (Alg. 3, linha 41), onde ocorre a ordenação dos indivíduos considerando os valores de energia (Alg. 3, linha 42), e o descarte das piores soluções excedentes (Alg. 3, linha 43), de forma a garantir N indivíduos na subpopulação. Todas estas operações de cruzamento e envio de soluções estão demonstradas no Algoritmo 3;
6. Cada nicho é capaz de reinicializar uma parcela de R indivíduos de sua população, caso esta atinja uma convergência prematura. A verificação de convergência se dá de forma independente em cada nicho, considerando apenas a sua própria subpopulação (Seção 4.4.4);

Nota-se que cada etapa descrita acima foi implementada de forma sequencial.

O Algoritmo 4 exibe o pseudocódigo do AM proposto, o qual recebe como parâmetro o número máximo de cálculos de energia a serem realizados e retorna um conjunto contendo a melhor solução de cada um dos nichos.

4.4.2 Representação de indivíduos

Conforme descrito na Seção 2.5 de definições do problema PSP, cada aa constituinte da proteína-alvo pode ser representado através de um vetor composto por sete

Algoritmo 3: Demonstração da função $CruzamentoEntreNichos$

Entrada: NS : número de soluções a serem enviadas aos pais

```
// Seleção e cruzamento do nível inferior
```

para $i \leftarrow 1 : 3$ **faça**
  $conjunto_{offsprings}[NS] \leftarrow \emptyset$
  $g_{inds} \leftarrow NS/3$
  $k \leftarrow 0$
  para $iteracao \leftarrow 1 : g_inds$ **faça**
    $j \leftarrow 1$
    $pai_1 \leftarrow$ *seleção por ranking de uma solução do* $nicho_{i*3+j}$
    $pai_2 \leftarrow$ *seleção por ranking de uma solução do* $nicho_{i*3+j+1}$
    $conjunto_{offsprings}[k] \leftarrow cruzamentoES(pai_1, pai_2)$
    $k \leftarrow k + 1$
    $pai_1 \leftarrow$ *seleção por ranking de uma solução do* $nicho_{i*3+j}$
    $pai_3 \leftarrow$ *seleção por ranking de uma solução do* $nicho_{i*3+j+2}$
    $conjunto_{offsprings}[k] \leftarrow cruzamentoES(pai_1, pai_3)$
    $k \leftarrow k + 1$
    $pai_2 \leftarrow$ *seleção por ranking de uma solução do* $nicho_{i*3+j+1}$
    $pai_3 \leftarrow$ *seleção por ranking de uma solução do* $nicho_{i*3+j+2}$
    $conjunto_{offsprings}[k] \leftarrow cruzamentoES(pai_2, pai_3)$
    $k \leftarrow k + 1$
  fim
  $nicho_i \leftarrow nicho_i + conjunto_{offsprings}$
  $OrdenaPopulacao(nicho_i)$
  $DescartaIndividuos(nicho_i)$
fim

```
// Seleção e cruzamento do nível superior
```

$conjunto_{offsprings}[NS] \leftarrow \emptyset$
$k \leftarrow 0$
para $iteracao \leftarrow 1 : g_inds$ **faça**
  $j \leftarrow 1$
  $pai_1 \leftarrow$ *seleção por ranking de uma solução do* $nicho_{0*3+j}$
  $pai_2 \leftarrow$ *seleção por ranking de uma solução do* $nicho_{0*3+j+1}$
  $conjunto_{offsprings}[k] \leftarrow cruzamentoES(pai_1, pai_2)$
  $k \leftarrow k + 1$
  $pai_1 \leftarrow$ *seleção por ranking de uma solução do* $nicho_{0*3+j}$
  $pai_3 \leftarrow$ *seleção por ranking de uma solução do* $nicho_{0*3+j+2}$
  $conjunto_{offsprings}[k] \leftarrow cruzamentoES(pai_1, pai_3)$
  $k \leftarrow k + 1$
  $pai_2 \leftarrow$ *seleção por ranking de uma solução do* $nicho_{0*3+j+1}$
  $pai_3 \leftarrow$ *seleção por ranking de uma solução do* $nicho_{0*3+j+2}$
  $conjunto_{offsprings}[k] \leftarrow cruzamentoES(pai_2, pai_3)$
  $k \leftarrow k + 1$
fim
$nicho_0 \leftarrow nicho_0 + conjunto_{offsprings}$
$OrdenaPopulacao(nicho_0)$
$DescartaIndividuos(nicho_0)$

Algoritmo 4: Pseudocódigo do AM proposto

Entrada: número máximo de cálculos de energia
Saída: $sol_{nichos}[13]$: vetor contendo a melhor solução de cada nicho

Processamento da etapa I de amostragem de indivíduos
Inicialização da população conforme a divisão em níveis da estrutura em árvore
enquanto *critério de parada não for satisfeito* **faça**
    para cada $nicho_i$, $i \leftarrow 0 : 12$ **faça**
        `// Execução do algoritmo ABC`
        $ABC(nicho_i)$
    fim
    `// Operações de cruzamento entre nichos gerando 9 soluções a serem enviadas a cada pai`
    $CruzamentoEntreNichos(9)$
    `// Definição das melhores soluções e verificação de convergência`
    para cada $nicho_i$, $i \leftarrow 0 : 12$ **faça**
        $sol_{nichos}[i] \leftarrow$ *melhor solução do* $nicho_i$
        se *atingida convergência prematura* **então**
            reinicialização de uma parcela R *da população do* $nicho_i$
        fim
    fim
fim

valores reais ($\mathbb{R}$). Três deles representam os ângulos diedros ϕ, ψ e ω da cadeia principal, e os valores restantes representam os ângulos diedros χ da cadeia lateral. Nota-se que nem todos os aminoácidos possuem os quatro ângulos χ.

Sendo assim, a representação computacional de uma dada solução X com n resíduos de aminoácidos é definida por um vetor de valores reais de tamanho $n \times 7$ (Eq. 4.3).

$$X = (x_{1_\phi}, x_{1_\psi}, x_{1_\omega}, x_{1_{\chi_1}}, x_{1_{\chi_2}}, \cdots, x_{n_\phi}, x_{n_\psi}, \cdots, x_{n_{\chi_3}}, x_{n_{\chi_4}}) \tag{4.3}$$

4.4.3 Seleção e operadores de cruzamento

As operações de cruzamento entre indivíduos são aplicadas de forma global entre as soluções integrantes dos diferentes nichos do AM, e internamente no algoritmo ABC. A seleção de indivíduos para o cruzamento é realizada através da aplicação da estratégia de seleção por *ranking*, que consiste em ordenar os indivíduos de acordo com seus valores de energia e atribuir probabilidades decrescentes de serem escolhidos. Estas probabilidades são proporcionais a razão entre a posição na ordenação das soluções e o número total de

indivíduos da população. Contudo, o melhor indivíduo (menor valor de energia) recebe a posição mais alta na ordenação e a pior solução recebe a posição de número 1, visto que os indivíduos com valores de energia menores devem apresentar mais chances de serem selecionados. Com isso, a escolha das soluções é feita de forma aleatória considerando as probabilidades estabelecidas. Nota-se que esta estratégia é uma variação da seleção por roleta viciada, que visa considerar a posição dos indivíduos na ordenação da população ao invés de considerar o valor de energia em si. Optou-se por esta estratégia devido à possibilidade de valores negativos na função de energia.

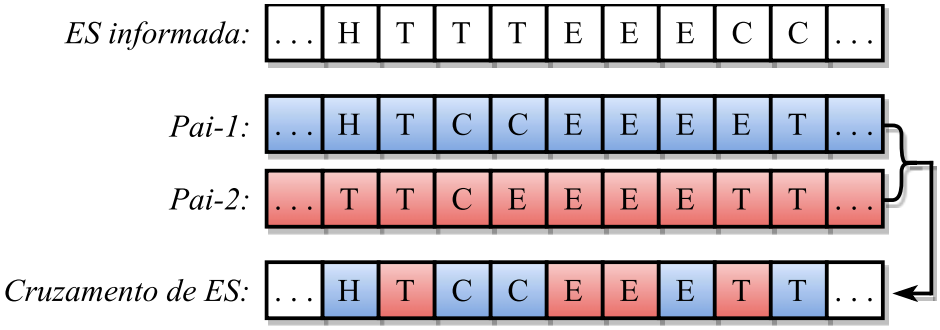
Com o objetivo de explorar de uma forma mais eficaz as propriedades intrínsecas relacionadas ao problema PSP, foi utilizada a operação de cruzamento proposta no trabalho de Corrêa et al. (2016). O operador de cruzamento uniforme de ES (*cruzamentoES()* no Algoritmo 3), objetiva manter a similaridade encontrada durante o processo de otimização entre as ESs das soluções que estão sendo otimizadas (pais na operação de cruzamento) e a sequência de ES informada previamente, como parâmetro de entrada do algoritmo. Esta estratégia tem por objetivo replicar nos filhos as formações corretas de ESs percebidas nos pais.

Primeiramente, dois indivíduos da população são selecionados através da estratégia por *ranking*, sendo que estes caracterizam os pais na operação de cruzamento que resultará em uma nova solução. Similar a ideia do cruzamento uniforme (SYSWERDA, 1989), para cada aa (conjunto de ângulos x_i no vetor X da Equação 4.3) da proteína-alvo, os ângulos são escolhidos tanto do pai 1 (com probabilidade de 50%) quanto do pai 2 (também com probabilidade de 50%). Este procedimento ocorre se ambas as ESs relacionadas aos aminoácidos dos pais forem iguais ou diferentes a ES informada para o respectivo aa. Se apenas um dos pais apresentar ES igual a ES informada para um aa específico, então o conjunto de ângulos correspondente ao aa em questão, pertencente a este pai, é escolhido para compor a solução resultante. Todos os aminoácidos da proteína são comparados durante o cruzamento. A Figura 4.6 ilustra o funcionamento do operador de cruzamento de ES.

4.4.4 Procedimento de reinicialização

O procedimento de reinicialização de populações é considerado um dos componentes básicos de um AM, sendo utilizado como meio de evitar a convergência prematura da população e auxiliar o método a escapar de mínimos locais (MOSCATO; COTTA,

Figura 4.6: Exemplo do operador de cruzamento de ES. *ES informada* representa a ES informada como entrada para determinada proteína-alvo, *Pai-1* e *Pai-2* representam as ESs dos indivíduos selecionados para o cruzamento, e *Cruzamento de ES* é o resultado da operação de cruzamento de ES aplicada sobre os pais selecionados



Fonte: do autor (2017).

2010).

No AM proposto, cada nicho é capaz de efetuar o reinício de uma parcela R da população, de forma independente dos outros. O critério empregado para determinar a aplicação do procedimento de reinicialização foi o coeficiente de variação (CV), que consiste em uma medida de dispersão relativa, utilizada para estimar o grau de diversidade de amostras (BEDEIAN; MOSSHOLDER, 2000). O CV representa o desvio-padrão de uma determinada amostra expresso como porcentagem da média, conforme a Equação 4.4. Valores de CV baixos indicam que o conjunto amostral é similar.

$$CV = \frac{S}{\bar{x}} \tag{4.4}$$

Utilizou-se o RG dos modelos estruturais como métrica para o cálculo do CV de uma determinada população de indivíduos, visto que os valores da função de energia podem assumir valores negativos, o que pode prejudicar no cálculo do coeficiente. O RG não é capaz de identificar diferenças conformacionais entre estruturas com valores similares, porém é um bom indicador do nível de empacotamento das mesmas. Sendo assim, populações que apresentam CV baixo tendem a englobar modelos estruturais com níveis de empacotamento similares, o que indica que a população está convergindo para uma determinada conformação. Com isso, a população de um determinado nicho é reinicializada caso apresente CV menor do que $p\%$, sendo que o CV é expresso em porcentagem. O parâmetro p, portanto, representa o limiar utilizado para expressar se determinada população é similar ou não, de acordo com o CV relativo ao RG dos integrantes da população.

A reinicialização de cada um dos nichos da árvore ocorre de maneira diferente,

porém todos descartam os R piores indivíduos, considerando os seus valores de energia, e inserem R novas soluções. Os nichos pertencentes ao nível inferior da estrutura (4-12), inserem novos indivíduos a partir dos grupos estruturais retornados pela etapa de amostragem de indivíduos. A escolha de soluções é feita de forma aleatória, porém cada nicho escolhe apenas as soluções contidas no seu próprio grupo. Os nichos 1, 2 e 3, localizados no nível superior da árvore, introduzem novos indivíduos na população a partir dos grupos estruturais pertencentes aos seus filhos na hierarquia. Sendo assim, o nicho 1 escolhe de forma aleatória soluções pertencentes aos grupos estruturais vinculados aos nichos 4, 5 e 6, o nicho 2 escolhe soluções oriundas dos grupos pertencentes aos nichos 7, 8 e 9, e o nicho 3 utiliza os grupos ligados aos nichos 10, 11, e 12. Já o nicho 0 insere de forma aleatória novos indivíduos oriundos dos grupos estruturais vinculados a qualquer um dos nove nichos do nível inferior.

Este esquema de acesso limitado aos grupos estruturais foi idealizado na tentativa de preservar em cada um dos nichos características específicas originárias da etapa I de amostragem. A reinicialização proposta é utilizada como forma de diversificar a população, mantendo as propriedades iniciais de cada subpopulação. Nota-se que na seção de discussão de resultados (Seção 6.1), será apresentada uma análise em relação ao limiar de reinicialização p, onde diferentes valores foram testados durante o processo de otimização.

4.4.5 Algoritmo colônia artificial de abelhas

Ao longo dos últimos anos, a área de inteligência de enxame (*swarm intelligence*) tem despertado o interesse de pesquisadores de diversas áreas (KARABOGA; BASTURK, 2008). Esta pode ser definida como qualquer tentativa de desenvolver algoritmos inspirados no comportamento coletivo de colônias sociais de insetos ou outras sociedades de animais (BONABEAU; DORIGO; THERAULAZ, 1999). O termo enxame pode ser empregado, de maneira geral, para se referir a qualquer grupo restrito de agentes que interagem entre si (KARABOGA; BASTURK, 2008). Porém, o comportamento inteligente de enxames provêm do conjunto de ações apresentadas por coletivos de insetos ou animais com o propósito de desempenhar tarefas específicas, sem qualquer tipo de supervisão. Este conjunto de comportamentos expressam populações que através de interações são capazes de se auto-organizar (KARABOGA; BASTURK, 2007). Baseadas na capacidade de auto-organização destas sociedades, diversas abordagens vêm sendo

aplicadas na modelagem de algoritmos voltados à resolução de problemas complexos do mundo real (BONABEAU; DORIGO; THERAULAZ, 1999; KENNEDY et al., 2001). Os algoritmos baseados em inteligência de enxame, assim como os AEs, possuem algumas vantagens, como a fácil escalabilidade, tolerância a falhas, adaptação, modularidade, autonomia e paralelismo (AKAY; KARABOGA, 2012).

Dessa forma, os componentes chave de algoritmos baseados em inteligência de enxames são a auto-organização e divisão do trabalho. Em uma sociedade auto-organizada, cada indivíduo deve responder individualmente a estímulos locais e agir de forma coletiva a fim de realizar tarefas por meio da divisão do trabalho, sem que haja nenhum tipo de supervisão centralizada. É fundamental a capacidade da sociedade conseguir se adaptar a mudanças internas ou externas do ambiente de forma eficiente (AKAY; KARABOGA, 2012). Segundo Bonabeau et al. (1999) existem quatro propriedades básicas inerentes à qualquer sistema auto-organizado: (*i*) reforço positivo (*positive feedback*), que caracteriza a ação de um indivíduo recrutar outros indivíduos por meio de alguma sinalização, como a dança das abelhas (*waggle dance*) que objetiva recrutar outras abelhas à explorarem uma fonte de alimentação específica; (*ii*) reforço negativo (*negative feedback*), propriedade que evita que todos os indivíduos da sociedade se empenhem na realização da mesma tarefa através do balanceamento negativo da atração, como a ação de abandonar fontes de alimentação muito exploradas praticada pelas abelhas; (*iii*) flutuações, que representam comportamentos aleatórios assumidos por alguns indivíduos com o objetivo de explorar novas áreas, como os voos aleatórios das abelhas exploradoras (*scouts*) em um enxame de abelhas; e (*iv*) interações múltiplas, que delineiam as bases das tarefas a serem realizadas por membros da sociedade seguindo algumas regras preestabelecidas.

Neste trabalho, o algoritmo ABC, proposto por Karaboga (2005), foi incorporado ao AM desenvolvido, com o intuito de otimizar, de modo independente, os diferentes nichos estruturados na árvore ternária. O ABC consiste em uma meta-heurística baseada no comportamento de forrageamento de uma colônia de abelhas, a qual foi proposta recentemente, sendo voltada à otimização de funções numéricas multivariáveis. Diversos estudos, desde então, foram publicados demonstrando a sua competitividade quando comparado a outras meta-heurísticas baseadas em população, como os AGs, algoritmos PSO e de evolução diferencial (DE, sigla em inglês) (KARABOGA; BASTURK, 2007; KARABOGA; BASTURK, 2008; KARABOGA; AKAY, 2009). Uma das principais vantagens do ABC, consiste na utilização de poucos parâmetros de controle (GAO; LIU; HUANG, 2012).

Em um enxame real de abelhas, algumas tarefas são executadas por indivíduos especializados, que tem por objetivo maximizar a quantidade de néctar armazenado na colmeia por meio da divisão do trabalho e auto-organização (AKAY; KARABOGA, 2012). A tarefa de forrageamento é de importância crucial para um enxame, e depende basicamente da aptidão de recrutamento dos indivíduos e abandono de fontes de alimentação esgotadas. O ABC simula os três tipos de abelhas existentes responsáveis pela tarefa de encontrar fontes de alimentação (KARABOGA; BASTURK, 2007): (*i*) abelhas empregadas (*employed bees*); (*ii*) abelhas observadoras (*onlooker bees*); e (*iii*) abelhas exploradoras (*scout bees*). Metade da colônia engloba abelhas empregadas e a outra metade abrange abelhas observadoras. As abelhas empregadas são responsáveis por procurar comida em torno das fontes de alimentação que elas previamente conhecem, enquanto compartilham informações acerca da qualidade das fontes de alimentação que estão sendo exploradas com as abelhas observadoras. Por sua vez, as abelhas observadoras aguardam na colmeia e decidem quais fontes de alimentação serão exploradas baseadas nas informações compartilhadas pelas abelhas empregadas. Por fim, as abelhas exploradoras eram abelhas empregadas que decidiram abandonar as suas fontes de alimentação com o objetivo de encontrar novas fontes.

O comportamento inteligente que emerge a partir da cooperação e interação das abelhas, necessário à realização da tarefa de forrageamento, simulado pelo ABC, pode ser resumido em três etapas (AKAY; KARABOGA, 2012):

1. Etapa das abelhas empregadas: As abelhas empregadas exploram as fontes de alimentação que elas previamente conhecem. Após algum tempo, elas retornam à colmeia para armazenarem o néctar coletado, e compartilharem informações sobre a qualidade das suas fontes de alimentação através da performance de uma dança (*waggle dance*), na área designada para isso. A natureza da dança é relativa à qualidade da fonte de alimentação e tem por objetivo recrutar novas abelhas, sendo que boas fontes de alimentação tendem a atrair mais abelhas observadoras;
2. Etapa das abelhas observadoras: As abelhas observadoras aguardam o retorno das abelhas empregadas na colmeia. Após assistirem as danças realizadas pelas abelhas empregadas, elas optam pelas fontes de alimentação que aparentam ser mais rentáveis dependendo do caráter da dança, e assim iniciam a exploração;
3. Etapa das abelhas exploradoras: Quando uma abelha empregada percebe que sua fonte de alimentação esgotou, ela abandona a fonte e torna-se uma abelha exploradora. As abelhas exploradoras procuram de forma aleatória por novas fontes de

alimentação.

A partir da interação entre estas três etapas, as abelhas designadas à realizar o forrageamento de alimentos são capazes de prover o sustento para o enxame. Com isso, define-se que as abelhas dos tipos empregadas e observadoras são responsáveis pela tarefa de refinar as fontes de alimentação já conhecidas, enquanto que as abelhas exploradoras devem explorar o ambiente em busca de novos locais de exploração.

No algoritmo ABC, a exploração e o refinamento de soluções são mecanismos extremamente importantes, assim como nos AMs. Contudo, o algoritmo apresenta algumas ineficiências, como a característica de ser bom na exploração de soluções, mas nem tanto no refinamento (AKAY; KARABOGA, 2012). Este fato acaba por tornar a convergência do algoritmo um pouco mais lenta, podendo ser um problema em alguns casos. O processo de exploração está relacionado à habilidade do método de procurar de forma independente os ótimos globais da função objetivo, já o refinamento representa a capacidade da aplicação de conhecimento existente, proveniente das soluções da população, para procurar melhores soluções em torno destas (LI; NIU; XIAO, 2012). Com o objetivo de encontrar um balanço entre estes dois processos e acelerar a otimização de soluções, algumas variações do ABC estão sendo propostas. Estas versões modificadas vêm obtendo melhores desempenhos do que o ABC original (LI; NIU; XIAO, 2012).

Portanto, neste trabalho será utilizada a combinação de duas abordagens propostas para o ABC. A primeira abordagem consiste no trabalho de Akay e Karaboga (2012), que propuseram modificações nos componentes que controlam a frequência de mutação de variáveis em uma solução, e uma análise acerca da parametrização mais adequada a ser utilizada na etapa de exploração do método. A segunda proposta consiste no trabalho de Zhu e Kwong (2010), que propuseram o algoritmo *gbest-guided ABC* (GABC). Basicamente, este algoritmo considera a informação da melhor solução da população na equação de geração de novas soluções, objetivando melhorar os processos de refinamento. Observa-se que os autores de ambos os trabalhos, concluíram que o ABC pode ser considerado um bom algoritmo em termos de otimizações globais e locais devido aos esquemas de alternância entre as especializações dos diferentes tipos de abelhas simulados. Por este motivo e por se tratar de uma meta-heurística relativamente nova, o algoritmo ABC foi escolhido para ser utilizado na otimização de cada subpopulação do AM, visando não só o refinamento das soluções de cada nicho mas também a exploração global do espaço de busca, corroborando com o propósito dos AMs.

4.4.5.1 Implementação do algoritmo colônia artificial de abelhas

No algoritmo ABC, cada fonte de alimentação representa uma solução para o problema em consideração, e a qualidade desta fonte de alimentação é expressa pelo seu valor de aptidão. Modelando em termos do problema PSP, cada fonte de alimentação representa uma diferente solução conformacional, sendo que a sua qualidade é definida pelo valor de energia. Cada fonte de alimentação é explorada por apenas uma abelha empregada, ou seja, o número de abelhas empregadas corresponde ao mesmo número de fontes de alimentação existentes em torno da colmeia (número de soluções da população). O número de abelhas observadoras no enxame equivale ao número de abelhas empregadas. Sendo SN o número de fontes de alimentação consideradas (soluções da população), ae o número de abelhas empregadas existentes no enxame, e ao o número de abelhas observadoras, convenciona-se que $SN = ae = ao$.

Conforme descrito anteriormente, o ABC simula as três etapas do processo de forrageamento das abelhas. Modelando em termos computacionais têm-se:

1. Cada solução da população é vista como uma fonte de alimentação, a qual é atualizada nesta etapa através de um processo de mutação. Cada abelha empregada está vinculada a somente uma fonte de alimentação, sendo que não existem soluções sem abelhas empregadas;
2. ao soluções da população são selecionadas de forma aleatória, simulando o comportamento das abelhas observadoras, e o mesmo processo de atualização da etapa anterior é aplicado nas soluções selecionadas. Neste trabalho, foi utilizada a seleção por *ranking*;
3. A solução mais inativa da população é descartada, e uma nova solução é gerada. Entende-se por solução mais inativa, aquela que não sofre melhoramentos em relação à função objetivo a um certo número de gerações.

A operação de atualização aplicada nas soluções das etapas 1 e 2 descritas acima, pode ser vista como a geração de uma nova fonte de alimentação em torno de uma fonte já conhecida. Sendo assim, a produção de uma nova solução $\upsilon_i = [\upsilon_{i1}, \upsilon_{i2}, ..., \upsilon_{iD}]$ a partir da i-ésima solução existente $X_i = [x_{i1}, x_{i2}, ..., x_{iD}]$, sendo que $X_i = \upsilon_i$, é dada pela expressão (Eq. 4.5):

$$\upsilon_{ij} = x_{ij} + \delta_{ij}(x_{ij} - x_{kj}) + \gamma_{ij}(y_j - x_{ij}) \tag{4.5}$$

Onde $i = 1, ..., SN$, $j = 1, ..., D$. SN denota o número de soluções na população e D representa o número de variáveis de otimização em cada solução. x_{ij} representa a j-ésima variável da solução X_i, v_{ij} representa o novo valor que essa varável irá assumir, x_{kj} representa a j-ésima variável da k-ésima solução da população ($k = 1, ..., SN$), selecionada de forma aleatória entre todas as soluções da população, e δ_{ij} é um número randômico no intervalo uniforme $[-1, +1]$. O último termo da equação integra ao cálculo a melhor solução da população. Sendo que y_j é a j-ésima variável da melhor solução da população, e γ_{ij} é um número randômico no intervalo uniforme $[0, +1, 5]$. O último termo, proposto por Zhu e Kwong (2010), tem o objetivo de guiar a nova solução em direção ao melhor indivíduo da população, aumentando o potencial de refinamento.

Ainda, cada variável j da solução X_i é atualizada conforme o parâmetro de controle MR. Neste trabalho, será utilizado $MR = 0, 4$, definido através do trabalho de Akay e Karaboga (2012). Portanto, uma variável será atualizada com probabilidade de 40%, se $NumeroAleatorio(0, 1) < 0, 4$, onde a função $NumeroAleatorio$ gera um número real entre 0 e 1. O processo de atualização termina com a escolha gulosa entre a nova solução gerada v_i e a anterior X_i, sendo que a melhor solução (menor energia) permanece na população.

Conforme a representação que será utilizada neste trabalho (Seção 4.4.2) e voltando-se para o problema PSP, cada variável representa um aa específico da proteína-alvo que contém no máximo 7 ângulos de torção, por exemplo, a variável $x_{ij} = (\phi_{ij}, \psi_{ij}, \omega_{ij}, \chi_{ij_{(0...4)}})$. Com isso, todos os ângulos pertencentes a mesma variável são atualizados da mesma forma, com exceção do ângulo ω que não sofre mutação.

Observa-se que foram realizadas duas modificações no funcionamento do algoritmo ABC, objetivando uma melhor adaptação a características do problema PSP. A primeira modificação consiste em uma verificação dos novos valores gerados pela operação de atualização. A cada mutação de um ângulo (ϕ ou ψ) da variável v_{ij}, é verificado se o novo valor gerado está contido na APL-1 que determina as preferências conformacionais do aa relacionado à v_{ij}, com o objetivo de restringir valores angulares desfavoráveis ou fora do intervalo de números reais $[-180, +180]$. Caso o novo valor assumido não esteja contido na APL-1, o mesmo não é considerado e o antigo valor permanece. Para os valores dos ângulos χ, é verificado se eles estão na faixa de valores reais que compreende o intervalo $[-180, +180]$.

Sabe-se que as estruturas irregulares de proteínas apresentam alto grau de flexibilidade devido à maior exposição ao solvente, e são consideradas as regiões mais difíceis

de serem preditas da estrutura 3-D de uma proteína (KOKKINIDIS; GLYKOS; FADOULOGLOU, 2012). Com isso, no AM apresentado por Corrêa et al. (2016), os esforços de refinamento local, realizados pela meta-heurística de arrefecimento simulado (*simulated annealing*), foram despendidos apenas em regiões irregulares da proteína, excluindo as regiões mais estáveis. Seguindo esta mesma ideia, a segunda modificação proposta no ABC, consiste em aplicar a atualização de soluções através do processo de mutação descrito acima, apenas nas variáveis que estiverem relacionadas aos aminoácidos com ESs irregulares (voltas e alças). As variáveis relacionadas aos aminoácidos com ESs regulares (hélices e folhas) são ignoradas no processo. Esta verificação é possível, pois o AM recebe como parâmetro de entrada, além da EP, a ES da proteína-alvo.

O pseudocódigo do Algoritmo 5 descreve o método ABC implementado neste trabalho, sendo que este representa a função $ABC(nicho)$ utilizada no AM para a otimização dos nichos (Alg. 4, linha 5). O ABC recebe como parâmetro de entrada a população do nicho a ser otimizado e executa 10 gerações a cada chamada de função do AM.

Em síntese, na primeira etapa (etapa das abelhas empregadas) (Alg. 5, linha 2), todas as soluções da população são atualizadas de acordo com a Equação 4.5 (Alg. 5, linha 6). Na segunda etapa (etapa das abelhas observadoras) (Alg. 5, linha 16), as soluções são selecionadas de forma aleatória considerando a probabilidade de cada uma ser escolhida com base no seu valor de aptidão. Nesta etapa, a escolha de uma solução é realizada através da seleção por *ranking*, explicada anteriormente (Alg. 5, linha 18). As soluções selecionadas são atualizadas seguindo o mesmo procedimento da etapa anterior (Eq. 4.5) (Alg. 5, linha 21). Nota-se que a única diferença entre as etapas das abelhas empregadas e observadoras é que na primeira etapa todas as soluções da população são atualizadas, e na segunda, apenas as soluções selecionadas são atualizadas, sendo que as melhores possuem mais chances.

Com o intuito de diversificar a população de indivíduos e aumentar o caráter exploratório do ABC, visto que as operações de mutação das etapas 1 e 2 são restritas as regiões irregulares da proteína-alvo, inseriu-se entre a etapa 1 e a etapa 2, uma operação de cruzamento entre duas soluções da população (Alg. 5, linha 11). Estas soluções são selecionadas através da seleção por *ranking* (Alg. 5, linhas 12 e 13), e a operação de cruzamento é realizada através do operador de cruzamento de ES (Alg. 5, linha 14), explicado na Seção 4.4.3. A solução resultante compete com os pais para ser inserida na população, ou seja, o pior pai é substituído, caso a solução filho seja melhor (Alg. 5, linha 15).

Algoritmo 5: Pseudocódigo do algoritmo ABC

Entrada: pop : conjunto de soluções do nicho a ser otimizado
Saída: pop : conjunto de soluções otimizadas

$MR \leftarrow 0,4$
para $g_{ABC} \leftarrow 1 : 10$ **faça**
  `// Etapa das abelhas empregadas`
  para cada $pop_i, i \leftarrow 1 : ae$ **faça**
    para cada $pop_{ij}, j \leftarrow 1 : D$ **faça**
      se $(NumeroAleatorio(0,1) < MR)$ *e* (pop_{ij} *representar um aa com ES irregular*) **então**
        gera um novo valor para a variável pop_{ij} *aplicando a Equação (4.5)*
        verifica se o novo valor é permitido
        `// Este procedimento é realizado para todos os ângulos da variável` pop_{ij}
      fim
    fim
  fim
  $OrdenaPopulacao(pop)$
  `// Operação de cruzamento realizada entre duas soluções da população`
  $pai_1 \leftarrow$ *seleção por ranking de uma solução da* pop
  $pai_2 \leftarrow$ *seleção por ranking de uma solução da* pop
  $filho \leftarrow cruzamentoES(pai_1, pai_2)$
  substitui o pior pai *pela solução* $filho$
  $OrdenaPopulacao(pop)$
  `// Etapa das abelhas observadoras`
  para $i \leftarrow 1 : ao$ **faça**
    $sol \leftarrow$ *seleção por ranking de uma solução da* pop
    para cada $sol_j, j \leftarrow 1 : D$ **faça**
      se $(NumeroAleatorio(0,1) < MR)$ *e* (sol_j *representar um aa com ES irregular*) **então**
        gera um novo valor para a variável sol_j *aplicando a Equação (4.5)*
        verifica se o novo valor é permitido
        `// Este procedimento é realizado para todos os ângulos da variável` sol_j
      fim
      insere a solução sol *na população se for melhor que a solução original*
    fim
  fim
  `// Etapa das abelhas exploradoras`
  $l \leftarrow 200$
  para cada $pop_i, i \leftarrow 1 : SN$ **faça**
    verifica se a solução pop_i *não recebeu melhoramentos a mais de* l *gerações*
    descarta pop_i
    insere nova solução na população
  fim
fim
$OrdenaPopulacao(pop)$

Na terceira e última etapa do algoritmo (etapa das abelhas exploradoras) (Alg. 5, linha 26), a solução que estiver sem sofrer melhorias a um certo número de gerações l é descartada (Alg. 5, linha 30), e uma nova solução é inserida na população (Alg. 5, linha 31). Sendo l o limiar de determinação para o descarte (Alg. 5, linha 29). Neste trabalho, será utilizado $l = 200$, definido através do trabalho de Akay e Karaboga (2012). Observa-se que a inserção de novas soluções depende do nicho que está sendo otimizado, e segue o mesmo procedimento explicado na seção de reinicialização (Seção 4.4.4). Por fim, após a execução de 10 gerações do algoritmo ABC para a população de um determinado nicho, o algoritmo retorna os indivíduos otimizados.

4.5 Resumo do capítulo

Neste capítulo foram apresentados os algoritmos e estratégias utilizados na elaboração desta dissertação, bem como a metologia e estruturação do método proposto.

Com isso, foi desenvolvido um AM multi populacional para lidar com o problema PSP. Este algoritmo pode ser dividido em duas etapas principais. A primeira etapa, denominada amostragem e inicialização de indivíduos, foi estruturada visando a geração e classificação de diversos modelos estruturais para determinada proteína-alvo, a partir da estratégia APL, buscando a definição de diferentes grupos estruturais e a criação de melhores estruturas a serem incorporadas à meta-heurística como soluções iniciais das multi populações de otimização. A APL foi utilizada como forma de inicialização de indivíduos, visando a redução da complexidade do espaço de busca, por meio da exploração do conhecimento prévio acerca de estruturas 3-D de proteínas determinadas experimentalmente e armazenadas no PDB. Esta etapa engloba os seguintes processos de execução: (*i*) amostragem de indivíduos inicializados a partir da APL-combinada; (*ii*) filtragem das soluções geradas através de limiares máximos de RG e SASA; e (*iii*) agrupamento das soluções resultantes, objetivando destacar os diferentes grupos estruturais gerados durante o processo de amostragem, utilizados na inicialização das multi populações do algoritmo de otimização da etapa II.

Dessa forma, a segunda etapa consiste na otimização das soluções oriundas da etapa I de amostragem de indivíduos, realizada pelo AM multi populacional desenvolvido. O cerne desta meta-heurística fundamenta-se na organização da população de indivíduos em uma estrutura hierárquica em árvore. Tal modelo de organização populacional favorece o gerenciamento da performance do processo de exploração do algoritmo de busca

sobre a superfície energética multimodal do problema, visto que cada nodo da árvore pode ser caracterizado como uma subpopulação independente. O algoritmo ainda engloba operadores genéticos modificados, como o operador de cruzamento de ES e a mutação baseada em regiões mais flexíveis da proteína, cujo objetivo consiste no aumento da capacidade de exploração do método voltando-se a características do PSP. Implementou-se também o algoritmo ABC, o qual foi incorporado ao AM com o propósito de ser utilizado como uma técnica de busca local, a ser aplicada em cada nodo da árvore. Este algoritmo recebeu algumas modificações em seu funcionamento, visando a melhor adaptação ao problema.

Sendo assim, o próximo capítulo apresentará os resultados obtidos e as análises realizadas a partir da etapa I de amostragem e classificação de indivíduos, descrita neste capítulo.

5 ANÁLISES E RESULTADOS - ETAPA I

Este capítulo tem por objetivo apresentar os resultados obtidos a partir da estratégia de amostragem e classificação de indivíduos, proposta como forma de prover soluções diversificadas e de qualidade à meta-heurística de otimização de estruturas 3-D de proteínas, assim como discorrer algumas análises concernentes a estes resultados, visando embasar as decisões tomadas na estruturação algorítmica da etapa I do método de otimização (Seção 4.3), conforme descrito anteriormente.

A etapa I do método de otimização foi idealizada com o intuito de auxiliar o AM proposto, responsável pela otimização dos modelos estruturais para determinada proteína-alvo, através dos processos de geração e classificação de diversos modelos estruturais de indivíduos, a partir da APL, buscando a definição de diferentes grupos estruturais e a criação de melhores estruturas para serem incorporadas à meta-heurística como soluções iniciais das multi populações de otimização.

Dessa forma, os resultados obtidos e as análises realizadas referentes a estes, envolvendo os diferentes passos de execução que constituem a etapa I, serão apresentados abaixo. Para cada seção de experimentos realizados, foi utilizado um conjunto de testes envolvendo 8 proteínas-alvo (Tab. 5.1). As diferentes sequências de aminoácidos foram obtidas a partir do PDB e utilizadas como estudos de caso relacionados à etapa I da abordagem desenvolvida. As proteínas foram selecionadas visando garantir que o conjunto de testes englobe sequências de aminoácidos de diferentes tamanhos e com variadas topologias estruturais. O detalhamento de cada proteína-alvo, como tamanho e composição de ES, pode ser observado na Tabela 5.1. Observa-se que a proteína-alvo T0820 foi extraída do conjunto de proteínas-alvo apresentado nos experimentos do CASP11[1], a qual foi referida pelo ID T0820-D1 e alocada na categoria FM (KINCH et al., 2016b).

Destaca-se aqui que as comparações de qualidade estrutural que foram realizadas entre os indivíduos gerados no processo de amostragem e as estruturas experimentais relativas às proteínas-alvo do conjunto de testes, deram-se através da aplicação da métrica RMSD (ZHANG; SKOLNICK, 2004). O RMSD é amplamente utilizado para avaliar o grau de semelhança entre duas estruturas 3-D de proteínas, e caracteriza uma função de minimização (RMSD igual a zero indica que as estruturas são idênticas). Esta métrica pode ser expressa através da Equação 5.1. Nota-se que nos cálculos de RMSD entre os modelos estruturais gerados e as estruturas determinadas experimentalmente foram

[1] <www.predictioncenter.org/casp11/targetlist.cgi>

Tabela 5.1: Conjunto de testes de proteínas-alvo empregado nos experimentos e análises relativos à primeira etapa do método proposto

ID-Proteína	Tamanho	Composição de ES
1AB1	46	1 folha/2 hélices
1ACW	29	1 folha/1 hélice
1AIL	70	3 hélices
1DFN	30	1 folha tripla
2MR9	44	3 hélices
2P5K	64	1 folha/3 hélices
3V1A	48	2 hélices
T0820-CASP11	90	3 hélices

Fonte: Do autor (2017).

considerados apenas os átomos de C_α do esqueleto polipeptídico das estruturas 3-D.

$$RMSD(a,b) = \sqrt{\frac{\sum_{i=1}^{n} \|r_{ai} - r_{bi}\|^2}{n}} \tag{5.1}$$

Onde a e b representam as duas estruturas a serem comparadas, n representa o tamanho da sequência de aminácidos, r_{ai} e r_{bi} são vetores que descrevem as posições Cartesianas do mesmo átomo i em cada uma das duas estruturas, a e b respectivamente. Considera-se no cálculo que a e b foram previamente sobrepostas de forma ótima.

5.1 Análise do potencial de inicialização das APLs, RG e SASA

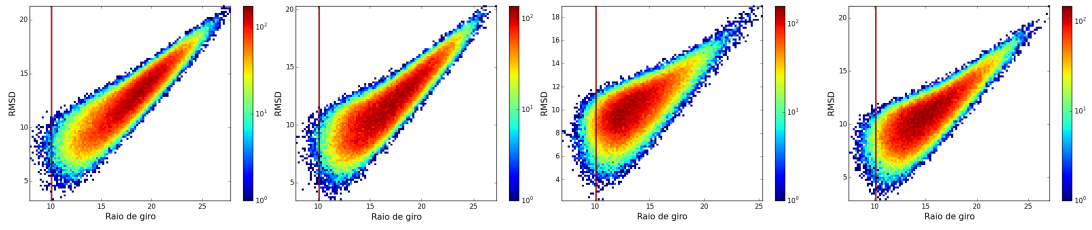
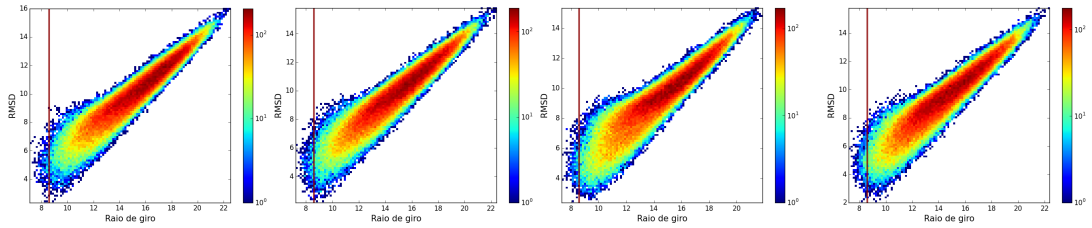
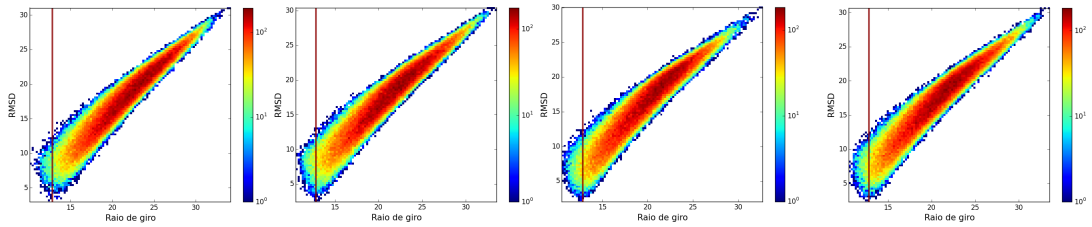
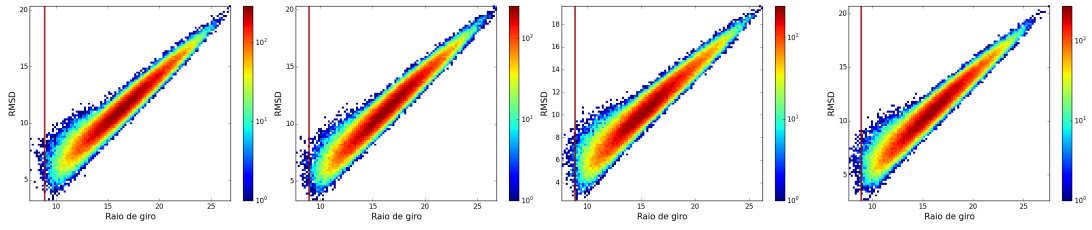
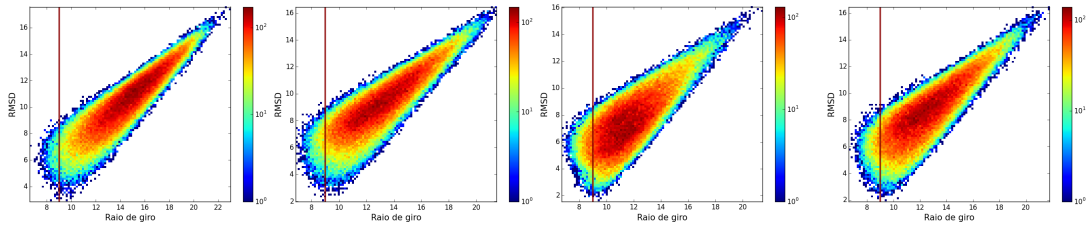
Com a finalidade de avaliar os diferentes tipos de APLs disponíveis e visualizar o potencial de cada um quando da inicialização de modelos estruturais para determinadas proteínas-alvo, e ainda analisar o comportamento dos indicadores de empacotamento estrutural (RG e SASA) empregados neste trabalho, realizou-se amostragens de indivíduos para cada proteína do conjunto de testes definido.

Sendo assim, para cada proteína-alvo foram gerados 100.000 indivíduos a partir de cada um dos tipos de APLs descritos anteriormente. As APLs consideradas nestes experimentos foram: (*i*) APL-1; (*ii*) APL-vizinhança, considerando os seus subtipos (APL-3, APL-2e, APL-2d e APL-1) e as probabilidades de cada um ser aplicado; (*iii*) APL-centroide, também considerando os seus subtipos (APL-9, APL-7 e APL-5) e as probabilidades de cada um ser aplicado; e (*iv*) APL-combinada, onde a APL-vizinhança e a APL-centroide foram utilizadas de forma conjunta. Nota-se que o detalhamento dos

diferentes tipos de APLs abordados neste trabalho estão descritos na seção de preferências conformacionais de aminoácidos (Seção 4.1), sendo que o modo como cada APL e subtipos são utilizados na inicialização de sequências de aminoácidos, assim como as probabilidades concernentes à aplicação de cada um deles, estão descritos na seção de inicialização de indivíduos (Seção 4.3.1). Destaca-se também que na abordagem de geração de indivíduos utilizada neste trabalho, são geradas 10.000 soluções. Contudo, nesta seção, a amostragem de indivíduos foi maior na tentativa de compreender amostras mais representativas. Os indivíduos gerados em cada processo de amostragem foram comparados de acordo com as métricas RMSD em relação às estruturas determinadas experimentalmente das proteínas-alvo, RG e SASA.

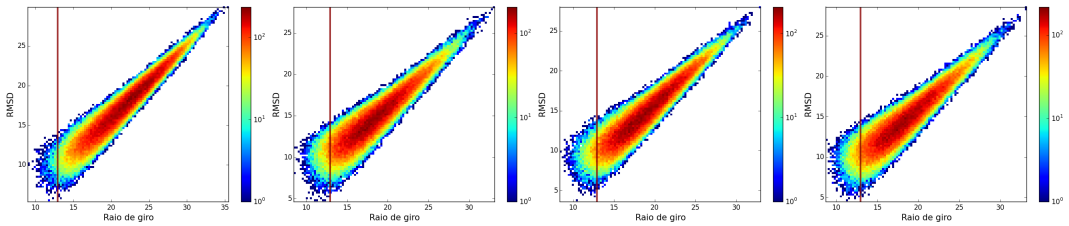
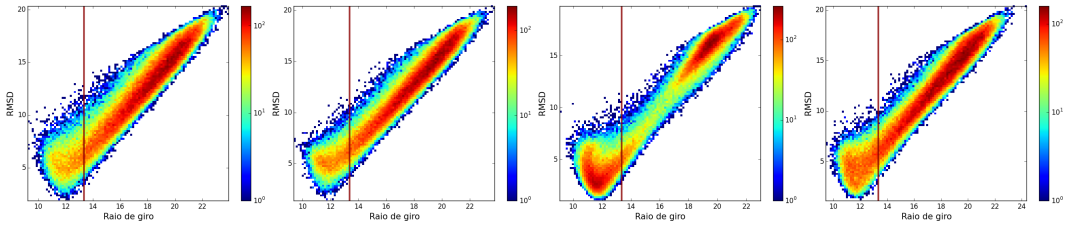
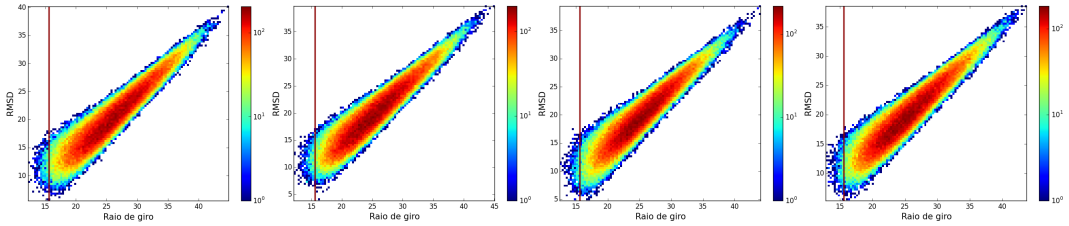
A fim de analisar a qualidade dos indivíduos gerados por cada tipo de APL em relação ao conjunto de proteínas-alvo, estes foram dispostos em dois tipos de gráficos. O primeiro relaciona os valores de RMSD dos indivíduos com os valores de RG, e o segundo associa os valores de RMSD com os valores de SASA de cada indivíduo. Ambos os gráficos buscam encontrar a correlação entre estado de empacotamento das estruturas geradas e qualidade de inicialização expressa pelo RMSD. Com isso, a Figura 5.2 ilustra a disposição dos 100.000 indivíduos gerados para o conjunto de testes de proteínas-alvo, a partir de cada tipo de APL disponível, relacionando RMSD e RG. Da mesma forma, a Figura 5.3 ilustra a relação entre RMSD e SASA dos 100.000 indivíduos criados para cada proteína-alvo, através dos diferentes tipos de combinações da APL. Nota-se que a linha vermelha em cada figura representa o valor de RG e SASA para as estruturas determinadas experimentalmente.

Figura 5.1: Análise do conjunto de testes de proteínas para amostragens de 100.000 estruturas geradas através de diferentes APLs (APL-1, APL-vizinhança, APL-centroide e APL-combinada), relacionando RMSD e RG. A linha vermelha na vertical indica o valor de RG referente à proteína determinada experimentalmente

APL-1	APL-vizinhança	APL-centroide	APL-combinada
			
1AB1			
			
1ACW			
			
1AIL			
			
1DFN			
			
2MR9			

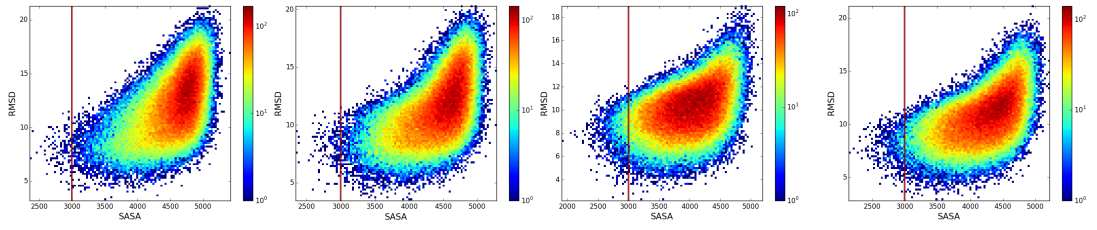
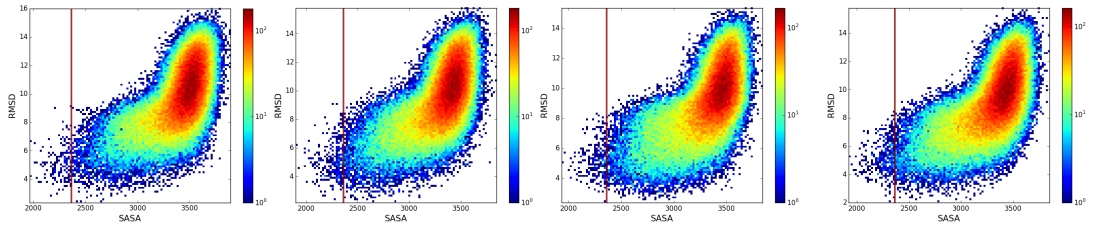
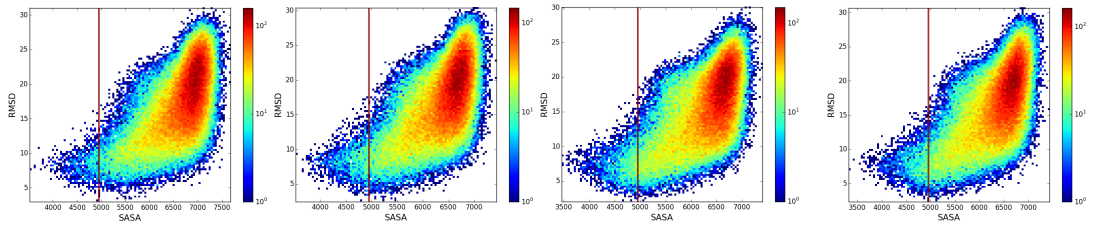
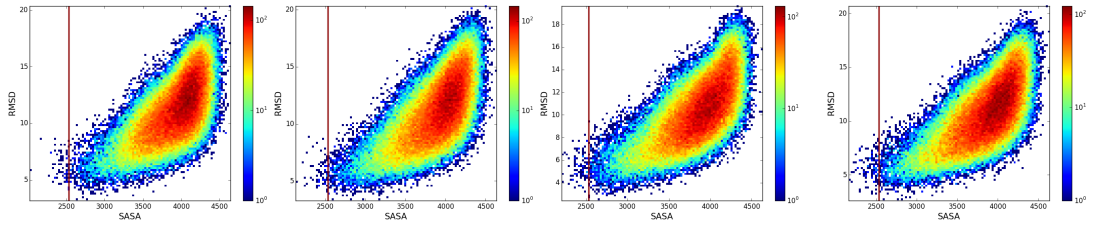
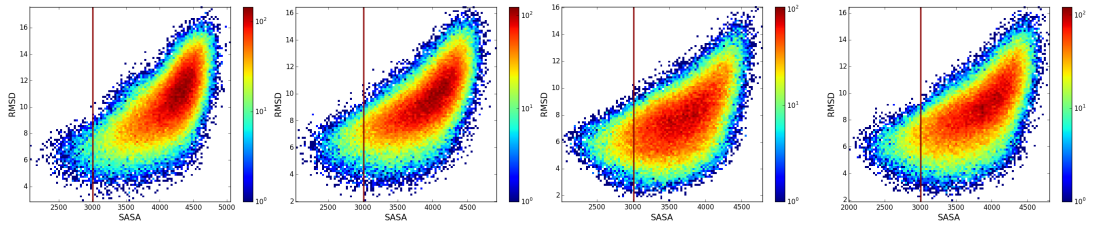
Continuação na próxima página

Figura 5.2: Continuação da tabela da página anterior

APL-1	APL-vizinhaça	APL-centroide	APL-combinada
			
2P5K			
			
3V1A			
			
T0820			

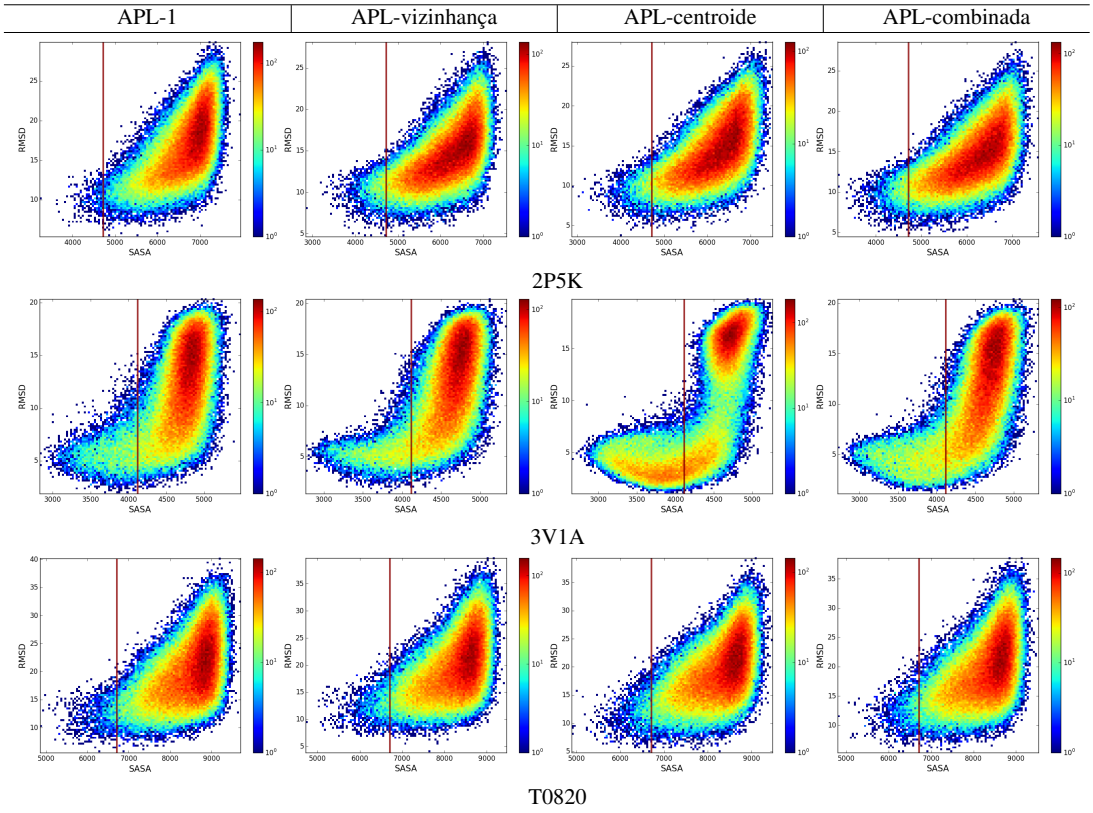
Fonte: do autor (2017).

Figura 5.3: Análise do conjunto de testes de proteínas para amostragens de 100.000 estruturas geradas através de diferentes APLs (APL-1, APL-vizinhança, APL-centroide e APL-combinada), relacionando RMSD e SASA. A linha vermelha na vertical indica o valor de SASA referente à proteína determinada experimentalmente

APL-1	APL-vizinhança	APL-9	APL-centroide
			
1AB1			
			
1ACW			
			
1AIL			
			
1DFN			
			
2MR9			

Continuação na próxima página

Figura 5.3: Continuação da tabela da página anterior



Fonte: do autor (2017).

Analisando a relação entre RMSD e RG em todos os gráficos apresentados na Figura 5.2, nota-se que, até certo ponto, valores mais baixos de RG tendem a gerar estruturas com RMSD melhores. Porém, quando estes valores aproximam-se dos valores relacionados às estruturas experimentais, o RG tende a não refletir de forma clara a qualidade dos indivíduos. Isto deve-se ao fato de que depois de um certo nível de empacotamento, as estruturas podem assumir inúmeras conformações, mantendo o mesmo grau de empacotamento. No entanto, verifica-se que a métrica de RG pode ser utilizada como um bom indicador de empacotamento e limitador de estruturas ruins nesta fase inicial de geração de indivíduos, visto que valores mais elevados desta métrica implicaram em soluções menos favoráveis, para todas as proteínas-alvo analisadas.

Quanto à relação entre RMSD e SASA, ilustrada nos gráficos da Figura 5.3, percebe-se que também há uma tendência de que soluções melhores sejam representadas por valores de SASA próximos ou menores do que os valores das estruturas experimentais. Contudo, nota-se que nesta relação, os indivíduos apresentaram uma variação maior de qualidade, independente do SASA assumido, posto que ocorreram estruturas com valores de SASA bastante elevados que assumiram RMSD baixos, o que não é per-

cebido na relação entre RMSD e RG. Neste ponto, a preferência por valores mais baixos de SASA pode ser explicada através da varição de RMSD ocorrida nas amostragens, onde as estruturas com SASA mais baixos apresentaram intervalos menores de RMSD. A título de ilustração desta explicação, no caso da proteína-alvo 1ACW, o valor máximo de RMSD assumido por algumas soluções localizadas nas regiões de SASA mais baixos, próximas do SASA atribuído à estrutura experimental da 1ACW, figurou em torno de 10Å, enquanto que os maiores valores de RMSD assumidos por indivíduos localizados em regiões de SASA mais elevado, ficaram em torno de 16Å.

Ainda, analisando ambas as figuras (Fig. 5.2 e 5.3), é possível perceber que todos os tipos de APLs conseguiram gerar indivíduos com valores de RG e SASA iguais e similares aos valores calculados para as estruturas experimentais. No entanto, sabendo que as áreas em vermelho denotam maior concentração de soluções, observa-se que as APLs vizinhança, centroide e combinada tenderam a concentrar mais indivíduos próximos aos valores das estruturas experimentais do que a APL-1, consequentemente gerando um número maior de soluções melhores. Esta observação pode ser percebida de uma forma mais clara nas amostragens relacionadas às proteínas-alvo 1AB1, 2MR9 e 2P5K, na Figura 5.2. Especificamente, no caso da 3V1A, a APL-centroide conseguiu concentrar um número maior de soluções de melhor qualidade quando comparada às outras APLs. Esta tendência também é evidenciada na Figura 5.3 para os processos envolvendo as mesmas proteínas, 1AB1, 2MR9 e 2P5K. E novamente, em relação a proteína 3V1A, a APL-centroide conseguiu concentrar mais soluções nas regiões de menor RMSD.

Sendo assim, a partir da análise da Figura 5.2, envolvendo todas as proteínas do conjunto de testes, conclui-se que a geração de indivíduos com valores de RG mais baixos, ou seja, estruturas mais empacotadas, são preferíveis aos valores mais elevados, pois fica evidente a correlação entre estruturas com RG menores e RMSD melhores. Tratando-se do SASA, através dos gráficos ilustrados na Figura 5.3, percebe-se que esta relação é menos evidente, visto que estruturas com SASA elevado também apresentaram valores de RMSD baixos. Contudo, nota-se que a preferência por valores menores de SASA ainda válida, devido ao fato de que a variação de qualidade das estruturas que assumiram valores mais elevados foi bem maior.

Em relação à qualidade das estruturas geradas por cada um dos tipos da APL, percebe-se que todos eles tenderam a gerar soluções de qualidade similares. Quanto à concentração de indivíduos em regiões de RMSD mais baixos, com exceção da APL-1, os outros três tipos, de maneira geral, produziram concentrações de soluções bem similares.

Dessa forma, constata-se que visualmente, analisando ambas as figuras (Fig. 5.2 e 5.3) e englobando todo o conjunto de testes, não percebe-se superioridade de nenhum tipo de APL quanto à qualidade de estruturas geradas.

Apesar disso, sabe-se que a APL-combinada detêm um conjunto maior de dados experimentais, visto que ela é o resultado da combinação das APLs vizinhança e centroide. Sendo assim, dependendo da proteína-alvo a ser otimizada, esta gama maior de informações pode se tornar interessante, pois cada sequência de aminoácidos apresenta propriedades particulares, e uma maior disponibilidade de dados experimentais proporciona explorações melhores destas características. Por esta razão, optou-se pela utilização da APL-combinada na etapa de amostragem e classificação de indivíduos do método de otimização proposto.

5.2 Análise de funções de energia

Esta seção tem por objetivo analisar o comportamento da função de energia composta que será utilizada neste trabalho, e da função de energia *Talaris2014*, a qual representa atualmente a função padrão do Rosetta para avaliação de modelos estruturais *all-atom*. Ambas as funções de avaliação foram descritas na Seção 2.6.

A função de energia composta que será empregada como função objetivo do método de otimização proposto, consiste na soma de três termos específicos. O primeiro termo representa a função *Talaris2014* do Rosetta, o segundo termo é o valor de SASA total calculado para a estrutura em avaliação, e o terceiro consiste em um termo de reforço de ES, que objetiva auxiliar no processo de formação correta de ESs de proteínas.

Objetivando harmonizar as escalas entre os diferentes termos da função de energia composta e analisar o comportamento resultante desta operação, inseriu-se uma terceira variação de função de avaliação nesta seção de análise, que consiste na normalização de cada um dos três termos da função composta. A normalização dos termos de energia representa a equiparação das diferentes escalas numéricas apresentadas pelos mesmos, a qual objetiva evitar que algum dos termos se sobreponha em relação aos outros, fazendo com que todos exerçam a mesma relevância no cálculo. Neste caso, esta análise se mostrou importante, devido ao fato de que a função composta é o resultado da agregação de três termos independentes que não foram desenvolvidos em conjunto. Embora esta função de energia tenha sido previamente aplicada no trabalho de Corrêa et al. (2016), nenhuma análise concernente ao comportamento da função foi realizada. Sendo assim, cada um

dos termos da função de avaliação foi normalizado segundo a amplitude do valor máximo assumido por este. A normalização para um dos termos de energia pode ser expressa através da Equação 5.2. Nota-se que os valores mínimos dos termos da função objetivo não foram incluídos na normalização devido ao elevado custo computacional atrelado ao recálculo do valor de energia (*fitness*) de cada um dos indivíduos da população a cada variação destes valores. Como trata-se de uma função de minimização, os valores mínimos dos termos tendem a variar frequentemente, implicando em muitos recálculos de *fitness*.

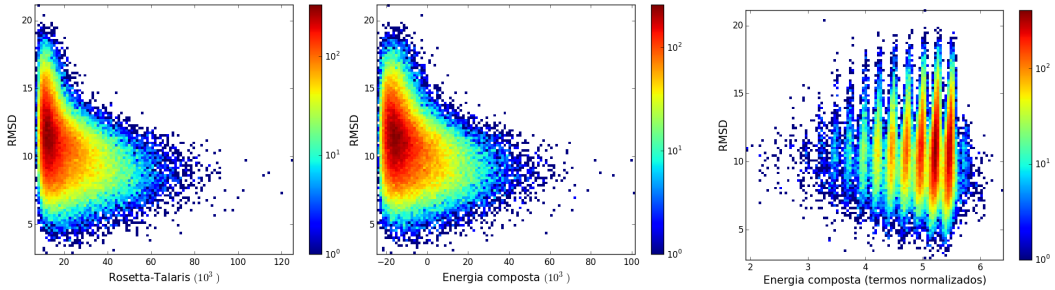
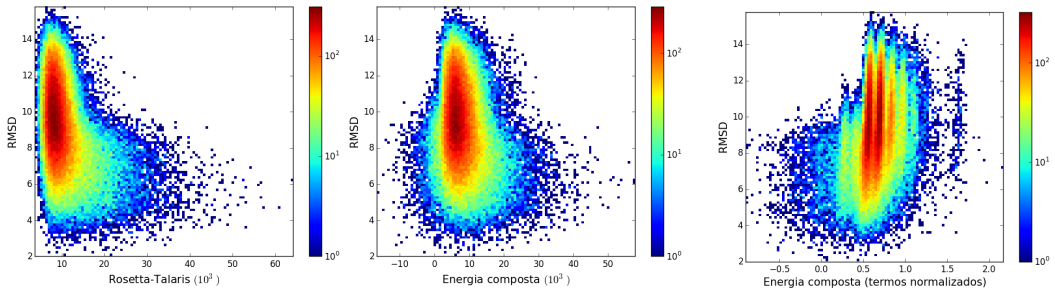
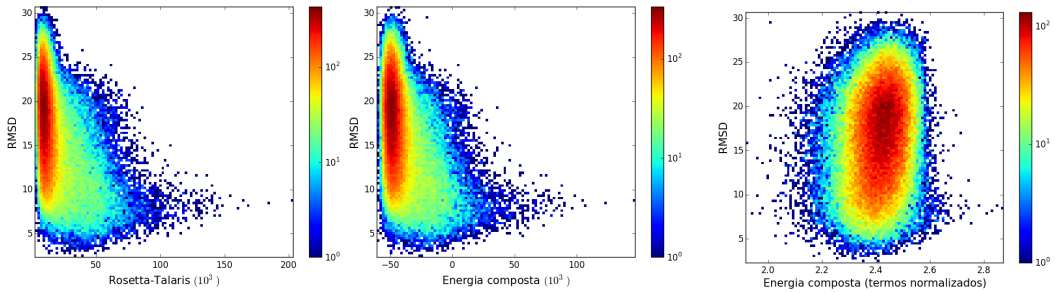
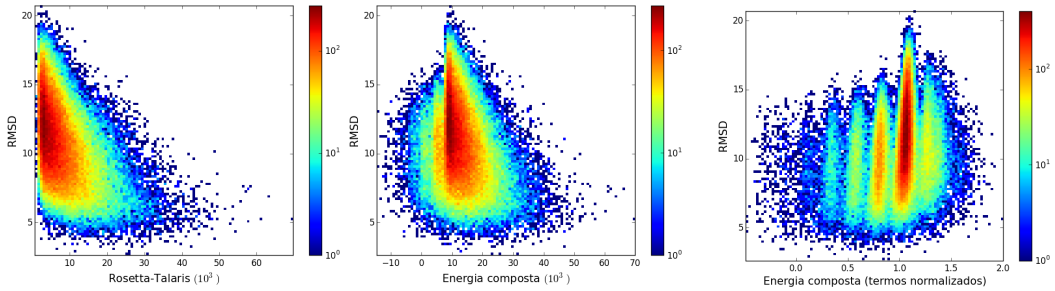
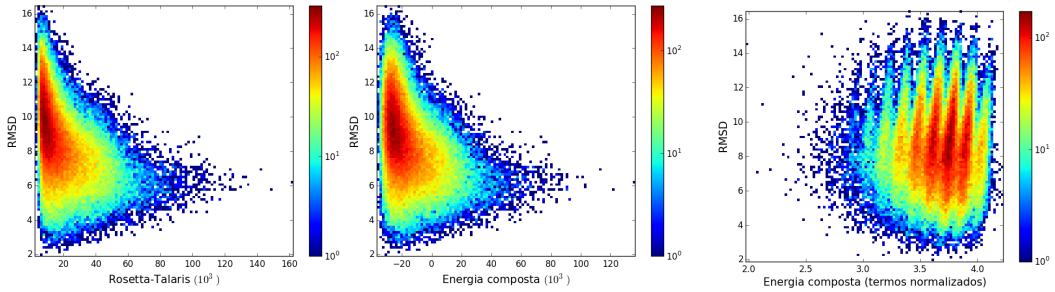
$$x_{norm} = \frac{x}{|x_{max}|} \tag{5.2}$$

Onde x representa um dos termos da função de energia composta e $|x_{max}|$ é o módulo do maior valor permitido ou conhecido para x.

Dessa forma, seguindo na mesma linha de análise da seção anterior, foram gerados 100.000 indivíduos para cada proteína-alvo do conjunto de testes (Tab. 5.1), a partir da APL-combinada, avaliados através destas três variações de funções de energia. Objetivando analisar o comportamento de cada uma destas variações, as soluções avaliadas foram dispostas em gráficos que relacionam os valores de energia calculados e o RMSD assumido por cada indivíduo em relação às estruturas determinadas experimentalmente. Neste sentido, os gráficos visam mostrar a distribuição das estruturas comparando energia e qualidade estrutural, expressa pelo RMSD.

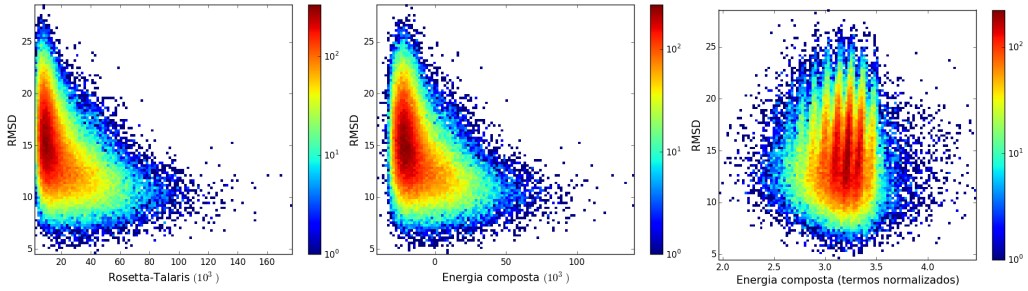
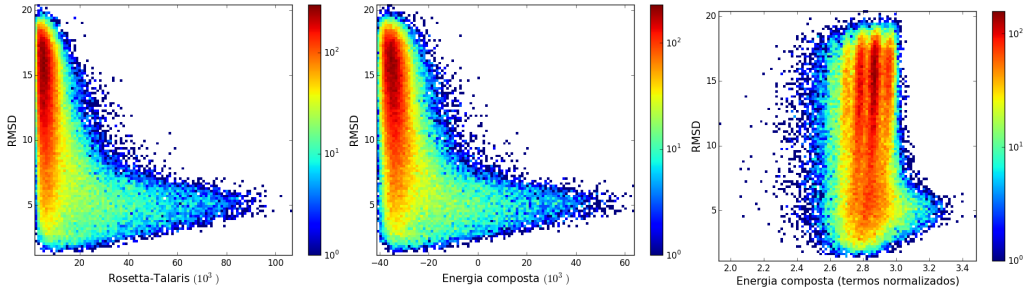
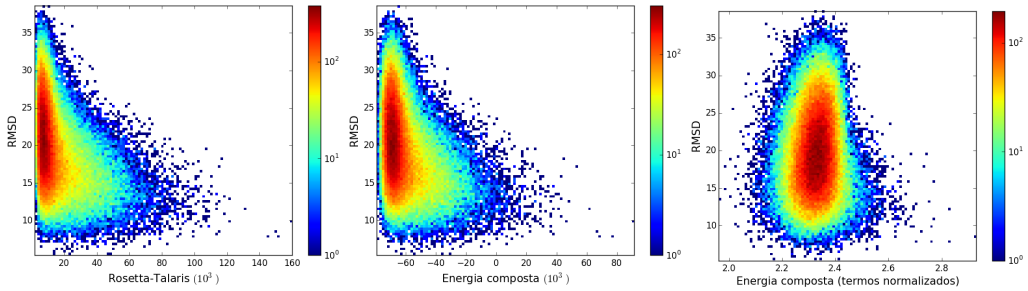
Com isso, a Figura 5.4 ilustra a disposição dos 100.000 indivíduos gerados para o conjunto de testes de proteínas-alvo, a partir da APL-combinada, relacionando RMSD e valores de energia.

Figura 5.4: Análise do conjunto de testes de proteínas-alvo para amostragens de 100.000 estruturas geradas através da APL-combinada (APLs vizinhança e centroide), relacionando RMSD e diferentes configurações de funções de energia

Rosetta-*Talaris2014*	Energia composta	Termos normalizados
		
1AB1		
		
1ACW		
		
1AIL		
		
1DFN		
		
2MR9		

Continuação na próxima página

Figura 5.4: Continuação da tabela da página anterior

Rosetta-*Talaris2014*	Energia composta	Termos normalizados
		
2P5K		
		
3V1A		
		
T0820		

Fonte: do autor (2017).

A partir da análise dos gráficos ilustrados na Figura 5.4, observa-se que para todas as variações de funções de energia e proteínas-alvo, a relação esperada entre energia e RMSD não é percebida. Sabendo que as funções de energia são funções de avaliação de minimização, espera-se que quanto menor for o valor de energia para um determinado modelo estrutural, menor também seja o seu RMSD. Analisando o comportamento das funções *Talaris2014* e energia composta, nota-se que ambas tendem a assumir comportamentos opostos ao que é de fato esperado, onde as regiões que concentram valores mais altos de energia tendem a diminuir o intervalo de variação dos valores de RMSD.

Ainda, no que se refere às funções *Talaris2014* e energia composta, é possível perceber que nas regiões de menores energias dos gráficos, ocorre a maior concentração de indivíduos, sendo que estes apresentam maiores variações de RMSD quando comparados aos indivíduos localizados nas regiões de valores mais altos de energia. Isto dificulta a

distinção entre soluções boas e ruins a partir dos valores de energia atrelados às soluções. Observa-se que a função de energia composta, especificamente para as proteínas-alvo 1ACW e 1DFN, conseguiu reduzir um pouco o intervalo de variação de RMSD entre os indivíduos das regiões de menores e maiores energias. No entanto, de forma geral, ambas tendem a produzir cenários bem similares relativos à relação entre energia estrutural e RMSD.

Quanto à função de energia de termos normalizados, observa-se que esta segue na mesma linha de análise realizada em relação à função composta para as proteínas-alvo 1ACW e 1DFN. A normalização dos termos contribuiu para que houvesse a redução na quantidade de indivíduos localizados em regiões de energias mais baixas, tornando o intervalo de variação dos valores de RMSD mais limitado nestas regiões. Nota-se ainda que esta forma de avaliação, de certo modo, conseguiu prover uma distinção maior entre as conformações estruturais geradas, visto que as áreas de maiores concentrações de indivíduos nos gráficos foram reduzidas, e as soluções apresentaram-se de uma forma mais distribuída. A redução do intervalo de RMSD em regiões de menores energias é interessante, pelo fato de que, sendo o método de otimização guiado pela função de energia, ao final do processo as soluções de menores energias tenderão a apresentar valores de RMSD mais similares. Acredita-se que o padrão diferenciado de distribuição de indivíduos, deve-se à normalização de cada um dos três termos da função composta, onde estes passaram a exercer a mesma relevância no cálculo de energia.

Contudo, a partir da análise da Figura 5.4 para todas as proteínas analisadas, percebe-se a ineficiência destas funções de energia quanto à representação de soluções de boa qualidade, visto que indivíduos com RMSD mais baixos apresentaram valores de energia bastante variados. Destaca-se que um mesmo valor de energia foi capaz de representar inúmeras conformações estruturais para a mesma proteína-alvo. Isto, durante o processo de otimização, pode fazer com que pontos ótimos encontrados não reflitam as estruturas de melhores qualidades, além de impossibilitar a diferenciação de qualidade entre estruturas oriundas de diferentes mínimos locais. Corroborando assim com o que foi posto anteriormente sobre as imperfeições relacionadas a funções de energia para o problema PSP.

Dessa forma, confirma-se que as técnicas de busca, como alvo principal de melhoramentos, precisam incorporar mecanismos mais robustos, capazes de gerar e manter estruturas energicamente aceitáveis e que ao mesmo tempo correspondam a distintas conformações, distribuídas em diferentes mínimos globais. Uma vez que ao final do processo

de otimização diferentes modelos provenientes dos diversos mínimos locais existentes sejam encontrados, obtêm-se entre as diversas conformações, soluções de boa qualidade. Esta observação reforça o propósito do AM multi populacional proposto, o qual objetiva a geração e preservação de um conjunto variado de soluções ao longo do processo, visando contornar a complexidade e ineficiências da função de avaliação.

Embora nenhuma das três funções de energia tenha sido capaz de refletir de forma satisfatória a relação esperada entre RMSD e energia, percebe-se que o processo de normalização dos termos conduziu a uma maior diferenciação entre os modelos estruturais gerados, e reduziu o intervalo de variação dos valores de RMSD relativos às regiões de menores energias, implicando na disposição de menos soluções com valores mais baixos de energia, porém com as mesmas qualidades. Com isso, nota-se que neste trabalho será utilizada a função de energia de termos normalizados como função objetivo do AM desenvolvido. Os valores máximos relativos à cada um dos três termos empregados no processo de normalização da função são definidos durante o processo de amostragem de indivíduos. Os maiores valores obtidos através da geração e avaliação dos indivíduos são utilizados na etapa de otimização de estruturas, como meio de normalizar os valores avaliados pela função de energia composta.

5.3 Análise da filtragem de soluções

O procedimento de filtragem de soluções ocorre após a etapa de amostragem de indivíduos, e objetiva desconsiderar do processo de otimização estruturas consideradas ruins, como forma de prevenir que estas sejam contabilizadas no processo de agrupamento de soluções e na definição dos grupos estruturais que serão utilizados na inicialização das populações do AM. Define-se como estruturas ruins, os modelos estruturais desprovidos de empacotamento. Esta falta de empacotamento é verificada através das métricas de RG e SASA.

Conforme referido anteriormente (Seção 4.3.2), as estruturas resultantes do processo de amostragem de indivíduos, são filtradas a partir de limiares máximos de RG e SASA. Onde, para uma determinada proteína-alvo, os limiares máximos destas métricas são definidos a partir dos maiores valores de RG e SASA vinculados a estruturas previamente conhecidas, retornadas pela consulta à base de dados de proteínas experimentais, a qual relaciona o tamanho da sequência de aminoácidos e a sua classe. Nota-se que a classe de uma proteína a ser predita é definida conforme as características da ES assumida

e informada previamente como parâmetro de entrada para o método, juntamente com a sequência linear de aminoácidos.

Sendo assim, objetivando analisar a etapa de filtragem de soluções, realizou-se processos de amostragem de 10.000 indivíduos, para cada proteína-alvo do conjunto de testes (Tab. 5.1), aplicando e não aplicando os limiares de filtragem. A Tabela 5.2 descreve os limiares máximos de RG e SASA definidos para cada uma das proteínas-alvo, bem como os valores experimentais calculados para cada uma delas. Os indivíduos gerados através dos processos de amostragem foram avaliados por meio da métrica RMSD em relação às estruturas determinadas experimentalmente, e dispostos em gráficos que relacionam os valores de RMSD, RG e SASA assumidos por cada indivíduo. Conforme já mencionado (Seção 4.3), os cálculos de RG e SASA para as estruturas de proteínas foram realizados através das bibliotecas fornecidas pelo PyRosetta. A Figura 5.5 ilustra a disposição dos 10.000 indivíduos gerados para o conjunto de testes de proteínas-alvo, a partir da APL-combinada, onde não foi aplicado o processo de filtragem de soluções. Da mesma forma, a Figura 5.6 demonstra a disposição dos 10.000 indivíduos gerados para cada proteína-alvo, sendo que nestes procedimentos foram realizadas as filtragens de soluções a partir dos limiares máximos de RG e SASA referentes à cada proteína analisada (Tab. 5.2). Os valores indicados entre parênteses representam o número de estruturas que não foram descartadas no procedimento de filtragem.

Tabela 5.2: Limares máximos de RG e SASA empregados no procedimento de filtragem de indivíduos concernentes às proteínas do conjunto de testes. Os valores foram definidos a partir da consulta à base de dados experimentais considerando o tamanho de resíduos de aminoácidos e a classe das proteínas-alvo

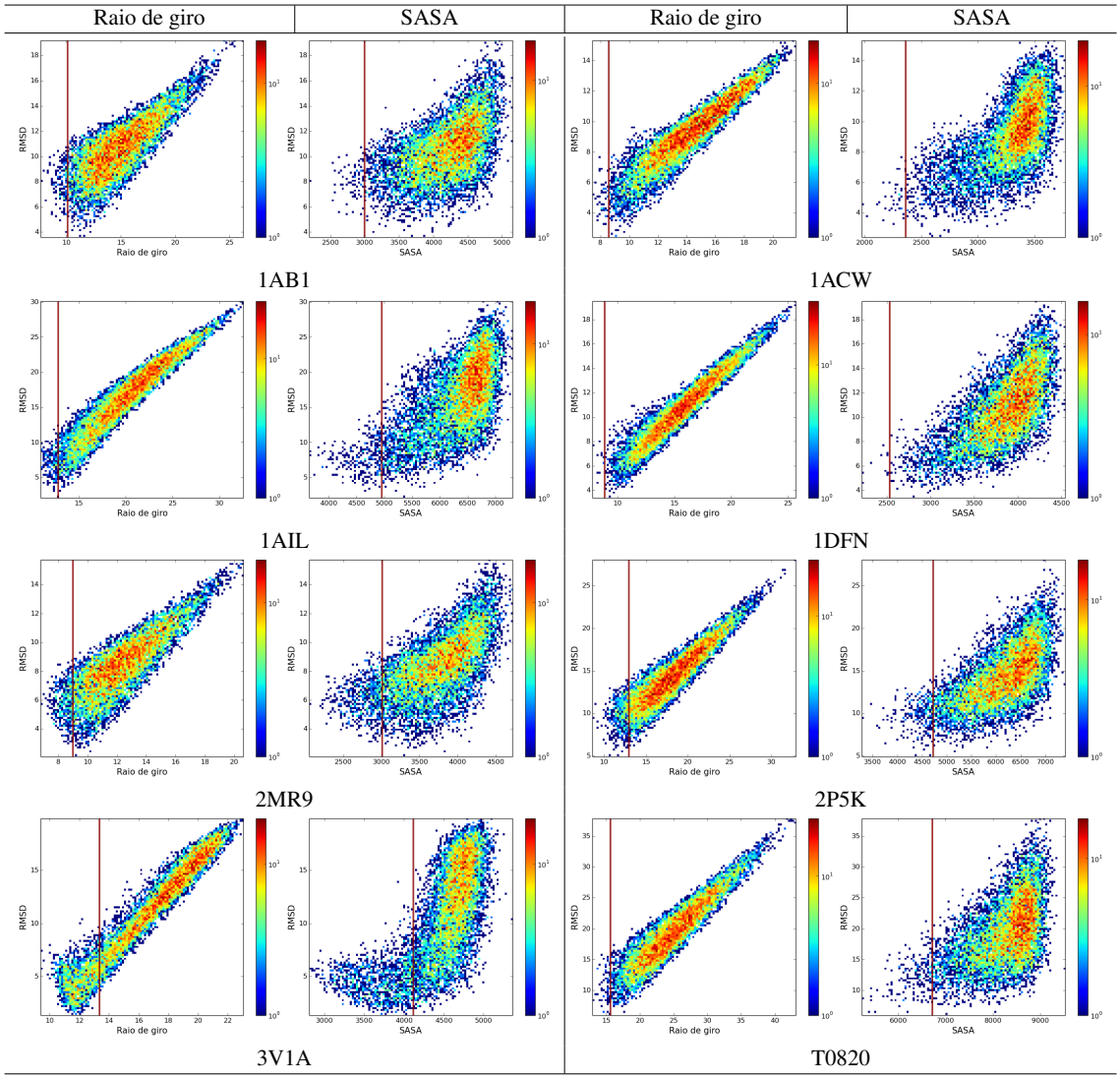
ID-Proteína	**Limiar de RG**	**RG experimental**	**Limiar de SASA**	**SASA experimental**
1AB1	23,81Å	10,11Å	5.728,53Å^2	2.997,46Å^2
1ACW	15,96Å	8,59Å	3.723,72Å^2	2.362,48Å^2
1AIL	18,22Å	12,83Å	6.342,87Å^2	4.951,45Å^2
1DFN	21,80Å	8,90Å	4.265,58Å^2	2.531,59Å^2
2MR9	23,88Å	9,01Å	5.422,31Å^2	3.008,75Å^2
2P5K	29,22Å	12,94Å	6.809,60Å^2	4.724,33Å^2
3V1A	22,91Å	13,35Å	5.868,81Å^2	4.118,30Å^2
T0820-CASP11	26,31Å	15,62Å	9.527,92Å^2	6.715,46Å^2

Fonte: Do autor (2017).

A partir dos limiares de filtragem de RG e SASA descritos na Tabela 5.2, observa-se que em nenhum dos casos apresentados os limiares assumiram valores menores do que os valores calculados para as proteínas experimentais. Isto demonstra que para este conjunto de proteínas-alvo, a base de dados criada mostrou-se robusta e grande o sufi-

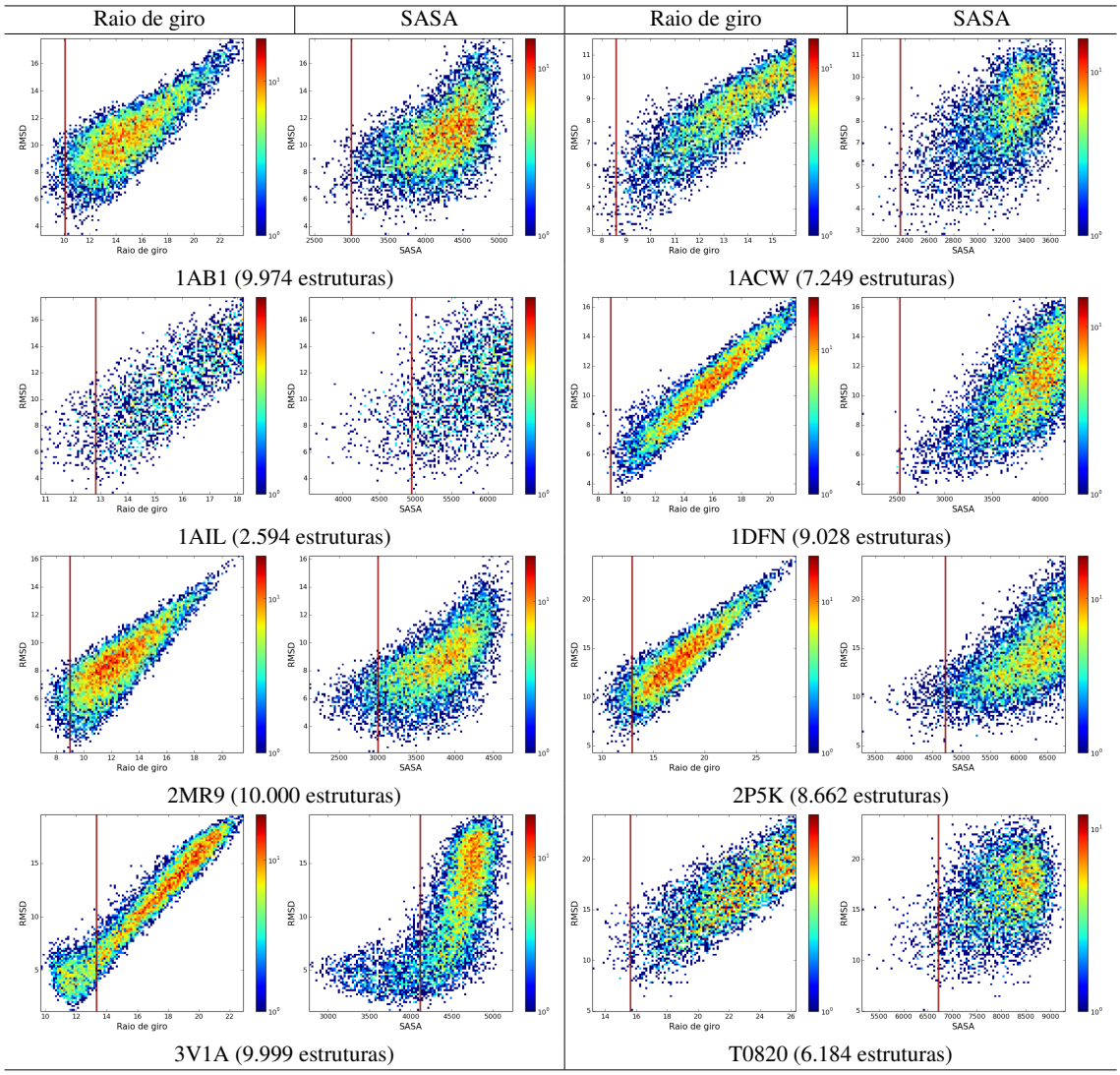
ciente, evitando que os limiares assumissem valores menores do que aqueles definidos através das estruturas determinadas experimentalmente. Limiares de filtragem menores do que os valores de RG e SASA definidos para as estruturas experimentais, implicariam no descarte de soluções que tendem a assumir o mesmo nível de empacotamento destas, fazendo com que o processo de agrupamento de indivíduos e descoberta de diferentes grupos estruturais fosse afetado. Nota-se que na consulta à base de dados para a definição dos valores máximos de RG e SASA, as estruturas experimentais das proteínas-alvo foram excluídas da busca.

Figura 5.5: Análise do conjunto de testes de proteínas-alvo para amostragens de 10.000 estruturas geradas através da APL-combinada (APLs vizinhança e centroide), sem a utilização dos limiares máximos de RG e SASA. Os gráficos relacionam RMSD, RG e SASA. A linha vermelha na vertical indica o valor referente a proteína determinada experimentalmente



Fonte: do autor (2017).

Figura 5.6: Análise do conjunto de testes de proteínas-alvo para amostragens de 10.000 estruturas geradas através da APL-combinada (APLs vizinhança e centroide) utilizando limiares máximos de RG e SASA para filtrar as soluções. Os gráficos relacionam RMSD, RG e SASA. A linha vermelha na vertical indica o valor referente a proteína determinada experimentalmente, e os valores entre parênteses representam o número de estruturas não descartadas



Fonte: do autor (2017).

A partir da análise dos resultados apresentados na Figura 5.6, observa-se que para algumas proteínas-alvo houve a redução significativa de soluções através do descarte, como pode ser percebido para as proteínas-alvo 1AIL, 1ACW e T0820. Considerando as relações previamente estabelecidas entre RMSD, RG e SASA, nota-se que a maior parte dos indivíduos descartados englobaram estruturas com RMSD mais elevado, o que garantiu a manutenção de um conjunto menor de soluções melhores quando comparado ao processo de amostragem que não utilizou os limiares de exclusão (Fig. 5.5). Por exemplo, para a proteína-alvo 1AIL, foram descartados 7.406 indivíduos e considerados apenas 2.594. Ao compararmos o intervalo de variação de RMSD dos gráficos gerados para ambas as figuras (Fig. 5.5 e 5.6), percebe-se que no procedimento de amostragem onde a filtragem de soluções não foi aplicada, os indivíduos chegaram a assumir valores de RMSD de até 30Å, sendo que no processo de amostragem considerando a filtragem de soluções, este valor máximo de RMSD assumido por alguns indivíduos foi reduzido para 18Å. Observa-se este comportamento de redução no valor máximo de RMSD em todas as proteínas-alvo onde um número considerável de soluções foi descartado. No caso da proteína-alvo T0820, este valor máximo de RMSD foi de 38Å para o procedimento que não efetuou o descarte de soluções, sendo reduzido para 25Å quando da aplicação do procedimento.

Ainda, comparando ambas as figuras (Fig. 5.5 e 5.6), percebe-se que em todos os casos analisados, os valores mínimos de RMSD assumidos por alguns indivíduos mantiveram-se similares tanto nos processos de amostragem que não consideraram os limiares de restrição de indivíduos, quanto nas amostragens que aplicaram a filtragem de soluções. Com isso, conclui-se que o descarte de soluções não prejudicou a inicialização do método por meio da restrição de boas soluções, confirmando assim que a etapa de filtragem de indivíduos possui o potencial de rejeitar estruturas de qualidade inferior, à medida que garante a manutenção de soluções melhores.

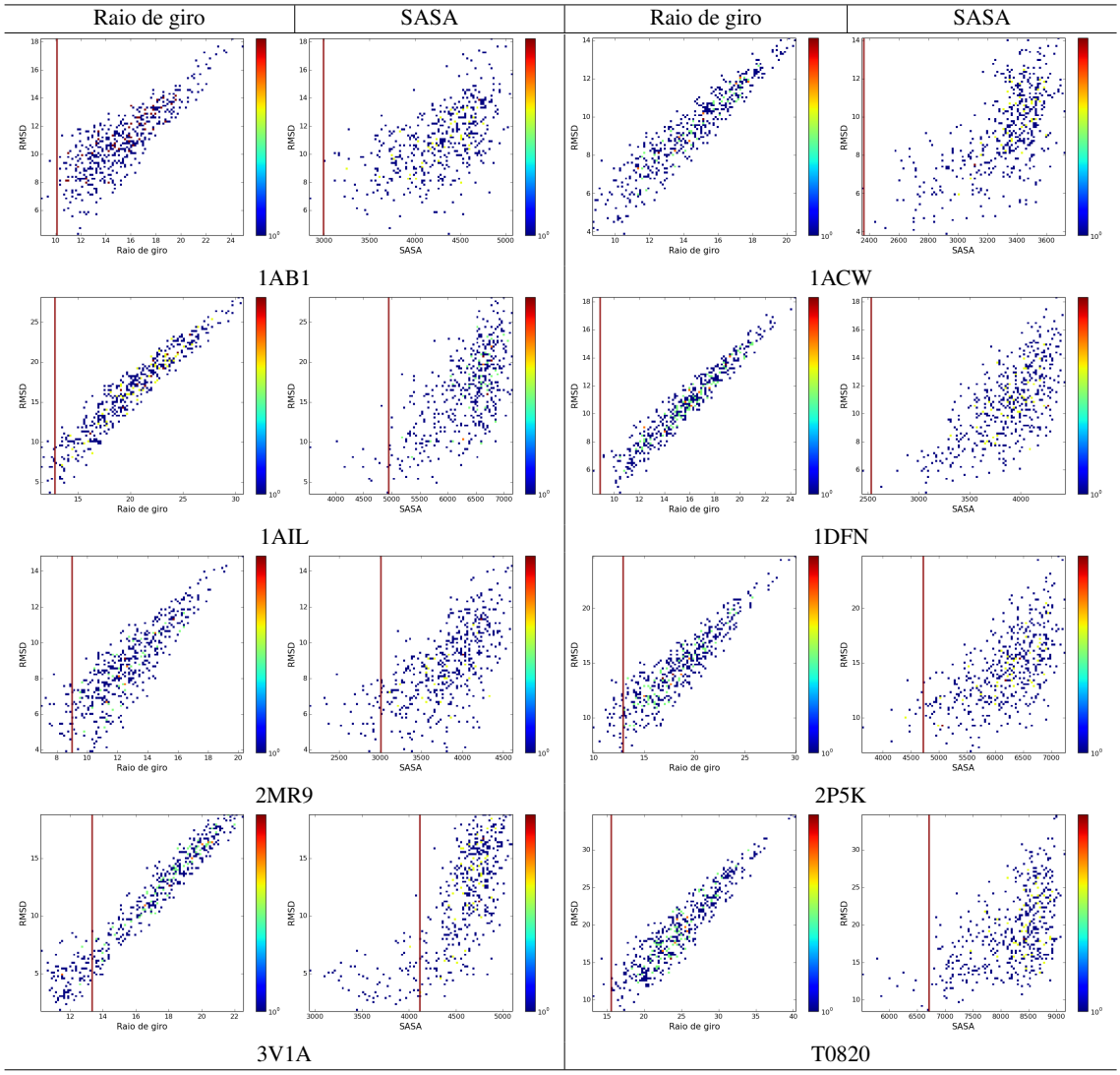
No entanto, analisando a Figura 5.6, nota-se que para algumas proteínas-alvo do conjunto de testes, a filtragem de soluções foi pouco eficiente a partir dos limiares de RG e SASA definidos. Por exemplo, para a proteína 3V1A, o processo de filtragem não foi capaz de descartar nenhuma solução gerada pela amostragem de indivíduos. No caso da proteína-alvo 2MR9, apenas uma solução foi descartada. Isto indica que talvez, dependendo da proteína-alvo considerada no processo, faça-se necessário a inclusão de mais algum fator de restrição estrutural aliado aos dois outros já utilizados.

Por fim, observa-se que apesar dos conjuntos de soluções resultantes dos proces-

sos de filtragem reunirem estruturas de RMSD relativamente mais baixos em comparação com as soluções descartadas, sabe-se que estas estruturas também estão contidas nos conjuntos resultantes dos processos que não consideraram os limiares de restrição. Todavia, estes conjuntos ainda consideram as soluções de RMSD mais altos que foram descartadas nos conjuntos com restrições de RG e SASA, sendo que estas também serão consideradas no processo de agrupamento e exercerão influência sobre a inicialização da meta-heurística.

Com o objetivo de verificar se há vantagens no procedimento de amostragem de muitos indivíduos (10.000, neste trabalho) em relação à geração de poucas soluções, realizou-se amostragens de 500 soluções para cada proteína-alvo do conjunto de testes, sem a aplicação do processo de filtragem. Nota-se aqui que a atribuição dos ângulos diedros para as estruturas geradas foi realizada da mesma forma em ambos os procedimentos, a partir das preferências conformacionais dos aminoácidos representadas na APL-combinada. Com isso, a Figura 5.7 ilustra a disposição dos 500 indivíduos gerados para o conjunto de testes de proteínas-alvo, relacionando as métricas de RMSD, RG e SASA.

Figura 5.7: Análise do conjunto de testes de proteínas-alvo para amostragens de 500 estruturas geradas através da APL-combinada (APLs vizinhança e centroide), relacionando RMSD, RG e SASA. A linha vermelha na vertical indica o valor referente a proteína determinada experimentalmente



Fonte: do autor (2017).

Comparando as Figuras 5.6 e 5.7, percebe-se que o processo de amostragem de 10.000 indivíduos para determinada proteína-alvo, mostra-se capaz de gerar soluções com valores de RMSD melhores quando da inicialização de apenas 500 soluções, visto que um conjunto maior de indivíduos é produzido. Além de conseguir concentrar mais soluções nas regiões de RMSD mais baixos. A amostragem de grandes quantidades de soluções acaba proporcionando ao método mais oportunidades quanto à exploração do conhecimento experimental contido nas APLs.

A título de exemplo desta constatação, para a proteína 1ACW, nota-se que o valor mínimo de RMSD assumido por alguns indivíduos ficou em torno de 4Å para o procedimento de amostragem de 500 soluções, sendo que este valor foi reduzido em mais de 1Å quando da geração de 10.000 indivíduos. Percebe-se a redução deste valor mínimo de RMSD em todos os casos analisados. Para a proteína-alvo T0820, por exemplo, a geração de 500 indivíduos atingiu um valor mínimo de RMSD em torno de 8Å, sendo que a amostragem de 10.000 soluções reduziu esta valoração para 5Å.

Dessa forma, infere-se que a amostragem de diversos indivíduos, aliada ao procedimento de filtragem de soluções, pode ser considerada uma eficiente abordagem para a inicialização de indivíduos de boa qualidade.

5.4 Análise do agrupamento de soluções

Esta seção pretende ilustrar os diferentes grupos estruturais encontrados após a execução do procedimento de filtragem de soluções. O agrupamento de modelos estruturais tem por objetivo destacar os diversos grupos oriundos de um mesmo processo de criação. Com isso, os indivíduos resultantes do processo de filtragem de soluções, gerados a partir da amostragem de 10.000 modelos estruturais, foram agrupados conforme as suas similaridades estruturais, visando a identificação de diferentes padrões conformacionais para cada proteína-alvo. O agrupamento de soluções representa a última etapa do processo de amostragem e classificação de indivíduos, e está descrito na Seção 4.3.3.

Neste trabalho, os diferentes grupos estruturais formados, são utilizados como forma de prover indivíduos iniciais às multi populações do AM (Seção 4.4), visando a diversidade das populações, na tentativa de contornar os problemas da multimodalidade da função objetivo (HANDL; LOVELL; KNOWLES, 2008), ao mesmo tempo que preocupa-se com a qualidade das soluções iniciais.

Dessa forma, objetivando analisar os grupos estruturais formados a partir do pro-

cesso de agrupamento, bem como as diferentes métricas possíveis de ranqueamento de cada conjunto de grupos, realizou-se, para cada proteína-alvo do conjunto de testes definido (Tab. 5.1), o procedimento de agrupamento das soluções oriundas do processo de filtragem.

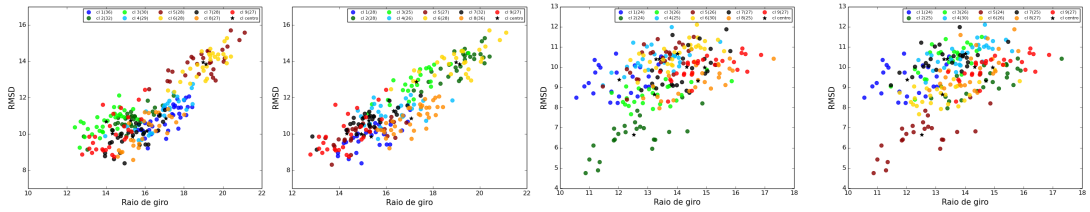
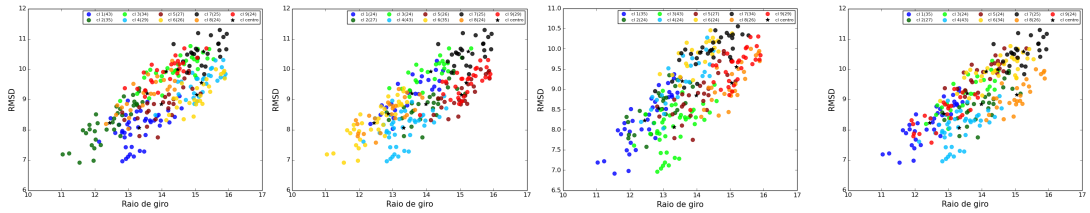
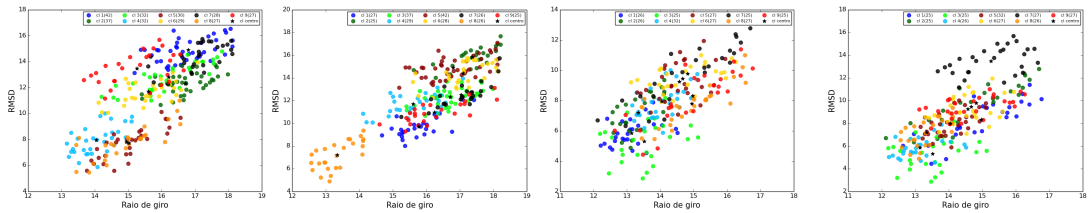
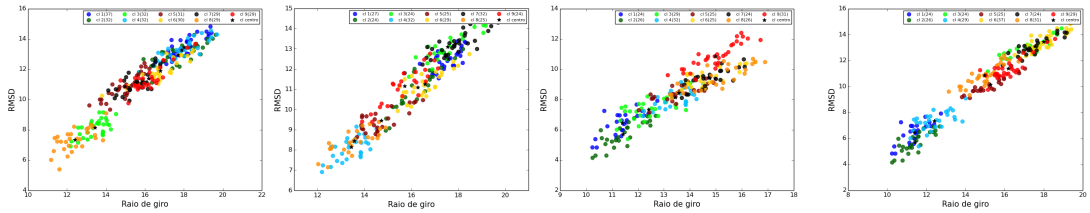
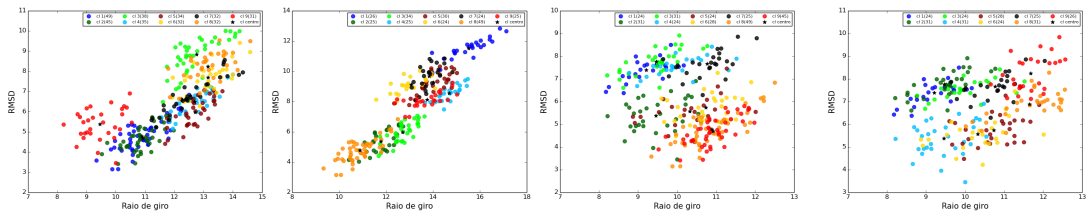
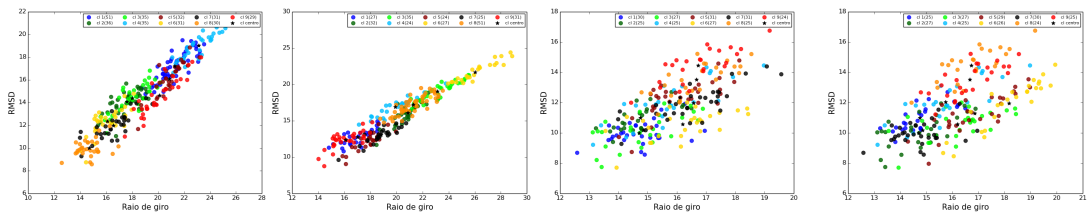
O agrupamento dos indivíduos foi realizado através da técnica de clusterização hierárquica aglomerativa, onde a Tabela 5.3 descreve os limiares de corte (*cut-off*) utilizados no agrupamento de cada proteína-alvo. Estes limiares representam a similaridade estrutural permitida entre os indivíduos de um mesmo grupo, sendo calculada e expressa através da métrica RMSD. Os limiares estabelecidos para cada proteína, respeitaram a formação de no mínimo 9 diferentes grupos, onde cada grupo devia conter 24 ou mais estruturas. Estes valores foram definidos em conformidade com a estruturação do AM de otimização, sendo que 9 representa o número de subpopulações inicializadas a partir dos grupos formados e 24 denota o número de indivíduos em cada subpopulação. Os indivíduos integrantes dos grupos resultantes do procedimento de agrupamento foram avaliados por meio da métrica RMSD em relação às estruturas determinadas experimentalmente, e dispostos em gráficos que relacionam os valores de RMSD e RG assumidos por cada indivíduo. A Figura 5.8 ilustra a disposição dos 9 melhores grupos gerados para cada proteína do conjunto de testes, sendo que estes foram ordenados considerando o número de indivíduos em cada grupo, valor de energia médio, valor de RG médio e valor de SASA médio de cada grupo. Quanto ao procedimento que classificou os grupos por meio do número de indivíduos existentes em cada grupo, foram escolhidos os 9 grupos mais populosos, visto que a presença de mais soluções em um grupo demonstra uma tendência conformacional maior. Em contrapartida, para os agrupamentos ordenados através das outras métricas, os grupos escolhidos foram aqueles que apresentaram os menores valores. Nota-se ainda que em cada um dos gráficos da Figura 5.8, os grupos formados estão representados por diferentes cores.

Tabela 5.3: Limiares de corte utilizados em cada agrupamento de dados após a amostragem de 10.000 estruturas, para o conjunto de testes de proteínas-alvo

ID-Proteína	**Limiar de similaridade (RMSD)**
1AB1	7,08Å
1ACW	4,57Å
1AIL	8,62Å
1DFN	6,11Å
2MR9	5,53Å
2P5K	9,98Å
3V1A	3,63Å
T0820-CASP11	10,98Å

Fonte: Do autor (2017).

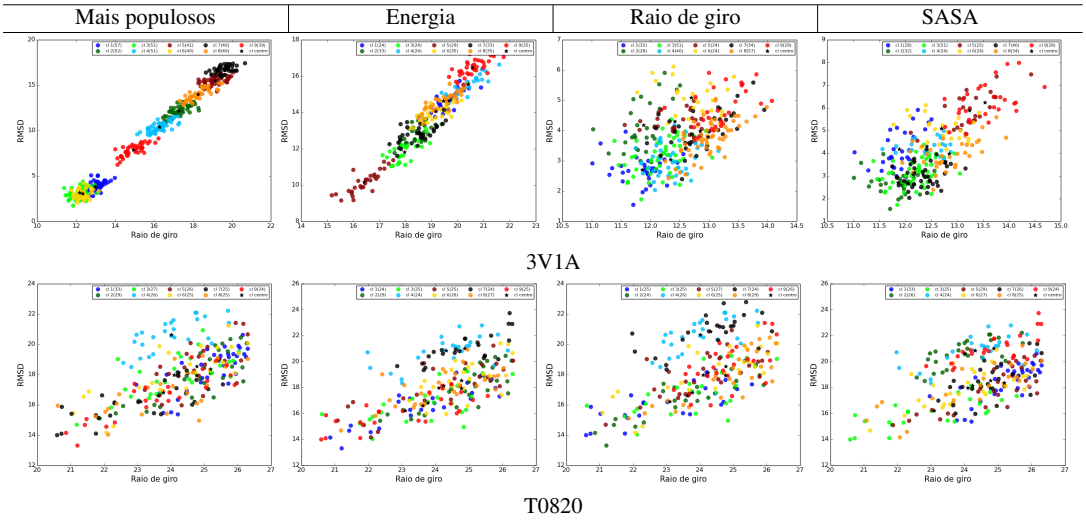
Figura 5.8: Análise do conjunto de testes de proteínas-alvo para o procedimento de agrupamento de indivíduos oriundos dos processos de filtragem de soluções, relacionando RMSD e RG. A ordenação dos diferentes grupos considerou o número de indivíduos em cada grupo (mais populosos), valor de energia médio, valor de RG médio e valor de SASA médio de cada grupo

Mais populosos	Energia	Raio de giro	SASA
			
1AB1			
			
1ACW			
			
1AIL			
			
1DFN			
			
2MR9			
			
2P5K			

Continuação na próxima página

Fonte: do autor (2017).

Figura 5.8: Continuação da tabela da página anterior



Fonte: do autor (2017).

Analisando os diferentes grupos formados em cada um dos agrupamentos ilustrados na Figura 5.8, percebe-se que para algumas proteínas-alvo a identificação de diferentes padrões conformacionais é bastante evidente, como nos casos das proteínas 1DFN e 2MR9. Ainda, com exceção da proteína T0820, nota-se que para todas as outras proteínas, obteve-se, de certa forma, a distinção estrutural dos indivíduos agrupados, mesmo que em alguns casos, eles se sobrepuseram em relação ao RMSD e RG nos gráficos de distribuição. Esta sobreposição exibida em alguns gráficos, como para as proteínas-alvo 1ACW e 1AB1, indica que os grupos identificados apresentaram graus de empacotamento e RMSD em relação às proteínas experimentais similares, porém com diferentes conformações. Especificamente para a proteína-alvo T0820, percebe-se que a identificação de diferentes grupos estruturais não ocorreu de forma clara. Talvez o motivo disto ter ocorrido, possa ser explicado pelo valor mais elevado do limiar de corte utilizado no agrupamento, combinado à sua difícil conformação.

Destaca-se que os ranqueamentos de grupos considerando os grupos mais populosos e o valor de energia médio, mostraram-se capazes de gerar grupos mais distintos uns dos outros, os quais apresentaram diferentes relações entre RMSD e RG, conforme pode ser observado nas proteínas 2P5K e 3V1A. No entanto, apesar destas duas estratégias terem sido capazes de identificar melhores grupos estruturais, percebe-se que a variação nos valores de RMSD em alguns casos é bem maior quando comparada com as classificações de grupos por RG médio e SASA médio. Por exemplo, para a proteína 1AB1, o intervalo de RMSD dos grupos de indivíduos ordenados pelas duas primeiras estratégias, compreendeu os valores em torno de 8Å a 16Å, enquanto que os intervalos de

RMSD para os outros dois tipos de classificação, variaram em torno de 5Å a 13Å. Para a proteína 1AIL, estas diferenças de intervalos de RMSD foram ainda maiores, sendo que as duas primeiras estratégias selecionaram grupos de indivíduos com valores de RMSD que figuraram no intervalo em torno de 5Å a 17Å, já os outros dois tipos de classificação escolheram grupos onde os indivíduos apresentaram variação nos valores de RMSD em torno de 3Å a 13Å para o RG médio e 3Å a 16Å para o SASA médio. Ainda, observa-se também esta tendencia na proteína 3V1A, onde estas diferenças no intervalo de valores de RMSD assumidos pelas diferentes classificações mostraram-se mais expressivas. Para a primeira estratégia, o intervalo de valores figurou em torno de 1Å a 17Å, e para a segunda estratégia este intervalo variou de 9Å até 18Å. As outras duas estratégias (RG e SASA) selecionaram grupos que implicaram na redução deste intervalo de valores de RMSD assumidos pelos indivíduos agrupados. A estratégia de RG médio apresentou intervalo de valores variando em torno de 2Å até 6Å e a estratégia de SASA médio escolheu grupos onde os indivíduos apresentaram valores de RMSD que variaram no intervalo de 2Å a 8Å. Nota-se aqui que a ordenação de grupos realizada através do valor de RG médio e SASA médio, tenderam a selecionar grupos bem similares.

Sendo assim, sabendo que o objetivo desta etapa de amostragem e inicialização de indivíduos consiste na identificação de diferentes grupos estruturais que englobem estruturas de melhores qualidades a serem utilizadas como soluções iniciais do AM, optou-se por realizar a classificação de grupos, após o processo de agrupamento, através dos valores de RG médio apresentados por estes. Embora as duas primeiras estratégias de ordenação tenham se mostrado capazes de selecionar grupos de indivíduos mais diversificados entre si, o ranqueamento através do RG médio foi capaz de ordenar os grupos gerados, de forma que as melhores soluções tenderam a ficar entre os primeiros grupos. Acredita-se que a ordenação realizada pelo RG médio dos grupos fornece um balanceamento ideal entre a identificação de diferentes grupos estruturais e a seleção das melhores soluções geradas pelos processos de amostragem e filtragem de modelos estruturais.

Dessa forma, definiu-se que os grupos identificados serão classificados considerando o valor de RG médio do grupo, e aqueles que apresentarem os menores valores serão utilizados como indivíduos iniciais da meta-heurística de otimização.

5.5 Resumo do capítulo

Este capítulo apresentou os resultados obtidos a partir da estratégia de amostragem e classificação de indivíduos, proposta como forma de prover soluções diversificadas e de qualidade à meta-heurística de otimização. Também forneceu análises concernentes a estes resultados, visando embasar as decisões tomadas na estruturação algorítmica da etapa I do método de otimização. As análises discorridas relativas à etapa I consistiram em: (*i*) análises do comportamento das métricas de RG e SASA em relação à métrica de qualidade estrutural RMSD; (*ii*) amostragem de indivíduos a partir dos diferentes tipos de APLs e discussão dos cenários apresentados; (*iii*) análises do comportamento de três diferentes funções de energia quanto à capacidade de representar boas estruturas; (*iv*) discussão sobre o comportamento da filtragem de soluções e implicações relativas aos indivíduos resultantes do processo; e (*v*) exemplificação do procedimento de agrupamento de indivíduos e discussão acerca dos diferentes tipos possíveis de classificação dos grupos estruturais gerados.

Em síntese, constatou-se que as métricas de avaliação estrutural de RG e SASA podem ser utilizadas como bons indicadores do nível de empacotamento de estruturas de proteínas. Verificou-se que não há diferenças entre os diferentes tipos de APLs analisados quanto à qualidade dos indivíduos gerados, porém optou-se pela utilização APL-combinada na etapa de amostragem de indivíduos do método proposto, visto a maior disponibilidade de dados experimentais. Concluiu-se que nenhuma das três funções de energia analisadas foram capazes de representar a relação esperada entre energia estrutural e RMSD, contudo, a função de energia de termos normalizados será utilizada como função objetivo, dada a sua maior capacidade de diferenciar modelos estruturais. Também foi observado que o procedimento de filtragem de soluções possui o potencial de descartar soluções ruins, enquanto garante a manutenção de boas soluções. E por fim, constatou-se que a estratégia de agrupamento de soluções consegue identificar diferentes grupos estruturais oriundos do processo de filtragem de soluções, e que a classificação dos grupos formados em decorrência do agrupamento será realizada pela métrica de RG médio de cada grupo, a qual demostrou um balanceamento ideal entre identificação de diferentes grupos estruturais e seleção de soluções de melhores qualidades.

O próximo capítulo visa descrever os resultados obtidos concernentes à otimização de um conjunto de proteínas-alvo, realizada pelo AM multi populacional desenvolvido, aliado à etapa I de amostragem e inicialização de indivíduos.

6 ANÁLISES E RESULTADOS - ETAPA II

Este capítulo objetiva a apresentação dos resultados obtidos relativos à execução e à eficácia da abordagem proposta neste trabalho, aplicada na otimização de um conjunto de proteínas-alvo. A abordagem proposta conta com a etapa I de amostragem e inicialização de indivíduos, descrita anteriormente na Seção 4.3, integrada ao AM multi populacional de otimização, que por sua vez, utiliza como indivíduos iniciais as soluções provenientes da etapa I, com o intuito de inicializar o método com soluções mais diversificadas e de qualidade. O AM proposto, bem como a meta-heurística ABC utilizada na otimização dos indivíduos constituintes de cada subpopulação, interações entre nichos e operadores de busca, estão detalhados na Seção 4.4.

Dessa forma, visando analisar a etapa II de otimização de estruturas 3-D de proteínas, este capítulo foi estruturado em duas partes. A primeira parte reúne experimentos de otimização realizados com o intuito de demonstrar as adaptações ocorridas no AM ao longo do trabalho para que este culminasse na abordagem proposta, assim como definir os parâmetros mais adequados para a meta-heurística. Nesta primeira etapa de experimentos, foi utilizado o mesmo conjunto de testes apresentado no capítulo anterior, o qual engloba 8 diferentes proteínas-alvo.

A partir dos resultados obtidos e análises discorridas relativos à primeira parte deste capítulo, a segunda parte de experimentos tem por objetivo testar o AM desenvolvido através da otimização de um conjunto maior de proteínas-alvo, bem como compará-lo a dois métodos de referência na área de predição de estruturas 3-D de proteínas, Rosetta e QUARK. Ambos descritos na Seção 3.1.

Com isso, a Tabela 6.1 descreve o conjunto composto por 24 proteínas-alvo utilizado neste capítulo de avaliação do método proposto. As proteínas destacadas em negrito, denotam as 8 proteínas-alvo empregadas como estudos de caso de otimização referentes à primeira parte de experimentos deste capítulo. Para os testes de otimização relacionados à segunda parte do capítulo, todas as proteínas-alvo do conjunto descrito na Tabela 6.1 foram empregadas. Nota-se que as proteínas foram selecionadas visando garantir que o conjunto de testes englobe sequências de aminoácidos de diferentes tamanhos e com variadas topologias estruturais. O detalhamento de cada proteína-alvo, como tamanho e composição de ES, pode ser observado na Tabela 6.1. Assim como no capítulo anterior, observa-se que a proteína-alvo T0820 foi extraída do conjunto de proteínas-alvo apresen-

tado nos experimentos do CASP11[1], a qual foi referida pelo ID T0820-D1 e alocada na categoria FM (KINCH et al., 2016b).

Tabela 6.1: Conjunto de testes de proteínas-alvo empregado nos experimentos e análises relativos à segunda etapa de otimização do método proposto. As proteínas-alvo destacadas em negrito denotam o conjunto de proteínas utilizado na primeira parte de análises deste capítulo

ID-Proteína	Tamanho	Composição de ES
1AB1	46	1 folha/2 hélices
1ACW	29	1 folha/1 hélice
1AIL	70	3 hélices
1CRN	46	1 folha/2 hélices
1D5Q	27	1 folha/1 hélice
1DFN	30	1 folha tripla
1ENH	54	3 hélices
1FNA	91	2 folhas (1 quádrupla)
1K43	14	1 folha
1L2Y	20	2 hélices
1OPD	85	1 folha (quádrupla)/3 hélices
1Q2K	31	1 folha/1 hélice
1ROP	56	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	76	3 hélices
3V1A	48	2 hélices
T0820-CASP11	90	3 hélices

Fonte: Do autor (2017).

6.1 Variações do algoritmo memético

Esta seção objetiva investigar o AM desenvolvido sob a perspectiva de resultados obtidos quando da otimização de estruturas 3-D de proteínas, bem como analisar os diversos componentes inseridos no método, como a estruturação da população em árvore e a etapa I de amostragem de indivíduos, através da execução de diferentes versões do algoritmo proposto. Objetiva-se também definir o valor mais adequado ao parâmetro relacionado à diversidade de soluções, empregado na etapa de reinicialização das subpopulações da meta-heurística.

O AM proposto neste trabalho, aliado ao método ABC, apresenta uma série de

[1]<www.predictioncenter.org/casp11/targetlist.cgi>

parâmetros necessários à sua execução. Estes parâmetros foram sendo apresentados ao longo do texto, e estão explicados na Seção 4. Destaca-se que estes tiveram seus valores previamente estabelecidos e não sofrerão variações nas diferentes versões do método que serão analisadas nesta seção. Com isso, o conjunto de parâmetros preestabelecidos concernente à abordagem envolvendo AM e ABC, está descrito abaixo:

- Condição de parada: a condição de parada do método é definida pelo número máximo de avaliações de energia permitido durante uma execução. Para todas as execuções definiu-se 1.000.000 de cálculos de energia;
- Tamanho de cada subpopulação do AM: o número de indivíduos em cada subpopulação do AM foi definido em 24 soluções. Este número foi estipulado com o objetivo de não tornar o tamanho da população total do método demasiadamente grande, o que implicaria na perda de eficiência computacional;
- Tamanho da população global do AM: o tamanho da população do AM representa a soma total de todos os indivíduos integrantes das subpopulações definidas. Sendo assim, o tamanho da população global é dada por 13×24, onde 13 representa o número de nichos da estrutura em árvore do algoritmo e 24 é o número de soluções em cada nicho. Totalizando 312 indivíduos na população global do AM;
- Soluções descartadas na etapa de reinício do AM: definiu-se que serão descartadas as 6 piores soluções de uma determinada subpopulação quando este procedimento for aplicado, o que representa 25% do total de indivíduos;
- g_{ABC}: este parâmetro representa o número de gerações que o ABC realiza quando aplicado na otimização de uma subpopulação. O g_{ABC} foi definido como 10;
- l: o parâmetro l representa o limiar de determinação para o descarte de uma solução no algoritmo ABC, modelando a etapa das abelhas exploradoras, onde ocorre o abandono de uma fonte de alimentação. Definiu-se $l = 200$, ou seja, se a mais de 200 gerações uma solução não tiver sido melhorada, ela é descartada;
- MR: o parâmetro MR é utilizado no controle de atualizações de variáveis no ABC para uma determinada solução. Definiu-se $MR = 0,4$. Portanto, cada variável será atualizada com probabilidade de 40%. Esta atualização ocorre apenas em variáveis relacionadas aos aminoácidos com ESs irregulares.

Sendo assim, restou apenas a definição do parâmetro p, relativo à etapa de reinicialização do AM, explicado na Seção 4.4.4. O parâmetro p é responsável por realizar o controle da diversidade de cada subpopulação através do procedimento de reinicialização.

Baseado no CV relativo ao RG dos integrantes de uma determinada subpopulação, esta será reinicializada caso apresente CV menor do que $p\%$. Valores mais altos de p indicam taxas de reinicializações maiores.

Com o objetivo de analisar os diferentes componentes incorporados ao AM, e definir a valoração mais razoável para p, criou-se algumas variações do método proposto. Os componentes que serão abordados nestas diferentes versões da meta-heurística são: (*i*) organização populacional do AM em nichos, baseada na estrutura em árvore ternária apresentada, e incorporação da etapa I de amostragem e inicialização de indivíduos; (*ii*) procedimento de reinicialização das subpopulações; e (*iii*) modificações relativas ao funcionamento do algoritmo ABC, como a operação de cruzamento inserida entre as etapas 1 e 2 de mutação de soluções e as atualizações de variáveis realizadas apenas em regiões mais flexíveis da proteína-alvo.

Assim sendo, foram implementadas 6 versões diferentes de métodos de otimização considerando a abordagem proposta. Observa-se que a descrição destas variações se dará através de um processo incremental, ou seja, a primeira versão descrita configura a versão mais básica do algoritmo. As próximas versões incorporam todos os componentes agregados nas versões anteriores. Nota-se também que a inicialização dos indivíduos de todas as variações do método foi realizada através da APL-combinada, e utilizou-se a função de energia de termos normalizados como função de avaliação. Os valores definidos acima, relativos à cada parâmetro do método, mantêm-se os mesmos, caso este parâmetro exista em determinada versão do algoritmo.

A primeira variação do método consiste apenas na utilização do algoritmo ABC, sem a integração da estrutura em árvore e a organização da população em nichos. Esta versão também não faz uso da etapa I de amostragem e inicialização de indivíduos. Contudo, as soluções da população são ainda inicializadas através da APL-combinada. O algoritmo ABC, implementado nesta versão, não recebeu as modificações descritas anteriormente, relacionadas a operações de cruzamento e mutações em regiões irregulares da proteína-alvo. Consequentemente, a otimização de estruturas é realizada em todas as variáveis ao invés de ser aplicada apenas nas regiões irregulares da proteína. Nota-se que o procedimento de verificação dos novos valores gerados através das operações de atualização de variáveis foi mantido. O tamanho da população do algoritmo é de 312 indivíduos. Este número foi utilizado para equiparar com o tamanho da população global do AM e proporcionar comparações mais justas. Esta versão será referida como ABC.

A segunda versão proposta também consiste apenas na utilização do algoritmo

ABC, sem a incorporação da estrutura em árvore, abordagem em nichos e a etapa I de amostragem de indivíduos. No entanto, nesta versão foram inseridas as modificações realizadas referentes ao funcionamento do algoritmo ABC. Tais modificações consistem na otimização apenas de variáveis que compreendem aminoácidos com ESs irregulares e operações de cruzamento inseridas entre as etapas 1 e 2 de atualização de variáveis das soluções. Esta versão foi proposta objetivando avaliar a influência destas duas modificações no processo de otimização de estruturas, e será referida como ABC-mod.

A terceira versão desenvolvida consiste basicamente no AM descrito anteriormente. Esta versão engloba todos os componentes já apresentados, com exceção do procedimento de reinicialização das subpopulações do algoritmo. Esta variação objetiva avaliar a influência dos componentes do AM na otimização de estruturas quando comparada com a otimização realizada apenas pelo algoritmo ABC, conforme as duas versões anteriores. Esta variação será mencionada como M-1.

Por fim, a quarta variação proposta consiste no AM completo, o qual abrange todas as etapas e componentes já explicados. Nesta versão, portanto, inseriu-se o procedimento de reinicialização das subpopulações. Com isso, o parâmetro p utilizado como limiar de aplicação do reinício populacional, recebeu o valor 15 ($p = 15$). Isto representa que se determinada população apresentar CV menor do que 15%, os 6 piores indivíduos (valores mais altos de energia) são descartados e novas soluções são inseridas. Nota-se que estas inserções ocorrem a partir dos grupos estruturais atrelados à cada subpopulação. Esta versão objetiva contrastar com a versão anterior quando da influência da aplicação do procedimento de reinicialização sobre a otimização de estruturas, e será referida como MABC-1.

As outras duas versões finais consistem apenas na variação do valor atribuído ao parâmetro p. Deste modo, a quinta versão, MABC-2, utiliza $p = 10$, e a última versão proposta, MABC-3, considera $p = 5$. Ressalta-se que quanto menor o valor de p, menos reinicializações são aplicadas, visto que a população de indivíduos leva mais tempo para convergir a este critério.

Dessa forma, para avaliar as diferentes versões descritas acima, foi realizado o processo de otimização para as 8 proteínas-alvo destacadas em negrito na Tabela 6.1. Para esta série de experimentos, cada variação do método foi executada 8 vezes, para cada proteína, utilizando como critério de parada o valor máximo de 1.000.000 de cálculos de energia. A Tabela 6.2 sumariza os resultados relativos aos valores de RMSD obtidos, a partir das 8 execuções de cada versão do método, aplicadas na otimização do conjunto de

testes de proteínas-alvo definido (Tab. 6.1).

Observa-se que a melhor solução considerada em cada uma das execuções dos métodos, foi a que apresentou o menor valor de RMSD, dentre a melhor solução de cada subpopulação do AM, definida através dos valores de energia. Assim, para uma determinada execução do método, as 13 melhores soluções (menores valores de energia), cada uma proveniente de uma subpopulação diferente, foram comparadas quanto aos seus valores de RMSD relativos à determinada proteína-alvo. A de menor valor foi contabilizada como sendo a melhor solução encontrada dentre as 13 de menores energias. No entanto, nota-se que as duas primeiras versões criadas, ABC e ABC-mod, não utilizam esta abordagem em nichos, portanto, para tornar a comparação mais justa, considerou-se as 13 melhores soluções finais da população do algoritmo ABC, considerando os menores valores de energia. Fez-se o mesmo procedimento de comparação das 13 soluções em relação aos valores de RMSD obtidos. A solução de menor RMSD foi escolhida como sendo a melhor solução da execução. Embora o ABC não incorpore nenhum tipo de divisão populacional, aplicou-se esta estratégia para avaliar a capacidade de otimização do AM quando comparado ao ABC.

Tabela 6.2: Resumo dos resultados obtidos em relação ao RMSD para as 8 execuções de cada versão do método proposto. Os valores em negrito denotam os melhores resultados

Método	**RMSD (Å)**							
	Mín.	**Médio**	**Mín.**	**Médio**	**Mín.**	**Médio**	**Mín.**	**Médio**
	1AB1		**1ACW**		**1AIL**		**1DFN**	
ABC	2,62	4,81 ± (1,06)	1,85	2,82 ± (0,70)	4,31	5,99 ± (1,16)	2,56	3,69 ± (0,61)
ABC-mod	3,35	4,09 ± (0,57)	1,68	2,09 ± (0,30)	2,54	5,61 ± (1,30)	3,34	4,21 ± (0,47)
M-1	3,64	4,05 ± (0,31)	1,83	2,20 ± (0,29)	4,38	5,21 ± (0,75)	2,44	**3,57** ± (0,64)
MABC-1	**2,45**	**3,50** ± (0,71)	1,74	2,00 ± (0,23)	**2,71**	**4,58** ± (1,28)	**1,84**	3,83 ± (1,02)
MABC-2	2,77	3,77 ± (0,67)	1,60	**1,85** ± (0,22)	3,33	5,70 ± (1,28)	2,64	3,65 ± (0,59)
MABC-3	2,80	4,07 ± (0,86)	**1,46**	1,87 ± (0,28)	3,09	5,68 ± (1,48)	3,83	4,18 ± (0,26)
Método	**2MR9**		**2P5K**		**3V1A**		**T0820**	
ABC	2,14	3,85 ± (1,77)	5,93	7,67 ± (0,96)	2,23	3,00 ± (0,57)	6,58	8,25 ± (1,02)
ABC-mod	1,95	2,74 ± (0,81)	3,93	7,01 ± (1,76)	1,13	1,57 ± (0,63)	6,06	7,80 ± (1,17)
M-1	**1,62**	1,91 ± (0,20)	3,54	**4,74** ± (0,76)	1,07	1,33 ± (0,21)	5,79	**6,28** ± (0,41)
MABC-1	1,71	2,22 ± (0,61)	**3,24**	5,62 ± (1,17)	**0,82**	**1,11** ± (0,13)	**4,45**	6,92 ± (1,25)
MABC-2	1,78	2,04 ± (0,26)	5,25	6,10 ± (0,71)	1,03	1,17 ± (0,13)	5,48	6,63 ± (0,75)
MABC-3	1,64	**1,90** ± (0,20)	3,43	4,89 ± (0,80)	0,90	1,23 ± (0,23)	6,10	7,39 ± (0,77)

Fonte: Do autor (2017).

Analisando os resultados mostrados na Tabela 6.2, percebe-se que tanto o menor valor mínimo de RMSD, quanto a menor média de RMSD, obtidos para cada proteína-alvo analisada, foram atingidos pelas variações do método que assumiram a estrutura algorítmica do AM proposto, sendo eles M-1, MABC-1, MABC-2 e MABC-3. Isto demonstra que a abordagem proposta, incluindo todos os componentes agregados, como a

etapa I de amostragem e inicialização de indivíduos e a organização da população em subpopulações independentes baseadas em nichos, possui o potencial de melhorar a otimização de estruturas de proteínas quando comparada com a aplicação de apenas um método de busca, como ocorreu nas versões ABC e ABC-mod. Comparando os resultados de menores valores mínimos de RMSD e médias de RMSD alcançados pelos métodos ABC e ABC-mod, percebe-se que as alterações realizadas no funcionamento do algoritmo ABC, conseguiram auxiliar o método na obtenção de melhores resultados, visto que o ABC-mod não obteve menor valor de RMSD apenas em dois estudos de caso, 1AB1 e 1DFN, e atingiu valores médios de RMSD menores do que o ABC em 7 casos.

Com o intuito de avaliar a influência do procedimento de reinicialização inserido no AM, desenvolveu-se a versão M-1, que não o incorporou. Comparando os resultados obtidos pelo método M-1 em relação aos outros 3 métodos, MABC-1, MABC-2 e MABC-3, os quais incorporam o procedimento assumindo diferentes limiares de reinício, percebe-se que há a tendência de melhora quanto aos valores de RMSD obtidos, porém esta melhora foi pouco perceptível, visto que o M-1 conseguiu atingir os menores valores de média de RMSD para 3 proteínas-alvo (1AB1, 1AIL e T0820). No entanto, os menores valores mínimos de RMSD alcançados, concentraram-se entre os três métodos finais, englobando 7 proteínas-alvo. Isto indica que o procedimento de reinicialização, de certo modo, foi capaz de diversificar as multi populações, a ponto de alcançar melhores soluções.

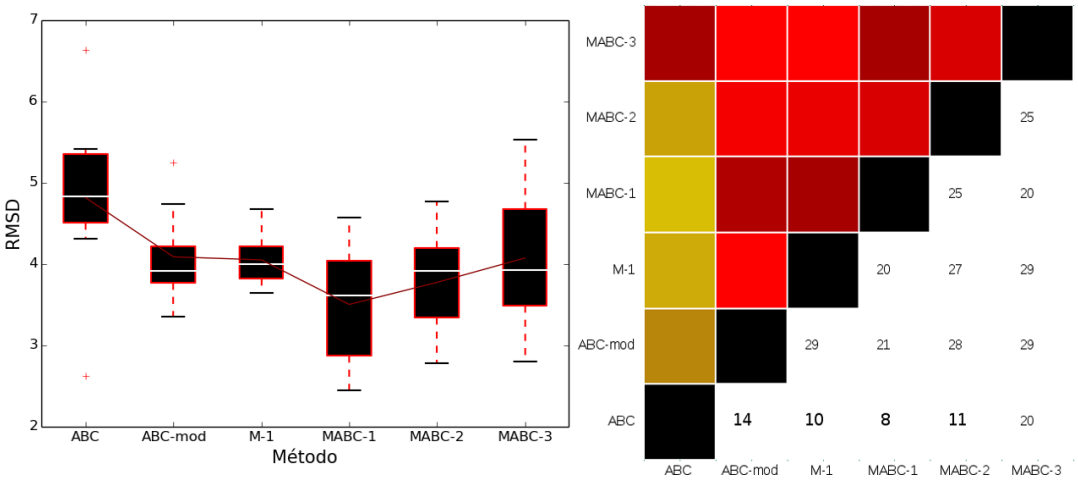
Por fim, observa-se na Tabela 6.2, que o método MABC-1, com limiar de reinicialização mais alto ($p = 15$), tendeu a obter os melhores resultados em comparação com todos os outros métodos apresentados. Este obteve o menor valor mínimo de RMSD para 6 proteínas-alvo do conjunto, e menor valor de média em relação a 3 proteínas-alvo. Isto demonstra que o AM, aliado a um procedimento de reinicialização de população com limiar de descarte mais alto, é capaz de prover melhores resultados quando comparado aos outros métodos. A versão MABC-1 tem a capacidade de prover mais diversidade as subpopulações do AM, e isto frente ao problema da multimodalidade do PSP é fundamental para se obter uma exploração mais adequada do espaço de busca conformacional, escapar de mínimos locais e em determinados momentos atingir melhores soluções.

Ainda, os resultados obtidos para cada versão de método implementada, descritos na Tabela 6.2 de resultados, foram dispostos em diagramas de caixa (*box plot*) para facilitar a visualização e comparação dos valores em relação à variação de menores valores de RMSD obtidos, mediana e média das 8 execuções de cada método, para cada

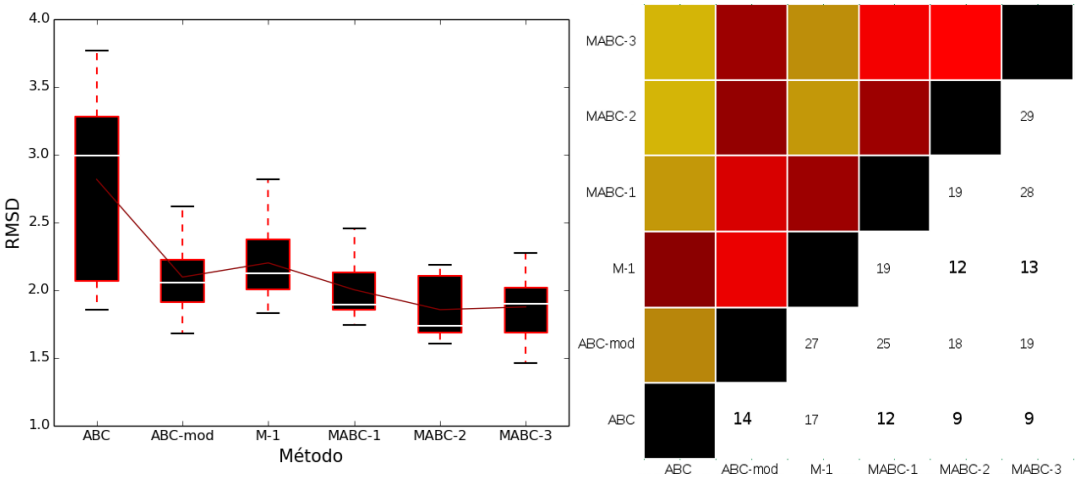
proteína-alvo. Os valores das 8 execuções de cada um dos métodos foram também comparados entre si, através do teste estatístico não-paramétrico de *Mann-Whitney* (teste U), para amostras independentes. Este teste objetiva verificar a igualdade das medianas entre duas amostras. Os valores estatísticos de U calculados no teste avaliam o grau de entrelaçamento entre duas amostras comparadas. A maior separação dos dados em conjunto indica que as amostras são distintas, rejeitando-se a hipótese de igualdade das medianas. Neste caso, uma amostra configura os 8 menores valores de RMSD obtidos a partir das 8 execuções de um método. Observa-se que por se tratar de uma amostra bastante pequena, os valores indicados pelo teste representam apenas indicativos de semelhança ou dissimilaridade. As amostras de resultados de cada um dos métodos foram comparadas em relação às amostras de todos os outros.

Com isso, a Figura 6.1 foi divida em duas partes para ilustrar os resultados de cada proteína-alvo. O lado esquerdo das imagens mostra os diagramas de caixa relativos a 8 execuções de cada um dos métodos, sendo que a linha em vermelho sobre o gráfico representa a média de RMSD destas execuções. A matriz de cores e números posicionada no lado direito das imagens descreve os resultados do teste U realizado par a par entre os métodos analisados. Observa-se, então, que para amostras de tamanho 8, conforme utilizado nestas comparações, valores de U menores do que 15 indicam que as amostras tendem a ser distintas, sendo que quanto mais distante de 15 for o valor, mais dissimilar é a amostra, enquanto que valores maiores do que 15 indicam que as amostras tendem a ser similares. Os valores de U estão indicados na parte inferior da matriz representada na imagem do lado direito da Figura 6.1, e na parte superior da matriz, estes valores de U estão mapeados em cores, onde a cor preta indica que duas amostras são totalmente iguais, as escalas em amarelo indicam que as amostras são distintas, e as escalas de cores representadas em vermelho indicam que as amostras tendem a ser similares.

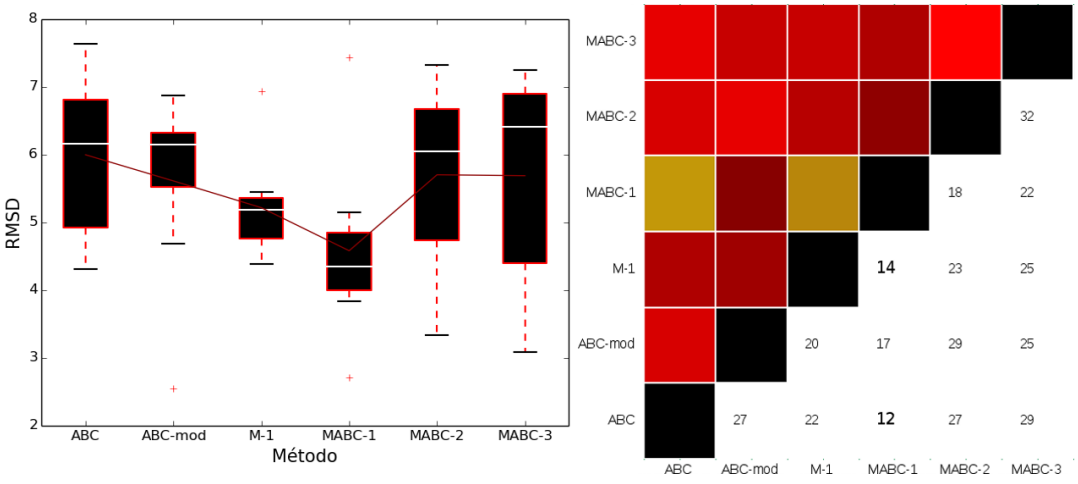
Figura 6.1: Resultados obtidos para a otimização de estruturas das proteínas-alvo do conjunto de testes. As imagens à esquerda, ilustram a representação em diagramas de caixa dos resultados obtidos a partir das 8 execuções de cada versão de método, considerando o menor valor de RMSD obtido em cada execução. A linha em vermelho sobre o gráfico representa a média de RMSD destas execuções. As imagens à direita, representam a aplicação do teste não-paramétrico de *Mann-Whitney*, através de comparações por pares dos resultados obtidos pelos métodos implementados. Valores destacados em negrito e representados em escala de amarelo denotam que os métodos são significativamente diferentes para U crítico de 15.



1AB1

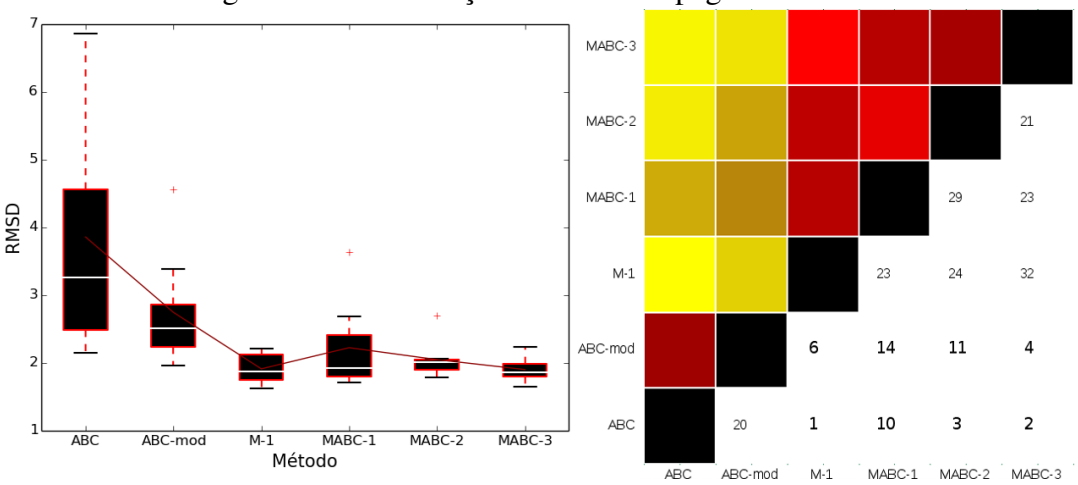


1ACW

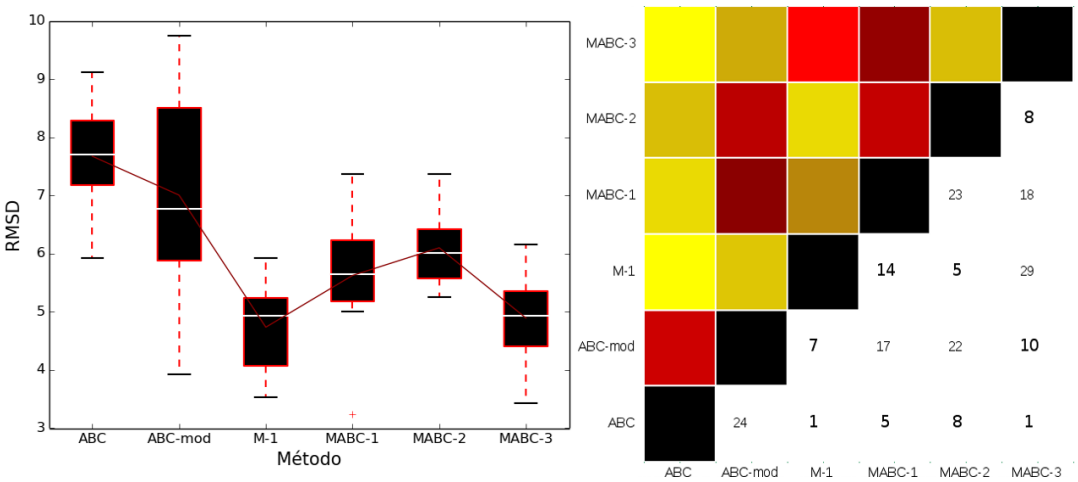


1AIL

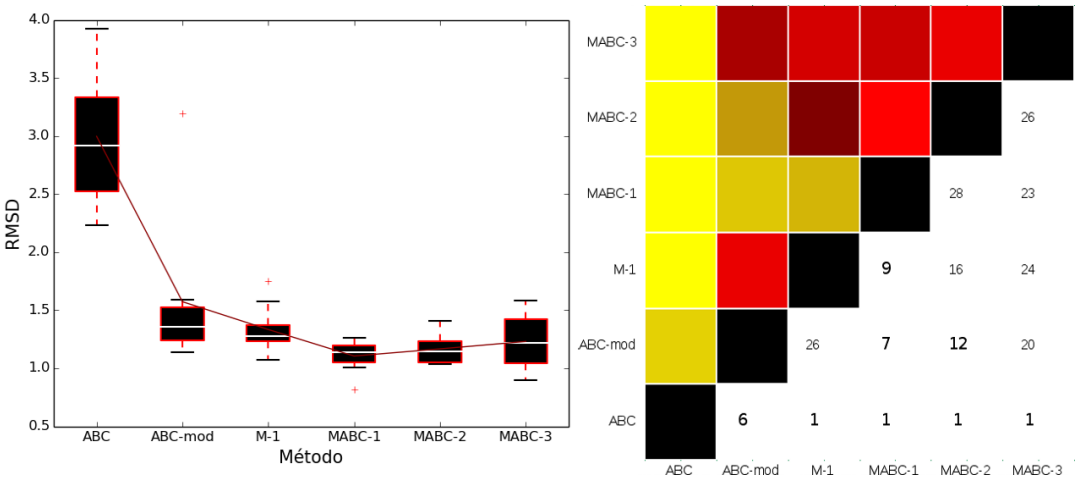
Figura 6.1: Continuação da tabela da página anterior



2MR9

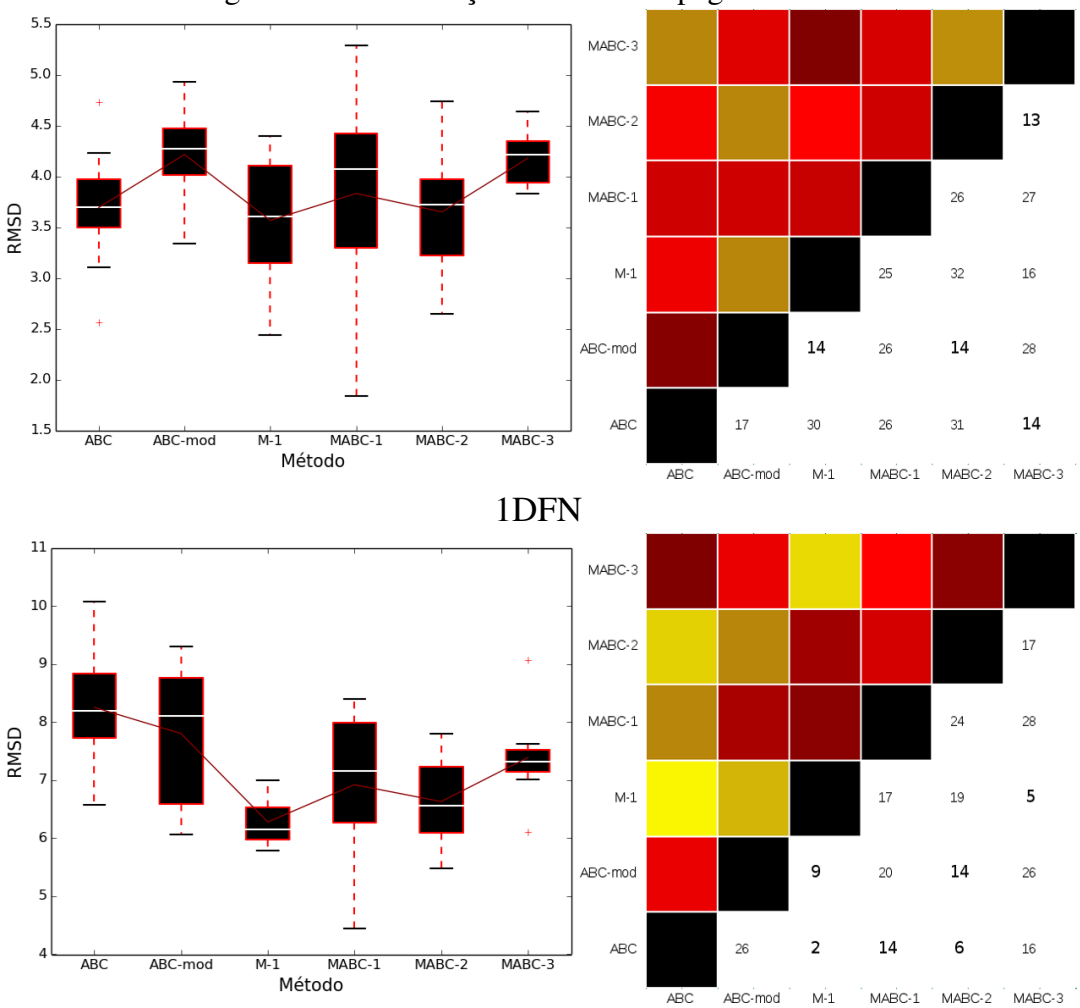


2P5K



3V1A

Figura 6.1: Continuação da tabela da página anterior



Fonte: Do autor (2017).

Analisando os diagramas de caixa representados na Figura 6.1, observa-se que os valores de RMSD assumidos por cada amostra de dados, tendem a diminuir à medida que as diferentes versões implementadas vão sendo incrementadas com novos componentes. Assim como, percebe-se que a linha vermelha nos gráficos, que denota a média dos menores valores de RMSD obtidos nas 8 execuções de cada método, também acompanha este decréscimo de RMSD, conforme as versões dos métodos evoluem. Por exemplo, para as proteínas-alvo 2P5K, 2MR9, 3V1A e T0820, este declínio amostral dos valores de RMSD é bastante perceptível. Dessa forma, pode-se dizer que os resultados ilustrados nos gráficos de caixa, corroboram com as análises realizadas anteriormente, referentes à Tabela 6.2 de resultados.

Quanto à análise de similaridades relacionadas aos resultados obtidos por cada

um dos métodos, percebe-se na Figura 6.1, que estes tenderam a não apresentar diferenças significativas em relação aos menores valores de RMSD obtidos para as 8 execuções das proteínas-alvo. Nota-se nas imagens à direita que a maior parte de cada uma das matrizes foi preenchida por diferentes escalas de vermelho, as quais indicam que duas amostras comparadas não diferem estatisticamente. No entanto, percebe-se um padrão nas cores apresentadas nestas matrizes. Os métodos ABC e ABC-mod tendem a apresentar diferenças em comparação com as outras versões de método implementadas, tornando o lado superior esquerdo das matrizes amarelo. O exemplo disto, pode ser observado nas representações de resultados da proteína 2MR9, onde os métodos ABC e ABC-mod apresentaram semelhança amostral entre eles, e diferiram significativamente de todos os outros. Observa-se, então, que as versões de métodos que constituem a abordagem de AM proposta (M-1, MABC-1, MABC-2 e MABC-3), tenderam a não apresentar diferenças significativas entre eles, porém percebe-se diferenças quando estes são comparados aos métodos ABC e ABC-mod. Sendo assim, este padrão percebido reforça a análise feita anteriormente, de que o AM e seus componentes possuem a capacidade de potencializar os resultados dos processos de otimização de estruturas de proteínas quando da comparação com utilizações de uma única meta-heurística, como por exemplo, o algoritmo ABC.

Por fim, apesar das variações do AM proposto, M-1, MABC-1, MABC-2 e MABC-3, não terem apresentado muitas diferenças significativas em relação às amostras de resultados obtidas, optou-se por utilizar como versão final da abordagem proposta, o algoritmo MABC-1, devido aos resultados mostrados na Tabela 6.2 e as análises realizadas concernentes a estes quando comparados com as outras versões.

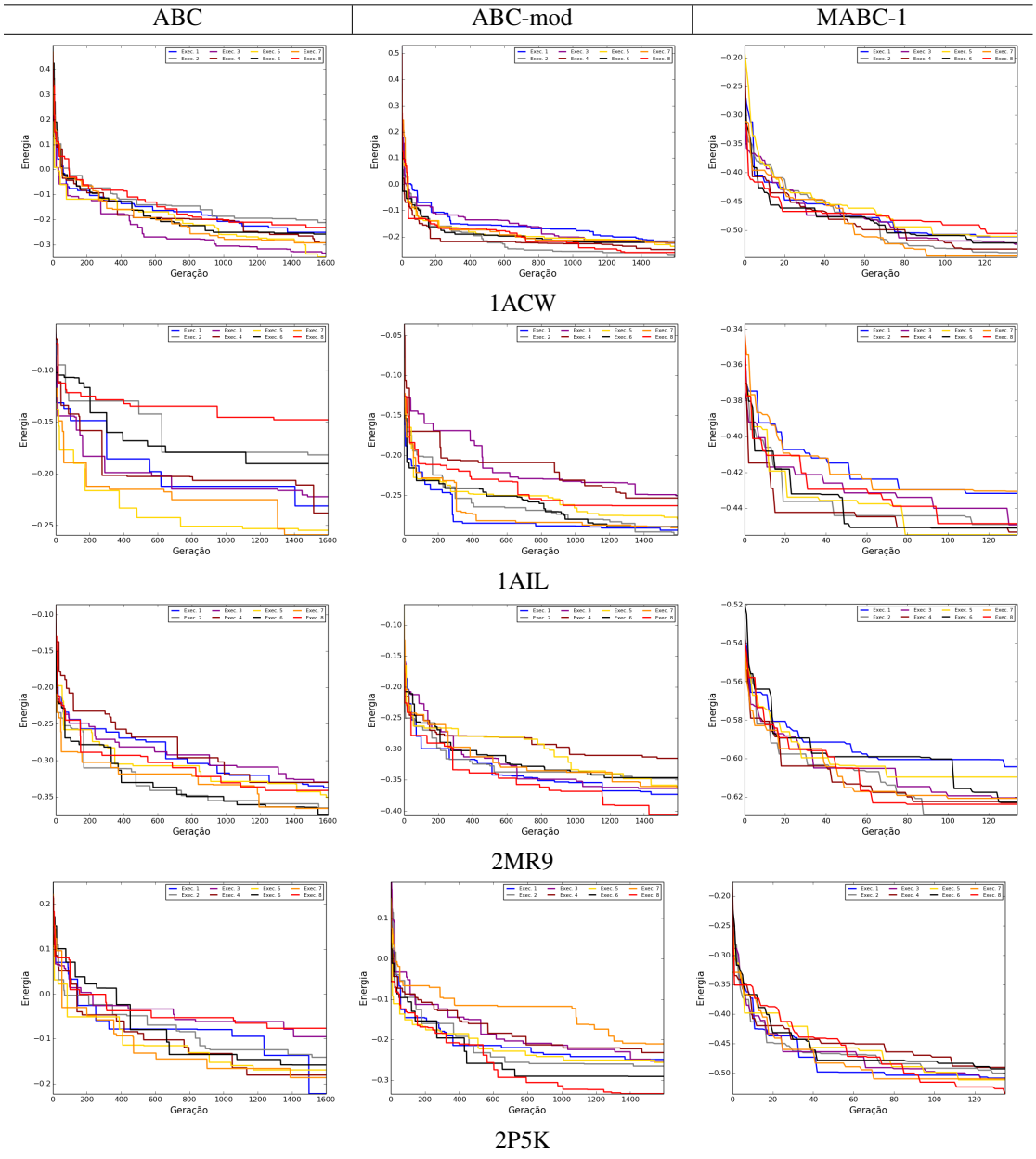
6.1.1 Análises de convergência do algoritmo

Com o objetivo de analisar a convergência do método MABC-1 em relação aos melhores valores de energia obtidos durante as execuções dos processos de otimização, estruturou-se alguns gráficos de convergência de energia, para as proteínas-alvo 1ACW, 1AIL, 2MR9 e 2P5K.

Dessa forma, a Figura 6.2 ilustra a convergência dos melhores valores de energia do método MABC-1 em comparação com os métodos ABC e ABC-mod, considerando os melhores valores de energia assumidos ao longo das gerações dos 8 processos de otimização. Nota-se que considerou-se também os métodos ABC e ABC-mod nesta análise, devido à diferença percebida entre eles na seção anterior, além de que as outras versões

do método também são baseadas nesta abordagem de AM proposta.

Figura 6.2: Convergência dos valores de energia durante os processos de otimização, relativos a 8 execuções dos métodos ABC, ABC-mod e MABC-1



Fonte: do autor (2017).

Analisando os gráficos de convergência de energia, apresentados na Figura 6.2, percebe-se que o método MABC-1, foi capaz de atingir um certo grau de convergência entre os valores finais de energia de cada execução, para a maioria das proteínas analisadas. Nota-se que esta convergência é mais perceptível nas proteínas 1ACW e 2P5K. Em relação à proteína 2MR9, observa-se que apenas duas execuções do método não conseguiram convergir conforme as outras execuções. Observa-se também, que os métodos ABC e ABC-mod tenderam a convergir em alguns casos, como no caso da otimização da

proteína-alvo 1ACW.

Ainda, observa-se a diferença nos valores de energia apresentados pelo método MABC-1 em comparação com os outros dois. Os valores de energia do MABC-1 são, em geral, para todas as proteínas-alvo, mais baixos do que os valores mostrados pelos outros métodos, e isto pode ser observado desde as primeiras gerações. Por exemplo, para a proteína-alvo 1ACW, o método MABC-1 iniciou o processo de otimização com os valores de energia das melhores soluções em torno de -0,2, enquanto os outros dois métodos apesentaram energias em torno de 0,5. Já os valores finais de energia do MABC-1, para a 1ACW, figuraram em torno de -0,5, sendo que para os métodos ABC e ABC-mod, as energias finais das melhores soluções variaram em torno de -0,2 a -0,3. Este comportamento observado em relação aos valores de energia, mantém-se em todas as outras proteínas-alvo.

Sendo assim, estas diferenças apresentadas pelos valores de energia entre os métodos, pode ser atribuída à etapa I de amostragem e classificação de indivíduos presente no algoritmo MABC-1, e ausente nas outras duas versões. Esta etapa proporciona indivíduos de melhor qualidade às populações iniciais da meta-heurística, fazendo com que antes mesmo do processo de otimização, estas assumam valores de energia mais baixos, pois possuem conformações melhores. Isto demonstra o potencial da etapa de amostragem e inicialização de indivíduos incorporada ao AM, a qual provendo soluções iniciais melhores ao método, permite que este através do processo de otimização de estruturas, alcance melhores resultados.

Finalmente, posto que a versão MABC-1 do AM, apresentada anteriormente, foi definida como sendo a abordagem final deste trabalho, na próxima seção serão apresentados os resultados obtidos quando da otimização do conjunto de proteínas-alvo descrito na Tabela 6.1. Tendo em vista a relevância do problema PSP para a área de Bioinformática Estrutural, estes resultados também serão comparados com abordagens ditas estado da arte para a área de predição de estruturas 3-D de proteínas voltada à categoria FM.

6.2 Otimizações finais do método proposto

Esta seção tem por objetivo realizar as últimas análises concernentes ao AM de otimização proposto neste trabalho. Para tanto, serão apresentados os resultados obtidos pelo método quando aplicado na otimização de um conjunto maior de proteínas-alvo. Dentre as diferentes versões de métodos apresentadas e analisadas na seção anterior, definiu-se

que o método final a ser abordado nesta seção, será o algoritmo memético MABC-1. Para fins de simplificação, este será referido como MABC.

Dessa forma, para avaliar o potencial do AM proposto, frente a um conjunto maior de proteínas-alvo, bem como situá-lo em relação aos métodos de referência na área de predição de estruturas 3-D de proteínas, o MABC foi aplicado na otimização do conjunto de testes de proteínas-alvo descrito na Tabela 6.1, o qual engloba 24 diferentes sequências de aminoácidos. Os resultados obtidos a partir destas otimizações, foram comparados com os métodos estado da arte Rosetta e QUARK. Estes dois métodos foram escolhidos de acordo com os resultados relacionados às últimas edições do CASP (TAI et al., 2014; KINCH et al., 2016a), os quais os postularam como sendo os métodos mais relevantes para a área, no que se refere a técnicas automáticas de predição FM, sem intervenção manual (*servers*), devido à sequência de melhores resultados alcançados.

Os algoritmos MABC e Rosetta foram executados 30 vezes para cada proteína-alvo do conjunto de testes definido (Tab. 6.1). O MABC utilizou como critério de parada o valor máximo de 1.000.000 de cálculos de energia. O método Rosetta[2] foi executado a partir da versão de software disponível *offline* (*Rosetta commons, Academic License, Version 3.4*) (ROHL et al., 2004). Para a execução do Rosetta, utilizou-se as configurações padrões do método recomendadas pelos desenvolvedores. O processo de execução do método QUARK foi realizado através do servidor online[3] fornecido pelos seus desenvolvedores, voltado à categoria FM. No entanto, nota-se que a submissão de sequências de aminoácidos a serem modeladas no sistema deve ser realizada de forma manual, sendo que o mesmo aceita que apenas uma única proteína-alvo seja submetida por vez. Devido as dificuldades apresentadas por esta dinâmica de processamento de sequências de aminoácidos, o QUARK foi executado uma única vez para cada proteína-alvo. Foram utilizadas as configurações padrões de execução do servidor.

Os resultados obtidos pelos métodos descritos, foram comparados através das métricas RMSD e GDT_TS em relação às estruturas determinadas experimentalmente das proteínas-alvo. A métrica de avaliação estrutural GDT_TS (ZHANG; SKOLNICK, 2004) objetiva avaliar a semelhança estrutural entre duas estruturas 3-D de proteínas, assim como o RMSD. Contudo, esta métrica pode ser considerada um tanto mais robusta, devido ao fato de não ser tão sensível à regiões discrepantes da proteína, como as regiões que compreendem as ESs irregulares (KUFAREVA; ABAGYAN, 2012). Nota-se que diferente do RMSD, o GDT_TS é uma métrica de maximização, sendo que as soluções

[2]<www.rosettacommons.org>

[3]<www.zhanglab.ccmb.med.umich.edu/QUARK>

mais similares apresentam valores mais elevados (GDT_TS igual a 100,0 indica que as estruturas são idênticas). O GDT_TS pode ser expresso pela Equação 6.1.

$$GDT_{TS} = \frac{(GDT_{P1} + GDT_{P2} + GDT_{P4} + GDT_{P8})}{4} \tag{6.1}$$

Onde GDT_{Pn} representa a percentagem de resíduos de aminoácidos sob o limiar de distância $\leq n$.

Observa-se que a melhor solução considerada em cada uma das execuções do algoritmo MABC, foi determinada conforme explicado na seção anterior. As 13 melhores soluções, cada uma oriunda de uma subpopulação do AM e determinadas através dos menores valores de energia, foram comparadas quanto aos valores de RMSD assumidos em relação à estrutura experimental de uma determinada proteína-alvo. A solução que apresentou o menor valor de RMSD foi escolhida como sendo a melhor solução, para dada execução. Quanto ao Rosetta, ao final da execução do processo de otimização de uma sequência de aminoácidos, o método retorna um conjunto contendo 10 soluções finais possíveis. Com isso, estas 10 soluções foram comparadas em relação aos valores de RMSD apresentados, sendo que a de menor valor foi escolhida como sendo a melhor solução desta execução. Da mesma forma, o método QUARK também retorna um conjunto com 10 soluções finais possíveis. O mesmo procedimento descrito acima foi aplicado para a determinação da melhor solução, dentre as 10 retornadas.

Sendo assim, a Tabela 6.3 resume os valores de RMSD e GDT_TS obtidos para as 30 execuções dos algoritmos MABC e Rosetta aplicados ao conjunto de testes de 24 proteínas-alvo. A tabela também descreve os valores de RMSD e GDT_TS para o método QUARK, sendo que este foi aplicado uma única vez em cada proteína. As amostras de melhores resultados para cada proteína-alvo obtidas pelos métodos MABC e Rosetta foram comparadas através do teste não-paramétrico de *Mann-Whitney*, utilizado na seção anterior. Estas comparações visam estabelecer a semelhança entre os resultados amostrais obtidos. O nível de significância utilizado foi de 5%. Nota-se que o valor-*p* resultante destas comparações determina a similaridade entre duas amostras. Valores acima de 5% rejeitam a hipótese de similaridade, e estas são consideradas diferentes estatisticamente. Com isso, nota-se que as amostras que diferiram significativamente foram destacadas com o símbolo "+" na Tabela 6.3.

Tabela 6.3: Resumo dos resultados obtidos em relação ao RMSD e GDT_TS para as 30 execuções dos algoritmos MABC e Rosetta. Os resultados de RMSD e GDT_TS referentes à única execução do QUARK estão ilustrados na última coluna da tabela. O símbolo "*" denota os casos em que o QUARK foi superior aos outros dois métodos. A coluna Valor-*p* representa o resultado de *p* oriundo da comparação de similaridade realizado através do teste não-paramétrico de *Mann-Whitney* entre os valores de RMSD e GDT_TS obtidos pelo MABC e o Rosetta. Sendo *p* significante em 5% ($p < 0,05$), o símbolo "+" denota as amostras que diferiram significativamente

ID-Proteína	**RMSD (Å)**			**GDT (%)**			**QUARK**	
	Mín.	**Médio**	**Valor-*p***	**Máx.**	**Médio**	**Valor-*p***	**RMSD (Mín)**	**GDT (Máx.)**
1AB1-MABC	2,56	3,75 ± (0,54)	0,07	75,00	66,74 ± (3,93)	0,04 +	2,91	72,28
1AB1-Rosetta	**2,29**	**3,56** ± (0,72)		**82,07**	**68,73** ± (4,98)			
1ACW-MABC	**1,43**	1,90 ± (0,23)	4,33e-05 +	81,03	75,43 ± (2,80)	1,92e-03 +	5,79	47,41
1ACW-Rosetta	1,44	**1,66** ± (0,18)		**82,76**	**77,82** ± (2,60)			
1AIL-MABC	**2,72**	**5,33** ± (1,44)	5,14e-07 +	68,57	**57,42** ± (5,05)	3,66e-09 +	0,62*	95,36*
1AIL-Rosetta	3,24	7,83 ± (1,73)		**78,57**	46,67 ± (6,99)			
1CRN-MABC	**2,31**	**3,69** ± (0,73)	0,46	76,09	67,68 ± (4,41)	1,39e-04 +	2,64	73,37
1CRN-Rosetta	2,45	**3,69** ± (0,63)		**77,17**	**72,01** ± (3,22)			
1D5Q-MABC	**1,00**	1,78 ± (0,39)	0,03 +	**88,89**	81,64 ± (3,98)	0,07	1,97	81,48
1D5Q-Rosetta	1,12	**1,58** ± (0,24)		**88,89**	**83,24** ± (2,75)			
1DFN-MABC	**2,28**	**3,88** ± (0,56)	2,32e-05 +	**59,17**	**50,33** ± (3,07)	3,91e-05 +	3,58	55,83
1DFN-Rosetta	3,13	4,75 ± (0,73)		53,33	46,89 ± (2,76)			
1ENH-MABC	1,79	2,76 ± (0,38)	1,91e-10 +	**50,46**	**46,90** ± (1,96)	0,27	1,55	46,30
1ENH-Rosetta	**0,75**	**1,66** ± (0,40)		49,54	46,59 ± (1,10)			
1FNA-MABC	10,41	12,57 ± (1,13)	2,25e-11 +	18,68	16,61 ± (1,02)	0,11	3,10*	20,33
1FNA-Rosetta	**3,25**	**7,12** ± (1,95)		**21,15**	**17,06** ± (1,42)			
1K43-MABC	**0,36**	**0,64** ± (0,11)	1,39e-05 +	**92,86**	**88,27** ± (1,99)	4,52e-10 +	-	-
1K43-Rosetta	0,58	0,77 ± (0,08)		87,50	83,69 ± (1,71)			
1L2Y-MABC	0,94	1,30 ± (0,26)	2,36e-04 +	88,75	83,08 ± (3,68)	2,87e-05 +	2,59	77,50
1L2Y-Rosetta	**0,62**	**1,03** ± (0,17)		**97,50**	**87,83** ± (3,78)			
1OPD-MABC	8,09	9,94 ± (0,93)	5,54e-07 +	36,18	32,48 ± (2,02)	1,48e-11 +	3,92	65,88
1OPD-Rosetta	**2,71**	**6,99** ± (2,31)		**72,94**	**51,56** ± (10,58)			
1Q2K-MABC	1,16	1,92 ± (0,54)	3,89e-09 +	86,29	76,96 ± (5,29)	6,80e-10 +	7,18	36,29
1Q2K-Rosetta	**0,51**	**0,98** ± (0,35)		**97,58**	**90,65** ± (5,31)			
1ROP-MABC	1,51	**1,85** ± (0,23)	5,54e-07 +	83,48	**78,23** ± (2,96)	0,08	1,21	87,50
1ROP-Rosetta	**0,95**	2,88 ± (1,09)		**90,62**	76,44 ± (6,27)			
1UTG-MABC	3,67	**5,49** ± (1,38)	0,45	58,21	51,39 ± (4,23)	0,08	3,71	58,93
1UTG-Rosetta	**3,04**	6,03 ± (2,56)		**63,57**	**53,01** ± (7,18)			
1WQC-MABC	2,00	2,64 ± (0,47)	2,54e-10 +	76,92	70,29 ± (3,11)	5,68e-07 +	3,11	68,27
1WQC-Rosetta	**1,62**	**1,93** ± (0,16)		**78,85**	**74,55** ± (2,35)			
1ZDD-MABC	1,34	1,95 ± (0,34)	1,51e-11 +	**45,59**	**44,36** ± (0,67)	6,17e-06 +	0,60*	44,85
1ZDD-Rosetta	**0,72**	**0,93** ± (0,12)		44,12	43,63 ± (0,40)			
2MR9-MABC	1,55	2,08 ± (0,29)	1,98e-08 +	85,23	73,92 ± (3,90)	8,09e-08 +	1,39	85,23
2MR9-Rosetta	**1,11**	**1,56** ± (0,25)		**88,64**	**82,35** ± (4,89)			
2MTW-MABC	**1,44**	**1,97** ± (0,22)	1,51e-11 +	**80,00**	**77,17** ± (1,37)	2,78e-10 +	4,21	66,25
2MTW-Rosetta	2,55	3,54 ± (0,52)		77,50	67,71 ± (5,02)			
2P5K-MABC	3,36	5,19 ± (0,95)	1,51e-11 +	48,02	42,08 ± (2,61)	1,46e-11 +	2,70	53,57
2P5K-Rosetta	**0,70**	**1,52** ± (0,29)		**56,35**	**53,62** ± (1,32)			
2P6J-MABC	2,24	3,04 ± (0,58)	3,65e-04 +	73,56	62,88 ± (4,58)	1,59e-05 +	5,38	59,62
2P6J-Rosetta	**2,17**	**2,56** ± (0,22)		**74,52**	**67,98** ± (3,02)			
2P81-MABC	**3,30**	**4,45** ± (0,57)	4,88e-10 +	**39,20**	**37,58** ± (0,76)	2,79e-11 +	4,00	34,66
2P81-Rosetta	4,67	5,77 ± (0,45)		36,93	35,19 ± (0,79)			
2PMR-MABC	2,02	2,89 ± (0,52)	9,98e-06 +	**51,97**	**47,60** ± (2,11)	9,44e-03 +	0,96	49,34
2PMR-Rosetta	**0,87**	**2,10** ± (0,65)		50,66	46,38 ± (1,92)			
3V1A-MABC	0,82	1,25 ± (0,24)	3,56e-09 +	**63,02**	**58,72** ± (2,23)	2,11e-10 +	0,65	55,73
3V1A-Rosetta	**0,61**	**0,79** ± (0,26)		55,73	55,11 ± (0,39)			
T0820-MABC	**3,49**	**6,89** ± (1,19)	7,21e-04 +	**51,11**	42,96 ± (3,20)	0,35	7,76	43,89
T0820-Rosetta	6,68	7,81 ± (0,89)		47,78	**43,04** ± (2,84)			
Resumo	37,50% (9/24)	37,50% (9/24)	-	41,67% (10/24)	41,67% (10/24)	-	54,17% (13/24)	66,67% (16/24)

Fonte: Do autor (2017).

Analisando os resultados apresentados na Tabela 6.3, percebe-se que o AM proposto neste trabalho, foi capaz de atingir valores comparáveis ao Rosetta e ao QUARK, e predizer estruturas de proteínas com conformações similares às estruturas determinadas experimentalmente. Observa-se que as estruturas preditas podem ser consideradas similares a estruturas experimentais quando apresentam valores de RMSD ≤ 4Å, devido ao fato de que as estruturas cristalográficas no PDB representam um estado momentâneo da proteína, sendo que esta apresenta-se em movimento (CARUGO, 2003). Este limiar de 4Å também foi adotado nas análises concernentes ao método proposto no trabalho de Corrêa et al. (2016).

Sendo assim, nota-se que para 17 proteínas-alvo dentre as 24 analisadas, o MABC conseguiu alcançar valores médios de RMSD abaixo de 4Å. Destas 17 proteínas, para 9 estudos de caso, o RMSD médio figurou abaixo de 2Å. Observa-se que estes resultados equiparam-se aos do Rosetta, visto que este também obteve valores médios de RMSD abaixo de 4Å para 17 proteínas, das quais 11 apresentaram valores menores do que 2Å. Em relação aos valores mínimos de RMSD obtidos, é possível perceber que o método MABC alcançou valores abaixo de 4Å para 22 proteínas-alvo, sendo que para 12 destas os valores obtidos foram menores do que 2Å. Estendendo esta análise para o Rosetta, verifica-se novamente que o método também alcançou valores menores do que 4Å para 22 proteínas-alvo, sendo que em 13 casos os valores ficaram abaixo de 2Å.

Com isso, pode-se afirmar que estes resultados destacam o potencial do MABC em predizer estruturas 3-D de proteínas que se assemelham a estruturas determinadas experimentalmente, e aos resultados obtidos pelo método Rosetta. Contudo, entende-se que melhoramentos ainda precisam ser realizados para que estes números sejam melhorados. Por exemplo, as duas proteínas-alvo (1FNA e 1OPD) em que o MABC não foi capaz de atingir valores mínimos de RMSD menores do que 4Å, delineiam possíveis pontos fracos do método que precisam ser abordados. Nota-se que este caso, especificamente, será discutido na sequência do texto.

Quanto à comparação com os resultados gerados pelo Rosetta, observa-se que o AM atingiu melhores valores mínimos e médios de RMSD em 9 das 24 proteínas testadas (37,5%). Entretanto, é possível perceber que ambos os métodos obtiveram valores tanto de média de RMSD, quanto de mínimo de RMSD, bastante similares, visto que os métodos apresentaram diferenças menores do que 1Å em 17 proteínas-alvo, o que caracteriza 71% do conjunto de testes. Dentre as proteínas-alvo testadas, em apenas 3 casos os resultados amostrais mostraram-se significativamente semelhantes, como pode ser ob-

servado nas proteínas: 1AB1, onde o Rosetta obteve a melhor média de RMSD; 1UTG, onde o MABC superou o Rosetta quanto a melhor média de RMSD; e 1CRN, onde ambos obtiveram a mesma média.

Este comportamento também é percebido ao analisarmos os valores de GDT_TS mostrados na Tabela 6.3. O método MABC conseguiu obter melhores resultados do que o Rosetta quanto aos valores máximos e médios de GDT_TS em 10 proteínas-alvo (41,67%). Esta diferença de 1 proteína percebida entre as métricas RMSD e GDT_TS, está relacionada ao fato do GDT_TS não ser tão sensível a pequenas mudanças na estrutura da proteína. Porém, de forma geral, observa-se que ambos os métodos também atingiram valores similares de GDT_TS. Em 6 casos, os resultados amostrais apresentaram similaridades significativas estatisticamente.

Em comparação com os resultados exibidos pelo QUARK, nota-se que o MABC atingiu melhores valores de menores RMSD para 13 proteínas-alvo (54,17%), e obteve melhores valores de GDT_TS para 16 proteínas (66,67%). No entanto, ressalta-se que o QUARK foi executado uma única vez para cada proteína-alvo. Dessa forma, estes resultados são indicativos de que o MABC tende a gerar estruturas similares ou melhores do que o QUARK.

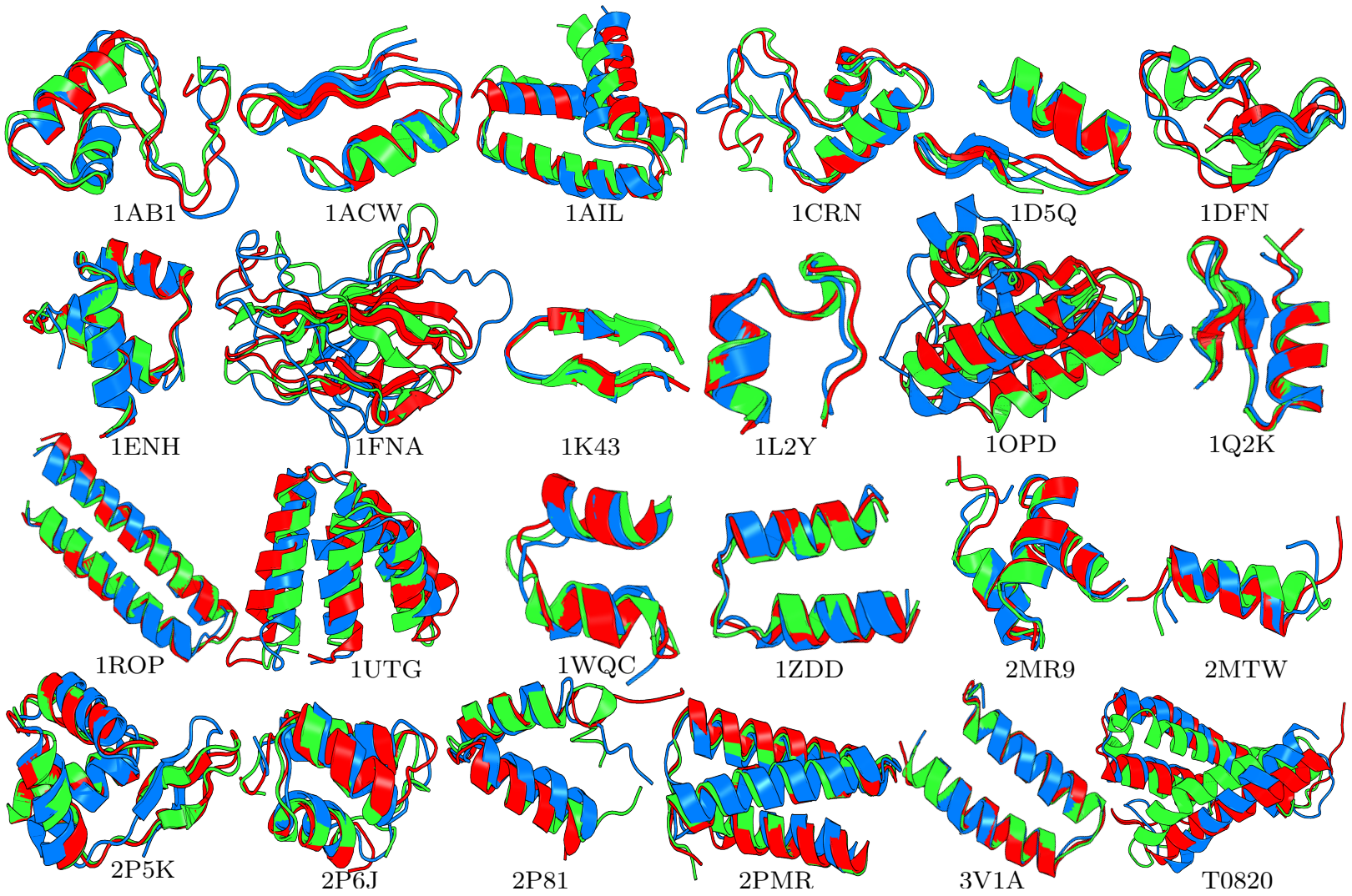
Analisando, em particular, os resultados alcançados para a proteína-alvo T0820, utilizada nos experimentos de predição de estruturas do CASP11 (KINCH et al., 2016b) e sendo a segunda maior proteína do conjunto de testes (Tab. 6.1), observa-se que o algoritmo MABC conseguiu atingir resultados de RMSD que superam os valores obtidos pelo Rosetta e o QUARK. Esta proteína é composta apenas de hélices e apresenta uma conformação bem empacotada, considerando o seu tamanho. Embora todas as proteínas do conjunto de testes assumam o caráter de serem empacotadas, observa-se que proteínas maiores implicam em maiores dificuldades de posicionamento correto de estruturas regulares e empacotamento. Outro exemplo que pode ser referido, é a proteína 2PMR, onde o MABC obteve valores médios de RMSD menores do que 3Å. Estes são dois casos de proteínas maiores com conformações compactas. Diante disso, destaca-se a capacidade do método em modelar corretamente conformações de proteínas mais empacotadas.

Por outro lado, destaca-se aqui dois casos, onde o MABC não conseguiu predizer de forma adequada as estruturas 3-D das proteínas. Para as proteínas 1FNA e 1OPD, o método atingiu valores médios de RMSD bastante elevados, sendo que os valores mínimos também acompanharam estes resultados. Nestes casos, o Rosetta embora tenha obtido média de RMSD superior a 4Å em ambas as proteínas, foi capaz de obter valo-

res mínimos de RMSD abaixo de 4Å. Da mesma forma, o QUARK também conseguiu atingir valores menores do que 4Å, para ambas as proteínas, em apenas uma única execução. Nota-se que estas duas proteínas configuram as maiores sequências de aminoácidos presentes no conjunto de testes, juntamente com a proteína T0820. Estas apresentam conformações bastante difíceis de serem preditas. A proteína 1FNA é composta apenas de folhas-β e regiões irregulares, e a 1OPD é uma proteína híbrida, em grande parte, constituída também por folhas. O problema da modelagem destas estruturas, se dá devido à dificuldade em modelar regiões irregulares da proteína, sendo que estas são as responsáveis pelo posicionamento correto das cadeias polipeptídicas que possibilitam a formação de folhas. Isto demonstra que, dependendo da complexidade das relações intermoleculares dos aminoácidos da proteína, apenas atingir o empacotamento ideal não é o suficiente. Neste sentido, entende-se que o MABC precisa ser melhorado quanto a modelagem de estruturas maiores, constituídas, em sua maioria, de folhas e regiões irregulares.

Por fim, objetivando visualizar as conformações assumidas pelas estruturas de menores valores de RMSD obtidas pelos métodos MABC e Rosetta, estas foram sobrepostas às suas respectivas estruturas experimentais. Com isso, a Figura 6.3 ilustra a representação gráfica das estruturas de proteínas preditas que obtiveram os menores valores de RMSD, sobrepostas às estruturas determinadas experimentalmente (vermelho), relativas aos métodos MABC (azul) e Rosetta (verde).

Figura 6.3: Representação gráfica da sobreposição das estruturas determinadas experimentalmente (vermelho), e de menores valores de RMSD obtidas pelos métodos MABC (azul) e Rosetta (verde). As estruturas 3-D preditas foram ajustadas em relação às experimentais considerando os átomos de C_α



Fonte: do autor (2017).

Analisando as estruturas 3-D das proteínas-alvo ilustradas na Figura 6.3, é possível verificar visualmente que o método proposto conseguiu alcançar conformações similares às estruturas determinadas experimentalmente. Além disso, observa-se que o método também atingiu conformações muito próximas às estruturas preditas pelo Rosetta, sendo que em alguns casos, o superou. A partir desta representação gráfica, constata-se a capacidade do método em atingir conformações empacotadas de acordo com os enovelamentos experimentais. Por exemplo, para as únicas duas proteínas-alvo em que o MABC não foi capaz de obter valores mínimos de RMSD menores do que 4Å, 1FNA e 1OPD, destaca-se que, de maneira geral, as conformações apresentadas conseguiram atingir o nível de empatamento correto, porém este enovelamento não ocorreu da forma esperada. Isto reforça o que foi posto anteriormente, que dependendo da natureza e complexidade da sequência de aminoácidos, apenas atingir o empacotamento ideal não é o suficiente.

Diante disto, percebe-se a capacidade do método em modelar corretamente conformações de proteínas, as quais se assemelhem a estruturas experimentais. No entanto, reforça-se a necessidade de adaptar os métodos de busca ao problema, visando corrigir pontos francos e deficiências dos mesmos através da investigação de características específicas das estruturas trabalhadas.

Finalmente, a partir das análises realizadas relativas aos resultados ilustrados na Tabela 6.3, e do que foi posto acerca das representações gráficas das estruturas 3-D de proteínas ilustradas na Figura 6.3, conclui-se que o método MABC pode ser considerado uma contribuição efetiva para a área de predição de estruturas 3-D de proteínas, contudo, entendendo as suas limitações, sabe-se que este ainda necessita de melhoramentos para que os resultados obtidos possam ser melhorados.

6.3 Resumo do capítulo

Este capítulo apresentou os resultados obtidos concernentes ao método de otimização proposto quando da otimização de um conjunto de testes de proteínas-alvo. A meta-heurística desenvolvida consiste da etapa I de amostragem e inicialização de indivíduos, incorporada ao AM multi populacional de otimização.

Com isso, o capítulo foi estruturado em duas partes. A primeira parte reuniu experimentos de otimização realizados com o objetivo de analisar as diferentes adaptações ocorridas no algoritmo ao longo da execução do trabalho, através da implementação de diferentes versões do método proposto. Também visou definir o valor mais adequado ao

parâmetro p, relacionado à diversidade de soluções, empregado na etapa de reinicialização das subpopulações da meta-heurística. Dessa forma, a partir dos resultados obtidos e das análises realizadas relativas às diferentes versões de métodos apresentadas nesta primeira parte, verificou-se que o método MABC-1 com limiar de reinicialização mais elevado ($p = 15$), tendeu a obter os melhores resultados em comparação com todos os outros métodos. Por esta razão, definiu-se o algoritmo MABC-1 como sendo a abordagem final a ser empregada nos experimentos da segunda parte do capítulo.

Sendo assim, a segunda parte do capítulo objetivou testar a abordagem final, definida como sendo o algoritmo memético MABC-1, por meio da otimização de um conjunto de 24 proteínas-alvo, bem como compará-lo a dois métodos de referência na área de predição de estruturas 3-D de proteínas, os quais consistiram nos métodos Rosetta e QUARK.

A partir dos resultados obtidos e das comparações realizadas entre o MABC-1, e os métodos Rosetta e QUARK, em relação às estruturas determinadas experimentalmente, constatou-se que o algoritmo proposto neste trabalho, foi capaz de atingir valores comparáveis ao Rosetta e ao QUARK, sendo que em alguns casos, os superou. Acrescenta-se a isso, o êxito obtido quanto à predição de estruturas 3-D de proteínas com conformações similares às estruturas determinadas experimentalmente. Concluiu-se, então, que o método MABC-1 pode ser considerado uma contribuição efetiva para a área de predição de estruturas 3-D de proteínas, mas que ainda necessita de melhoramentos para que os resultados obtidos possam ser melhorados.

7 CONCLUSÕES

Apesar dos avanços dos métodos computacionais para a área de predição de estruturas 3-D de proteínas, sabe-se que ainda existe uma demanda crescente pela exploração de novas estratégias que objetivem extrair e manipular dados estruturais oriundos de estruturas determinadas experimentalmente. Posto ainda que atualmente não existem métodos capazes de obter a solução ótima para o problema PSP, o desenvolvimento de novos métodos robustos, a adaptação e investigação destas metodologias, visando reunir e aplicar todas estas informações experimentais disponíveis, de maneira eficaz nos processos de predição, são claramente uma necessidade, além de se mostrar um relevante campo de pesquisa relacionado à Bioinformática Estrutural.

Ainda, sabendo que o sucesso da predição de estruturas 3-D de proteínas voltada à categoria FM, requer uma função de energia acurada, aliada a um método de busca eficiente na exploração e manutenção da diversidade do espaço de soluções, e a estratégias de incorporação de conhecimento acerca de estruturas determinadas experimentalmente que busquem contornar as adversidades impostas pela complexidade da função de avaliação, acredita-se que através do estudo e desenvolvimento de meta-heurísticas voltadas aos problemas complexos de otimização e a adaptação das mesmas para lidar com as questões endereçadas pelo PSP, seja possível extrair o potencial máximo dos métodos de busca, enquanto obtêm-se melhores resultados para o problema.

Neste sentido, foi proposto nesta dissertação um método de otimização voltado ao problema PSP. O método consiste em um AM multi populacional baseado em conhecimento. O algoritmo em questão, foi divido em duas partes principais de processamento: (*i*) etapa de amostragem e inicialização de indivíduos; e (*ii*) etapa de otimização dos modelos estruturais provenientes da etapa anterior. Serão apresentadas abaixo as considerações finais sobre o método proposto, a partir dos resultados e análises concernentes a estes, realizados anteriormente.

A criação da etapa de amostragem e classificação de indivíduos objetivou a geração e classificação de diversos modelos estruturais para a proteína-alvo, buscando a definição de diferentes grupos estruturais e a criação de melhores estruturas para serem incorporadas à meta-heurística como soluções iniciais das multi populações de otimização. Esta etapa concebeu alguns procedimentos necessários à sua execução, bem como análises relacionadas ao desempenho e potencial dos diferentes componentes incorporados:

- O primeiro procedimento consiste na amostragem e inicialização de 10.000 indivíduos a partir da técnica APL-combinada (APL-vizinhança e APL-centroide) e seus subtipos, utilizada para restringir o espaço conformacional de aminoácidos, considerando as probabilidades de ocorrências e preferências conformacionais de aminoácidos em estruturas previamente conhecidas. Em relação à este procedimento, realizou-se amostragens de indivíduos para cada proteína-alvo do conjunto de testes definido, com a finalidade de avaliar: (*i*) o potencial dos diferentes tipos de APLs disponíveis; (*ii*) o comportamento dos indicadores de empacotamento estrutural (RG e SASA); e (*iii*) o comportamento de três diferentes funções de energia quanto à capacidade de representar boas estruturas. Concluiu-se, então:

  1. Verificou-se que não há diferenças entre os diferentes tipos de APLs analisados quanto à qualidade dos indivíduos gerados. Nota-se que todos os tipos de APLs tenderam a gerar soluções de qualidade similares, porém optou-se pela utilização da APL-combinada na etapa de amostragem de indivíduos do método proposto, visto a maior disponibilidade de dados experimentais;
  2. Constatou-se que as métricas de avaliação estrutural de RG e SASA podem ser utilizadas como bons indicadores do nível de empacotamento de estruturas de proteínas. A partir dos processos de amostragem, verificou-se a evidente correlação entre estruturas com RG menores e RMSD melhores. Quanto ao SASA, percebeu-se que esta relação mostrou-se menos evidente. Contudo, a preferência por valores menores de SASA ainda foi considerada válida, devido ao fato de que a variação de qualidade (intervalo de RMSD) das estruturas que assumiram valores mais elevados foi bem maior quando comparados aos indivíduos de menores SASA;
  3. Observou-se que nenhuma das três funções de energia analisadas foram capazes de representar a relação esperada entre energia estrutural e RMSD. As funções de energia analisadas foram a função de energia composta definida neste trabalho, a função de energia *Talaris2014* do Rosetta, e a função de energia composta, tendo cada um dos três termos normalizados; Embora nenhuma das três funções de energia tenha sido capaz de refletir de forma satisfatória esta relação, percebeu-se que o processo de normalização dos termos conduziu a uma maior diferenciação entre os modelos estruturais gerados, e reduziu o intervalo de variação dos valores de RMSD relativos às regiões de menores energias, implicando na disposição de menos soluções com valores mais bai-

xos de energia, porém com as mesmas qualidades. Por esta razão, a função de energia de termos normalizados foi escolhida para ser utilizada como função objetivo do método proposto, dada a sua maior capacidade de diferenciar modelos estruturais.

- O segundo procedimento desta etapa consiste no processo de filtragem de soluções oriundas do procedimento anterior de amostragem de indivíduos, e objetiva desconsiderar do processo de otimização estruturas consideradas ruins, como forma de prevenir que estas sejam contabilizadas no processo de definição dos grupos estruturais que serão utilizados na inicialização das populações do AM. Definiu-se como estruturas ruins, os modelos estruturais desprovidos de empacotamento, sendo que esta falta de empacotamento foi verificada através das métricas de RG e SASA. Concluiu-se que o procedimento de filtragem de soluções possui o potencial de descartar soluções ruins, enquanto garante a manutenção de boas soluções, visto que a estratégia conseguiu reduzir de forma significativa o número de soluções geradas pela amostragem de soluções através do processo de descarte;
- O último procedimento desta etapa consiste no agrupamento de soluções oriundas do processo de filtragem de indivíduos. Os diferentes grupos estruturais identificados no processo de agrupamento, são utilizados como forma de prover indivíduos iniciais às multi populações do AM, visando a diversidade das populações, na tentativa de contornar os problemas da multimodalidade da função objetivo, ao mesmo tempo que preocupa-se com a qualidade das soluções iniciais. Constatou-se que a estratégia de agrupamento de soluções conseguiu identificar diferentes grupos estruturais oriundos do processo de filtragem de soluções, e que a classificação dos grupos formados em decorrência do agrupamento seria realizada pela métrica de RG médio de cada grupo, a qual demostrou um balanceamento ideal entre identificação de diferentes grupos estruturais e seleção de soluções de melhores qualidades.

Dessa forma, conclui-se que a etapa de amostragem e inicialização de indivíduos possui o potencial de ser utilizada como forma de identificar diferentes grupos estruturais e prover soluções diversificadas e de mais qualidade para serem utilizadas como indivíduos iniciais das populações da meta-heurística de otimização.

A segunda etapa do método proposto consistiu no processo de otimização das estruturas oriundas da etapa de amostragem e inicialização de modelos estruturais. Nesta etapa, foi desenvolvido um AM multi populacional para o problema PSP, idealizado como

seguimento do algoritmo proposto no trabalho de Corrêa et al. (2016), o qual é fundamentado na organização da população de indivíduos em uma estrutura em árvore, onde cada nodo é visto como uma subpopulação independente, que ao longo do processo interage com outros nodos, por meio de operações de cruzamento adaptadas ao problema, buscando o compartilhamento de informações, a diversificação da população, e a exploração mais eficaz do espaço de busca conformacional. Foi desenvolvida ainda uma versão modificada do algoritmo ABC, onde cada nodo da árvore incorpora uma execução independente como forma de refinamento local.

Objetivando analisar a etapa II de otimização de estruturas, dividiu-se esta parte de experimentações em duas. A primeira parte reuniu experimentos realizados com o intuito de demonstrar as adaptações ocorridas no AM ao longo do trabalho para que este culminasse na abordagem proposta, assim como definir o valor mais adequado ao parâmetro p relacionado à diversidade de soluções, empregado na etapa de reinicialização das subpopulações da meta-heurística. Para tanto, foram propostas 6 variações do método proposto. Os componentes analisados nas diferentes versões da meta-heurística foram: (*i*) organização populacional do AM em nichos, e incorporação da etapa I de amostragem e inicialização de indivíduos; (*ii*) procedimento de reinicialização das subpopulações; e (*iii*) modificações relativas ao funcionamento do algoritmo ABC.

A partir dos resultados obtidos e das análises realizadas relativas às diferentes versões de métodos apresentados na primeira parte de experimentos, verificou-se que a versão de método MABC-1 com limiar de reinicialização mais elevado ($p = 15$), tendeu a obter os melhores resultados em comparação com todos os outros métodos. Por esta razão, definiu-se o algoritmo MABC-1 como sendo a abordagem final a ser empregada nos experimentos da segunda parte de experimentos.

Dessa forma, a segunda parte de testes objetivou testar o AM desenvolvido através da otimização de um conjunto maior de proteínas-alvo, bem como compará-lo a dois métodos de referência na área de predição de estruturas 3-D de proteínas, Rosetta e QUARK. Fundamentado pelos resultados obtidos e pelas comparações realizadas entre o MABC-1, e os métodos Rosetta e QUARK, em relação às estruturas determinadas experimentalmente, avaliados por meio das métricas estruturais de RMSD e GDT_TS, constatou-se que o algoritmo proposto foi capaz de atingir valores comparáveis ao Rosetta e ao QUARK, sendo que em alguns casos, os superou. Acrescenta-se a isso, o êxito obtido quanto à predição de estruturas 3-D de proteínas com conformações similares às estruturas determinadas experimentalmente.

Conclui-se, então, que o AM multi populacional desenvolvido nesta dissertação, pode ser considerado uma contribuição efetiva, porém inicial, para a área de predição de estruturas 3-D de proteínas. Entende-se que o método ainda necessita de melhoramentos, testes e validações mais robustos e próximos aos métodos estado da arte na área, para que os resultados obtidos possam ser melhorados. Por exemplo, a otimização de regiões irregulares em proteínas maiores, utilização de variados conjuntos de proteínas-alvo fornecidos pelo CASP e adequação às métricas de avaliação definidas também pelo CASP. Observa-se que a conclusão deste trabalho aponta diretrizes interessantes de tópicos de pesquisa relacionados a meta-heurísticas de otimização multimodal, com vasta abrangência em Bioinformática, como por exemplo, a etapa de amostragem de indivíduos, filtragem e restrição de soluções, e a divisão da população em nichos relacionados a diferentes grupos estruturais, visando a diversidade de indivíduos e a melhor exploração do modelo de busca.

Como trabalhos futuros relacionados à área de predição de estruturas 3-D de proteínas, a partir do estudo e desenvolvimento de meta-heurísticas evolutivas, é possível delinear algumas perspectivas como seguimento à esta pesquisa:

1. Realizar novos estudos em relação a funções de energia para o problema PSP, visando analisar diferentes tipos de campos de energia potencial existentes, quando da relação entre RMSD e energia estrutural. A motivação para isto surge a partir da conclusão de que nenhuma função de energia analisada neste trabalho foi capaz de refletir de forma satisfatória esta relação;
2. Adaptação da meta-heurística proposta aos conceitos de múltiplos objetivos aplicados ao problema PSP, visando contornar as ineficiências das funções de energia apresentadas;
3. Testar a abordagem de otimização proposta quando da otimização de um conjunto de proteínas de tamanho maior, visto que a maior proteína testada compreendeu 91 resíduos de aminoácidos. Proteínas maiores implicam em complexidades maiores. Esta análise torna-se interessante, pois pode apontar novas diretrizes a serem investigadas e melhoradas no método de otimização;
4. Investigar novas métricas estruturais e conhecimentos experimentais a serem incorporados no método de busca, com o intuito de contribuir com o melhoramento dos resultados obtidos, e ainda auxiliar a contornar as complexidades impostas pelo problema PSP;

5. Expandir a estrutura em árvore por meio do aumento de nodos, aumento do número de indivíduos em cada subpopulação, modificação de ordem, ou ainda através da reorganização das conexões e relações estabelecidas entre as diferentes subpopulações, visando a análise de comportamento destas novas estruturas aplicadas ao problema;
6. Alterações ou substituição do método de busca ABC empregado na otimização dos indivíduos de cada subpopulação, como forma de avaliar os resultados da abordagem como um todo. Nota-se aqui que esta possível modificação reforça uma das características mais relevantes dos AMs, que consiste na flexibilização de componentes e estratégias de busca utilizados.

8 PUBLICAÇÕES E PRODUÇÃO TÉCNICA

Neste capítulo serão apresentados os trabalhos desenvolvidos durante o Mestrado, abrangendo as áreas de inteligência artificial, otimização e meta-heurísticas, e predição de estruturas 3-D de proteínas.

8.1 Artigos completos publicados em periódicos

- BESKOW, SAMUEL; ROGÉRIO DE MELLO, CARLOS; VARGAS, MARCELLE M.; **CORRÊA, LEONARDO DE L.**; CALDEIRA, TAMARA L.; DURÃES, MATHEUS F.; DE AGUIAR, MARILTON S.. Artificial intelligence techniques coupled with seasonality measures for hydrological regionalization of Q90 under Brazilian conditions. Journal of Hydrology (Amsterdam), v. 541, p. 1406-1419, 2016.
- **CORRÊA, L. L.**; BORGUESAN, B.; FARFÁN, C.; INOSTROZA-PONTA, M.; DORN, M.. A Memetic Algorithm for 3-D Protein Structure Prediction Problem. IEEE/ACM Transactions on Computational Biology and Bioinformatics, v. PP, p. 1-1, 2016.

8.2 Capítulos de livros publicados

- **CORRÊA, L. L.**; DORN, M.. Multi-Agent Systems in Three-Dimensional Protein Structure Prediction. In: Diana Adamatti. (Org.). Multi-Agent Based Simulations Applied to Biological and Environmental Systems. 1ed. Hershey: IGI Global, 2016, v. 1, p. 241-278.

8.3 Trabalhos completos publicados em anais de eventos

- BORGUESAN, B.; BOHRER, J. S.; BARBACHAN e SILVA, M.; **CORRÊA, L. L.**; DORN, M.. Improving protein tertiary structure prediction with conformational propensities of amino acid residues. In: IEEE CONGRESS ON EVOLUTIONARY COMPUTATION. **Proceedings...** Vancouver, Canada: IEEE, 2016. p. 9-15.

- VARGAS, M. M.; BESKOW, S.; **CORRÊA, L. L.**; DURAES, M. F.; CALDEIRA, T.; MELLO, C. R.; AGUIAR, M. S.. Técnicas de Inteligência Artificial para regionalização hidrológica: uma análise da Q90 no Rio Grande Do Sul. In: XXI SIMPÓSIO BRASILEIRO DE RECURSOS HÍDRICOS. **Proceedings...** Brasília, Brasil: 2015. p. 1-8.

8.4 Resumos publicados em anais de eventos

- **CORRÊA, L. L.**; DORN, M.. A multi-agent approach for the 3-D protein structure prediction problem. In: I ESCOLA GAÚCHA DE BIOINFORMÁTICA. **Proceedings...** Porto Alegre, Brasil: 2015.
- VARGAS, M. M.; **CORRÊA, L. L.**; DURAES, M. F.; AGUIAR, M. S.; CALDEIRA, T.; BESKOW, S.. Regionalização de vazões de estiagem no estado do Rio Grande do Sul empregando técnicas de Inteligência Artificial. In: XXIV CONGRESSO DE INICIAÇÃO CIENTÍFICA DA UNIVERSIDADE FEDERAL DE PELOTAS. **Proceedings...** Pelotas, Brasil: 2015.

8.5 Artigos completos submetidos/em revisão

- **CORRÊA, L. L.**; BORGUESAN, B.; KRAUSE, M. J.; DORN, M.. Three-Dimensional Protein Structure Prediction based on the MA-SW-Chains Algorithm. Computers & Operations Research, submetido. Status: *Em revisão*.
- **CORRÊA, L. L.**; INOSTROZA-PONTA, M.; DORN, M.. An evolutionary multi-agent algorithm to explore the high degree of selectivity in three-dimensional protein structures. In: IEEE Congress on Evolutionary Computation, 2017, submetido. Status: *Aceito para publicação*.

REFERÊNCIAS

AKAY, B.; KARABOGA, D. A modified artificial bee colony algorithm for real-parameter optimization. **Inf. Sci.**, Elsevier, v. 192, p. 120–142, 2012.

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

ANFINSEN, C. B. et al. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 47, n. 9, p. 1309–1314, 1961.

BACK, T. **Evolutionary Algorithms in Theory and Practice: Evolution Strategies, Evolutionary Programming, Genetic Algorithms**. 1. ed. Oxford, UK: Oxford University Press, 1996. 328 p.

BADER, D. A. et al. Bioperf: A benchmark suite to evaluate high-performance computer architecture on bioinformatics applications. In: IEEE INTERNATIONAL WORKLOAD CHARACTERIZATION SYMPOSIUM. **Proceedings...** [S.l.]: IEEE, 2005. p. 163–173.

BAXEVANIS, A. D.; OUELLETTE, B. F. **Bioinformatics: a practical guide to the analysis of genes and proteins**. 2. ed. New York, USA: John Wiley & Sons, Inc., 2004.

BEDEIAN, A. G.; MOSSHOLDER, K. W. On the use of the coefficient of variation as a measure of diversity. **Organ. Res. Methods**, Sage Publications, v. 3, n. 3, p. 285–297, 2000.

BELDA, I. et al. Evolutionary computation and multimodal search: A good combination to tackle molecular diversity in the field of peptide design. **Mol. Diversity**, Springer, v. 11, n. 1, p. 7–21, 2007.

BERMAN, H. M. et al. The protein data bank. **Nucleic Acids Res.**, Oxford Univ Press, v. 28, n. 1, p. 235–242, 2000.

BONABEAU, E.; DORIGO, M.; THERAULAZ, G. **Swarm intelligence: from natural to artificial systems**. 1. ed. New York, USA: Oxford university press, 1999. 320 p.

BORGUESAN, B.; INOSTROZA-PONTA, M.; DORN, M. Nias-server: Neighbors influence of amino acids and secondary structures in proteins. **J. Comput. Biol.**, Mary Ann Liebert, Inc, PP, p. 1–9, 2016.

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. **Inf. Sci.**, 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.

BRADLEY, P.; MISURA, K. M.; BAKER, D. Toward high-resolution de novo structure prediction for small proteins. **Science**, American Association for the Advancement of Science, v. 309, n. 5742, p. 1868–1871, 2005.

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

BRASIL, C. R. S.; DELBEM, A. C. B.; SILVA, F. L. B. da. Multiobjective evolutionary algorithm with many tables for purely ab initio protein structure prediction. **J. Comput. Chem.**, Wiley Online Library, v. 34, n. 20, p. 1719–1734, 2013.

CARUGO, O. How root-mean-square distance (r.m.s.d.) values depend on the resolution of protein structures that are compared. **J. Appl. Crystallogr.**, v. 36, n. 1, p. 125–128, 2003.

CAVANAGH, J. et al. **Protein NMR spectroscopy: principles and practice**. 2. ed. New York, USA: Academic Press, 2006. 912 p.

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

CHIVIAN, D. et al. Ab initio methods. In: **Structural Bioinformatics**. New Jersey, USA: John Wiley & Sons, Inc, 2003. v. 44, chp. 27, p. 547–557.

CHOU, K.-C. Structural bioinformatics and its impact to biomedical science. **Curr. Med. Chem.**, Bentham Science Publishers, v. 11, n. 16, p. 2105–2134, 2004.

CHOU, K.-C.; ZHANG, C.-T. Prediction of protein structural classes. **Crit. Rev. Biochem. Mol. Biol.**, Taylor & Francis, v. 30, n. 4, p. 275–349, 1995.

COMBS, S. et al. Small-molecule ligand docking into comparative models with rosetta. **Nat. Protoc.**, Nature Publishing Group, v. 8, n. 7, p. 1277–1298, 2013.

CONNOLLY, M. L. Solvent-accessible surfaces of proteins and nucleic acids. **Science**, American Association for the Advancement of Science, v. 221, n. 4612, p. 709–713, 1983.

CONSORTIUM, . G. P. et al. A global reference for human genetic variation. **Nature**, Nature Publishing Group, v. 526, n. 7571, p. 68–74, 2015.

CONTE, L. L. et al. Scop: a structural classification of proteins database. **Nucleic Acids Res.**, Oxford Univ Press, v. 28, n. 1, p. 257–259, 2000.

COOK, S. A. An overview of computational complexity. **Commun. ACM**, ACM, v. 26, n. 6, p. 400–408, 1983.

CORRÊA, L. et al. A memetic algorithm for 3-D protein structure prediction problem. **IEEE/ACM Trans. Comput. Biol. Bioinf.**, IEEE, PP, n. 99, p. 1–1, 2016.

CORRÊA, L. de L.; DORN, M. Multi-agent systems in three-dimensional protein structure prediction. In: **Multi-Agent-Based Simulations Applied to Biological and Environmental Systems**. [S.l.]: IGI Global, 2017. p. 241–278.

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

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

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

DAS, A.; CHAKRABARTI, B. **Quantum annealing and related optimization methods**. 1. ed. usa: Springer Science & Business Media, 2005.

DAS, S. et al. Real-parameter evolutionary multimodal optimization-a survey of the state-of-the-art. **Swarm Evol. Comput.**, Elsevier, v. 1, n. 2, p. 71–88, 2011.

DAWKINS, R. **The selfish gene**. 1. ed. Oxford, UK: Oxford university press, 1976. 224 p.

DEB, K. et al. A fast and elitist multiobjective genetic algorithm: NSGA-II. **IEEE Trans. Evolut. Comput.**, IEEE, v. 6, n. 2, p. 182–197, 2002.

DESJARLAIS, J. R.; CLARKE, N. D. Computer search algorithms in protein modification and design. **Curr. Opin. Struct. Biol.**, Elsevier, v. 8, n. 4, p. 471–475, 1998.

DILL, K. A.; MACCALLUM, J. L. The protein-folding problem, 50 years on. **Science**, American Association for the Advancement of Science, v. 338, n. 6110, p. 1042–1046, 2012.

DORN, M.; BURIOL, L. S.; LAMB, L. C. A hybrid genetic algorithm for the 3-d protein structure prediction problem using a path-relinking strategy. In: IEEE. **Evolutionary Computation (CEC), 2011 IEEE Congress on**. [S.l.], 2011. p. 2709–2716.

DORN, M. et al. A knowledge-based genetic algorithm to predict three-dimensional structures of polypeptides. In: IEEE. **Evolutionary Computation (CEC), 2013 IEEE Congress on**. [S.l.], 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.

DRÉO, J. et al. **Metaheuristics for hard optimization: methods and case studies**. 1. ed. USA: Springer Science & Business Media, 2006.

DUNBRACK, R. L.; COHEN, F. E. Bayesian statistical analysis of protein side-chain rotamer preferences. **Protein Sci.**, Wiley Online Library, v. 6, n. 8, p. 1661–1681, 1997.

DUNKER, A. et al. Intrinsically disordered protein. **J. Mol. Graphics Modell.**, v. 19, n. 1, p. 26–59, 2001.

DUNKER, A. K. et al. The unfoldomics decade: an update on intrinsically disordered proteins. **BMC genomics**, BioMed Central, v. 9, n. 2, p. S1, 2008.

DUNKER, A. K. et al. Function and structure of inherently disordered proteins. **Curr. Opin. Struct. Biol.**, Elsevier, v. 18, n. 6, p. 756–764, 2008.

DUPUIS, F.; SADOC, J.-F.; MORNON, J.-P. Protein secondary structure assignment through voronoi tessellation. **Proteins: Struct. Funct. Bioinf.**, Wiley Online Library, v. 55, n. 3, p. 519–528, 2004.

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.

FARAGGI, E.; KLOCZKOWSKI, A. A global machine learning based scoring function for protein structure prediction. **Proteins: Struct. Funct. Bioinf.**, Wiley Online Library, v. 82, n. 5, p. 752–759, 2014.

FLOUDAS, C. et al. Advances in protein structure prediction and de novo protein design: A review. **Chem. Eng. Sci.**, Elsevier, v. 61, n. 3, p. 966–988, 2006.

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.

FOX, N. K.; BRENNER, S. E.; CHANDONIA, J.-M. The value of protein structure classification information—surveying the scientific literature. **Proteins: Struct. Funct. Bioinf.**, Wiley Online Library, v. 83, n. 11, p. 2025–2038, 2015.

GAO, W.; LIU, S.; HUANG, L. A global best artificial bee colony algorithm for global optimization. **J. Comput. Appl. Math.**, Elsevier, v. 236, n. 11, p. 2741–2753, 2012.

GARZA-FABRE, M. et al. Generating, maintaining and exploiting diversity in a memetic algorithm for protein structure prediction. **Evol. Comput.**, MIT Press, v. 24, n. 4, p. 577–607, 2016.

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

GLIBOVETS, N.; GULAYEVA, N. A review of niching genetic algorithms for multimodal function optimization. **Cybern. Syst. Anal.**, Springer US, v. 49, n. 6, p. 815–820, 2013.

GLOVER, F. Genetic algorithms and scatter search: unsuspected potentials. **Stat. Comput.**, Springer, v. 4, n. 2, p. 131–140, 1994.

GRAND, S. M. L.; JR, K. M. M. The application of the genetic algorithm to the minimization of potential energy functions. **J. Global Optim.**, Springer, v. 3, n. 1, p. 49–66, 1993.

GUNASEKARAN, K. et al. Extended disordered proteins: targeting function with less scaffold. **Trends Biochem. Sci.**, Elsevier, v. 28, n. 2, p. 81–85, 2003.

GUYEUX, C. et al. Is protein folding problem really a np-complete one? first investigations. **J. Bioinf. Comput. Biol.**, World Scientific, v. 12, n. 01, p. 1350017, 2014.

HANDL, J.; LOVELL, S. C.; KNOWLES, J. Investigations into the effect of multiobjectivization in protein structure prediction. In: SPRINGER. **International Conference on Parallel Problem Solving from Nature**. [S.l.], 2008. p. 702–711.

HAO, M.-H.; SCHERAGAT, H. A. Designing potential energy functions for protein folding. **Curr. Opin. Struct. Biol.**, Elsevier, v. 9, n. 2, p. 184–188, 1999.

HEINIG, M.; FRISHMAN, D. Stride: a web server for secondary structure assignment from known atomic coordinates of proteins. **Nucleic Acids Res.**, Oxford Univ Press, v. 32, n. suppl 2, p. W500–W502, 2004.

HOLLAND, J. H. **Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence.** 1. ed. Ann Arbor, USA: Univesity of Michigan Press, 1975. 183 p.

HOVMÖLLER, S.; ZHOU, T.; OHLSON, T. Conformations of amino acids in proteins. **Acta Crystallogr. Sect. D-Biol. Crystallogr.**, International Union of Crystallography, v. 58, n. 5, p. 768–776, 2002.

HUANG, Y. J. et al. Assessment of template-based protein structure predictions in casp10. **Proteins: Struct. Funct. Bioinf.**, Wiley Online Library, v. 82, n. S2, p. 43–56, 2014.

INOSTROZA-PONTA, M.; FARFÁN, C.; DORN, M. A memetic algorithm for protein structure prediction based on conformational preferences of aminoacid residues. In: GENETIC AND EVOLUTIONARY COMPUTATION CONFERENCE (GECCO 2015). **Proceedings...** New York, USA: ACM, 2015. p. 1403–1404.

ISLAM, M. K.; CHETTY, M. Novel memetic algorithm for protein structure prediction. In: **AI 2009: Advances in Artificial Intelligence**. [S.l.]: Springer, 2009. p. 412–421.

ISLAM, M. K.; CHETTY, M. Clustered memetic algorithm with local heuristics for ab initio protein structure prediction. **IEEE Trans. Evol. Comput.**, v. 17, n. 4, p. 558–576, 2013.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data clustering: a review. **ACM Comput. Surv.**, Acm, v. 31, n. 3, p. 264–323, 1999.

JONES, G. et al. Development and validation of a genetic algorithm for flexible docking. **J. Mol. Biol.**, Elsevier, v. 267, n. 3, p. 727–748, 1997.

JORGENSEN, W. L.; TIRADO-RIVES, J. Potential energy functions for atomic-level simulations of water and organic and biomolecular systems. **Proceedings of the National Academy of Sciences of the United States of America**, National Acad Sciences, v. 102, n. 19, p. 6665–6670, 2005.

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

KARABOGA, D. **An idea based on honey bee swarm for numerical optimization**. [S.l.], 2005.

KARABOGA, D.; AKAY, B. A comparative study of artificial bee colony algorithm. **Appl. Math. Comput.**, Elsevier, v. 214, n. 1, p. 108–132, 2009.

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.

KARABOGA, D.; BASTURK, B. On the performance of artificial bee colony (abc) algorithm. **Appl. Soft Comput.**, Elsevier, v. 8, n. 1, p. 687–697, 2008.

KAUFMANN, K. W. et al. Practically useful: what the rosetta protein modeling suite can do for you. **Biochemistry-Us**, ACS Publications, v. 49, n. 14, p. 2987–2998, 2010.

KENNEDY, J. et al. **Swarm intelligence**. 1. ed. San Francisco, USA: Morgan Kaufmann, 2001. 512 p.

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 Res.**, v. 32, n. Web Server issue, p. W526–531, Jul 2004.

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

KINCH, L. N. et al. Casp 11 target classification. **Proteins: Struct. Funct. Bioinf.**, Wiley Online Library, v. 84, n. S1, p. 20–33, 2016.

KITCHEN, D. B. et al. Docking and scoring in virtual screening for drug discovery: methods and applications. **Nat. Rev. Drug Discovery**, Nature Publishing Group, v. 3, n. 11, p. 935–949, 2004.

KLEYWEGT, G. J.; BRÜNGER, A. T. Checking your imagination: applications of the free r value. **Structure**, Elsevier, v. 4, n. 8, p. 897–904, 1996.

KOKKINIDIS, M.; GLYKOS, N.; FADOULOGLOU, V. Protein flexibility and enzymatic catalysis. In: CHRISTOV, C.; KARABENCHEVA-CHRISTOVA, T. (Ed.). **Structural and Mechanistic Enzymology Bringing Together Experiments and Computing**. [S.l.]: Academic Press, 2012, (Advances in Protein Chemistry and Structural Biology, v. 87). p. 181 – 218.

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

KONAK, A.; COIT, D. W.; SMITH, A. E. Multi-objective optimization using genetic algorithms: A tutorial. **Reliab. Eng. Syst. Safe.**, Elsevier, v. 91, n. 9, p. 992–1007, 2006.

KORTEMME, T.; MOROZOV, A.; BAKER, D. An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein–protein complexes. **J. Mol. Biol.**, Elsevier, v. 326, n. 4, p. 1239–1259, 2003.

KRASNOGOR, N. et al. Multimeme algorithms for protein structure prediction. In: **Parallel Problem Solving from Nature VII**. [S.l.: s.n.], 2002. p. 769–778.

KRASNOGOR, N.; SMITH, J. A tutorial for competent memetic algorithms: model, taxonomy, and design issues. **IEEE Trans. Evol. Comput.**, IEEE, v. 9, n. 5, p. 474–488, 2005.

KRYSHTAFOVYCH, A. et al. Challenging the state of the art in protein structure prediction: Highlights of experimental target structures for the 10th critical assessment of techniques for protein structure prediction experiment casp10. **Proteins: Struct. Funct. Bioinf.**, Wiley Online Library, v. 82, n. S2, p. 26–42, 2014.

KUFAREVA, I.; ABAGYAN, R. Methods of protein structure comparison. In: **Homology Modeling: Methods and Protocols**. Totowa, USA: Humana Press, 2012. v. 857, p. 231–257.

KUHLMAN, B.; BAKER, D. Native protein sequences are close to optimal for their structures. **Proc. Natl. Acad. Sci. U.S.A.**, National Acad Sciences, v. 97, n. 19, p. 10383–10388, 2000.

LANDER, E. S. et al. Initial sequencing and analysis of the human genome. **Nature**, Nature Publishing Group, v. 409, n. 6822, p. 860–921, 2001.

LASKOWSKI, R. A.; WATSON, J. D.; THORNTON, J. M. Profunc: a server for predicting protein function from 3d structure. **Nucleic Acids Res.**, Oxford University Press, v. 33, p. 89–93, 2005.

LAZARIDIS, T.; KARPLUS, M. Effective energy functions for protein structure prediction. **Curr. Opin. Struct. Biol.**, Elsevier, v. 10, n. 2, p. 139–145, 2000.

LEAVER-FAY, A. et al. Scientific benchmarks for guiding macromolecular energy function improvement. **Methods Enzymol.**, NIH Public Access, v. 523, p. 109, 2013.

LEAVER-FAY, A. et al. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. **Methods Enzymol.**, v. 487, p. 545–574, 2011.

LEE, J.; WU, S.; ZHANG, Y. Ab initio protein structure prediction. In: **From protein structure to function with bioinformatics**. 1. ed. Dordrecht, Netherlands: Springer Netherlands, 2009. chp. 1, p. 3–25.

LESK, A. **Introduction to protein science: architecture, function, and genomics**. 2. ed. New York, USA: Oxford university press, 2010. 455 p.

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

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

LEVITT, M. et al. Protein folding: the endgame. **Annu. Rev. Biochem.**, Annual Reviews, v. 66, n. 1, p. 549–579, 1997.

LI, G.; NIU, P.; XIAO, X. Development and investigation of efficient artificial bee colony algorithm for numerical function optimization. **Appl. Soft Comput.**, Elsevier, v. 12, n. 1, p. 320–332, 2012.

LI, Z. et al. Adaptive molecular docking method based on information entropy genetic algorithm. **Appl. Soft Comput.**, Elsevier, v. 26, p. 299–302, 2015.

LOBANOV, M. Y.; BOGATYREVA, N.; GALZITSKAYA, O. Radius of gyration as an indicator of protein structure compactness. **J. Mol. Biol.**, Springer, v. 42, n. 4, p. 623–628, 2008.

LODISH, H. et al. **Molecular cell biology**. 6. ed. New York, USA: W.H. Freeman, 2007. 973 p.

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

MACKERREL, A. Empirical force fields. In: XU, Y.; XU, D.; LIANG, J. (Ed.). **Computational methods for protein structure prediction and modeling**. 1. ed. New York, USA: Springer, 2010. chp. 2, p. 45–69.

MAHFOUD, S. W. Niching methods for genetic algorithms. **Urbana**, Citeseer, v. 51, n. 95001, p. 62–94, 1995.

MARTÍ-RENOM, M. A. et al. Comparative protein structure modeling of genes and genomes. **Annu. Rev. Biophys. Biomol. Struct.**, Annual Reviews, v. 29, n. 1, p. 291–325, 2000.

MAULIK, U.; BANDYOPADHYAY, S.; MUKHOPADHYAY, A. **Multiobjective Genetic Algorithms for Clustering: Applications in Data Mining and Bioinformatics**. [S.l.]: Springer Science & Business Media, 2011.

MCREE, D. E. **Practical protein crystallography**. 2. ed. London, UK: Academic press, 1999. 477 p.

MIRNY, L.; SHAKHNOVICH, E. Protein folding theory: from lattice to all-atom models. **Annu. Rev. Biophys. Biomol. Struct.**, Annual Reviews, v. 30, n. 1, p. 361–396, 2001.

MITCHELL, M. **An introduction to genetic algorithms**. [S.l.]: MIT press, 1998.

MOSCATO, P. **On Evolution, Search, Optimization, Genetic Algorithms and Martial Arts: Towards Memetic Algorithms**. Pasadena, California, USA, 1989.

MOSCATO, P.; COTTA, C. A modern introduction to memetic algorithms. In: **Handbook of Metaheuristics**. 1. ed. Boston, MA: Springer US, 2010. v. 146, p. 141–183.

MOULT, J. et al. Critical assessment of methods of protein structure prediction (casp)—round x. **Proteins: Struct. Funct. Bioinf.**, Wiley Online Library, v. 82, n. S2, p. 1–6, 2014.

MOULT, J. et al. Critical assessment of methods of protein structure prediction: Progress and new directions in round xi. **Proteins: Struct. Funct. Bioinf.**, Wiley Online Library, v. 84, n. S1, p. 4–14, 2016.

MURZIN, A. G. et al. Scop: a structural classification of proteins database for the investigation of sequences and structures. **J. Mol. Biol.**, Elsevier, v. 247, n. 4, p. 536–540, 1995.

NERI, F.; COTTA, C. A primer on memetic algorithms. In: **Handbook of Memetic Algorithms**. Heidelberg, Germany: Springer, 2012. v. 379, p. 43–52.

NERI, F.; COTTA, C.; MOSCATO, P. **Handbook of memetic algorithms**. 1. ed. Heidelberg, Germany: Springer, 2012.

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

OLSON, B.; SHEHU, A. Multi-objective stochastic search for sampling local minima in the protein energy surface. In: ACM CONFERENCE ON BIOINFORMATICS, COMPUTATIONAL BIOLOGY, AND HEALTH INFORMATICS (ACM BCB). **Proceedings...** [S.l.]: ACM, 2013. p. 430.

O'MEARA, M. et al. A combined covalent-electrostatic model of hydrogen bonding improves structure prediction with rosetta. **J. Chem. Theory Comput.**, ACS Publications, 2015.

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. **Proc. Natl. Acad. Sci. U. S. A.**, 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. **Proc. Natl. Acad. Sci. U. S. A.**, National Acad Sciences, v. 37, n. 4, p. 205–211, 1951.

PHAM, D. et al. The bees algorithm – a novel tool for complex optimisation problems. In: INTERNATIONAL VIRTUAL CONFERENCE ON INTELLIGENT PRODUCTION MACHINES AND SYSTEMS (IPROMS 2006). **Proceedings...** [S.l.]: Elsevier, 2006. p. 454–459.

PRUITT, K. D.; TATUSOVA, T.; MAGLOTT, D. R. Ncbi reference sequence (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins. **Nucleic Acids Res.**, Oxford University Press, v. 33, n. suppl 1, p. D501–D504, 2005.

QU, B.-Y.; SUGANTHAN, P. N.; DAS, S. A distance-based locally informed particle swarm model for multimodal optimization. **IEEE Trans. Evolut. Comput.**, IEEE, v. 17, n. 3, p. 387–402, 2013.

RICHARDSON, J. S. The anatomy and taxonomy of protein structure. **Adv. Protein Chem.**, Elsevier, v. 34, p. 167–339, 1981.

RICHMOND, T. J. Solvent accessible surface area and excluded volume in proteins: Analytical equations for overlapping spheres and implications for the hydrophobic effect. **J. Mol. Biol.**, Elsevier, v. 178, n. 1, p. 63–89, 1984.

ROCHA, G. K. et al. Using crowding-distance in a multiobjective genetic algorithm for protein structure prediction. In: GENETIC AND EVOLUTIONARY COMPUTATION CONFERENCE (GECCO 2016). **Proceedings...** New York, USA: ACM, 2016. p. 1285–1292.

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

ROSE, G. D. et al. Hydrophobicity of amino acid residues in globular proteins. **Science**, American Association for the Advancement of Science, v. 229, p. 834–839, 1985.

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 Ltd, v. 13, n. Suppl 1, p. S4, 2013.

SANCHO, D. de; REY, A. Energy minimizations with a combination of two knowledge-based potentials for protein folding. **J. Comput. Chem.**, Wiley Online Library, v. 29, n. 10, p. 1684–1692, 2008.

SCHEEF, E. D.; FINK, J. L. Fundamentals of protein structure. In: **Structural Bioinformatics**. 2. ed. New Jersey, USA: John Wiley & Sons, Inc., 2009. chp. 2, p. 15–40.

SEVIER, C. S.; KAISER, C. A. Formation and transfer of disulphide bonds in living cells. **Nat Rev Mol Cell Bio**, Nature Publishing Group, v. 3, n. 11, p. 836–847, 2002.

SHEHU, A.; KAVRAKI, L. E.; CLEMENTI, C. Multiscale characterization of protein conformational ensembles. **Proteins: Struct. Funct. Bioinf.**, Wiley Online Library, v. 76, n. 4, p. 837–851, 2009.

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

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

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.

STILLINGER, F. H.; HEAD-GORDON, T.; HIRSHFELD, C. L. Toy model for protein folding. **Phys Rev E Stat Nonlin Soft Matter Phys**, APS, v. 48, n. 2, p. 1469–1477, 1993.

SUN, S. A genetic algorithm that seeks native states of peptides and proteins. **Biophys. J.**, Elsevier, v. 69, n. 2, p. 340–355, 1995.

SYSWERDA, G. Uniform Crossover in Genetic Algorithms. In: INTERNATIONAL CONFERENCE ON GENETIC ALGORITHMS. **Proceedings...** San Mateo, California: Morgan Kaufmann Publishers, Inc., 1989. p. 2–9.

TAI, C.-H. et al. Assessment of template-free modeling in casp10 and roll. **Proteins: Struct. Funct. Bioinf.**, Wiley Online Library, v. 82, n. S2, p. 57–83, 2014.

TALBI, E.-G. Common concepts for metaheuristics. In: **Metaheuristics: from design to implementation**. [S.l.]: John Wiley & Sons, Inc., 2009. v. 74, chp. 1, p. 1–86.

THOMASSON, K. A.; APPLEQUIST, J. Bond-optimized ring closure for proline: Comparison of conformations and semiempirical energies with small molecule x-ray structures. **Biopolymers**, Wiley Online Library, v. 30, n. 3-4, p. 437–450, 1990.

THOMSEN, R. Multimodal optimization using crowding-based differential evolution. In: **2004 IEEE Congress on Evolutionary Computation (CEC)**. [S.l.: s.n.], 2004. v. 2, p. 1382–1389.

TOMPA, P. Intrinsically unstructured proteins. **Trends Biochem. Sci.**, Elsevier, v. 27, n. 10, p. 527–533, 2002.

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

UNGER, R.; MOULT, J. Finding the lowest free energy conformation of a protein is an np-hard problem: proof and implications. **Bull. Math. Biol.**, Springer, v. 55, n. 6, p. 1183–1198, 1993.

UNWIN, P. N. T.; HENDERSON, R. Molecular structure determination by electron microscopy of unstained crystalline specimens. **J. Mol. Biol.**, Elsevier, v. 94, n. 3, p. 425IN13433–432IN18440, 1975.

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.

WHISSTOCK, J. C.; LESK, A. M. Prediction of protein function from protein sequence and structure. **Q. Rev. Biophys.**, Cambridge University Press, v. 36, n. 03, p. 307–340, 2003.

WONG, K.-C.; LEUNG, K.-S.; WONG, M.-H. Protein structure prediction on a lattice model via multimodal optimization techniques. In: GENETIC AND EVOLUTIONARY COMPUTATION CONFERENCE (GECCO 2010). **Proceedings...** New York, USA: ACM, 2010. p. 155–162.

XU, D.; ZHANG, Y. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. **Proteins: Struct. Funct. Bioinf.**, Wiley Online Library, v. 80, n. 7, p. 1715–1735, 2012.

XU, R.; WUNSCH, D. Survey of clustering algorithms. **IEEE Trans. Neural Netw.**, IEEE, v. 16, n. 3, p. 645–678, 2005.

YANG, X.-S. **Nature-Inspired Metaheuristic Algorithms: Second Edition**. 2. ed. Cambridge, UK: Luniver Press, 2010. 160 p.

YURIEV, E.; HOLIEN, J.; RAMSLAND, P. A. Improvements, trends, and new ideas in molecular docking: 2012–2013 in review. **J. Mol. Recognit.**, Wiley Online Library, v. 28, n. 10, p. 581–604, 2015.

ZHANG, X. et al. 3d protein structure prediction with genetic tabu search algorithm. **BMC Syst. Biol.**, v. 4, n. Suppl 1, p. S6, 2010.

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

ZHOU, A. et al. Multiobjective evolutionary algorithms: A survey of the state of the art. **Swarm Evol. Comput.**, Elsevier, v. 1, n. 1, p. 32–49, 2011.

ZHU, G.; KWONG, S. Gbest-guided artificial bee colony algorithm for numerical function optimization. **Appl. Math. Comput.**, Elsevier, v. 217, n. 7, p. 3166–3173, 2010.