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

CUT-REMD: UMA NOVA ABORDAGEM PARA PREDIÇÃO DE ESTRUTURAS TERCIÁRIAS DE PROTEÍNAS BASEADA EM RAIO DE CORTE INCREMENTAL

THIAGO LIPINSKI PAES

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

Orientador: Prof. Dr. Osmar Norberto de Souza

Porto Alegre
2017

Ficha Catalográfica

P126c Paes, Thiago Lipinski

CuT-REMD : uma nova abordagem para predição de estruturas terciárias de proteínas baseada em raio de corte incremental / Thiago Lipinski Paes . – 2017.

199 f.

Tese (Doutorado) – Programa de Pós-Graduação em Ciência da Computação, PUCRS.

Orientador: Prof. Dr. Osmar Norberto de Souza.

1. Replica Exchange Molecular Dynamics. 2. Raio de Corte Incremental. 3. Predição de Estruturas de Proteínas. 4. Amostragem. I. Norberto de Souza, Osmar. II. Título.

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

Thiago Lipinski Paes

CuT-REMD: uma nova abordagem para predição de estruturas terciárias de proteínas baseada em raio de corte incremental

Tese apresentada como requisito parcial para obtenção do grau de Doutor em Ciência da Computação do Programa de Pós-Graduação em Ciencia da Computação, Faculdade de Informática da Pontifícia Universidade Católica do Rio Grande do Sul.

Aprovado em 27 de março de 2017.

BANCA EXAMINADORA:

Prof. Dr. Adriano Velasque Werhli (FURG)

Prof. Dr. Laurent Emmanuel Dardenne (LNCC)

Prof. Dr. Rafael Andrade Caceres (UFCSPA)

Prof. Dr. Osmar Norberto de Souza (PPGCC/PUCRS - Orientador)

DEDICATÓRIA

Dedico este trabalho aos meus pais, Adão Paes e Heloisa Lipinski Paes.

“The art of simplicity is a puzzle of complexity.”
(Douglas Horton)

AGRADECIMENTOS

De coração, agradeço a todos que contribuíram para este trabalho, direta ou indiretamente.

Aos amigos, pela força e aconselhamento nos momentos turbulentos. Não os nomearei aqui com receio de esquecer algum. Meus verdadeiros amigos se sentirão citados na frase anterior.

Aos colegas de laboratório Carlos Sequeiros, Eduardo Reder, Fernando Bachega, Gustavo Migott, Luís Fernando Saraiva, Vanessa Paixão-Côrtes, Walter Paixão-Côrtes, Michele Tanus e Rafael Cauduro, pelas sadias discussões, incontáveis risadas, e pela disposição infinita de vocês em ajudar.

Aos membros do LAD Rafael Bellé e Bruno Mendes, pela ajuda dada sempre que requisitada.

Aos colegas que a FACIN me apresentou e que hoje considero também amigos: Aline Zanin, Bernardo José, Odorico Mendizabal e Samuel Souza. Certamente, este trabalho tem um pouco de cada um de vocês.

Aos professores e membros de bancas ao longo desta caminhada, pelas críticas e sugestões de melhoria.

Ao orientador, pelos anos de convívio, paciência, e principalmente, pelo aprendizado.

E, por fim e mais importante, ao esforço sem precedentes de minha família, que, desde novo, me proporcionou a possibilidade de colocar o estudo como prioridade em minha vida. Serei eternamente grato a vocês.

Muito obrigado!

CUT-REMD: UMA NOVA ABORDAGEM PARA PREDIÇÃO DE ESTRUTURAS TERCIÁRIAS DE PROTEÍNAS BASEADA EM RAIO DE CORTE INCREMENTAL

RESUMO

Dentre os principais métodos computacionais aplicados atualmente ao estudo de proteínas, a dinâmica molecular clássica realiza importante papel, especialmente sua variação intitulada *Replica Exchange Molecular Dynamics* ou REMD, a qual provê amostragem conformacional eficiente. Elementos de Estruturas Secundárias (EES) regulares de proteínas são formados e mantidos através de estabilização por ligações de hidrogênio dentro de hélices e entre fitas de uma folha β. O empacotamento desses elementos estruturais, permitido por voltas e laços flexíveis conectando-os, leva à formação de uma estrutura que, nos casos bem sucedidos, representa o estado nativo, funcional de uma proteína. Interações iônicas, dipolo-dipolo, de van der Waals e hidrofóbicas, além de ligações de hidrogênio, são fundamentais para esses eventos. A maioria dessas forças é mais forte até uma distância de 4,0 Å. Assim, essas (de 0,0 Å a 4,0 Å) são as distâncias envolvidas na formação de estruturas locais, que podem ainda se propagar e formar elementos inteiros de estrutura secundária. A prática comum ao se executar simulações por DM é, no entanto, manter um raio de corte fixo em valores maiores ou iguais a 8,0 Å. Esta tese apresenta o método CuT-REMD, uma nova abordagem de REMD com base em raio de corte incremental (variando de 4,0 Å a 8,0 Å) testando a hipótese de que tal abordagem pode otimizar a predição de estruturas terciárias de proteínas. Primeiramente, foi utilizada a proteína *villin headpiece* humana (código PDB 1UNC), como estudo de caso, e nove diferentes protocolos de simulação foram testados, todos em triplicata. Posteriormente, com base nos resultados obtidos, um protocolo-padrão foi escolhido como protocolo CuT-REMD, e um conjunto de nove proteínas adicionais foi testado, sendo os resultados comparados com o método REMD convencional. A utilização de raio de corte incremental provou-se uma abordagem eficaz para melhorar a qualidade e velocidade das predições de estruturas de proteínas via REMD. Aplicando o método ao conjunto teste de proteínas, embora de tamanho limitado, CuT-REMD mostrou bom desempenho em relação aos métodos *ab initio*, colocando-se na grande maioria das vezes ou como o melhor método de predição ou com resultados próximos aos melhores métodos. Isso possibilitou compará-lo também com métodos *de novo* e, embora com mais dificuldade, CuT-REMD manteve bom desempenho, inclusive superando certos servidores em todas as ocasiões. Os resultados obtidos, em suma, mostram-se encorajadores, com o surgimento de novos questionamentos a serem abordados futuramente.

Palavras-Chave: Replica Exchange Molecular Dynamics, Raio de Corte Incremental, Predição de Estruturas de Proteínas, Amostragem.

CUT-REMD: A NOVEL APPROACH FOR TERTIARY PROTEIN STRUCTURE PREDICTION BASED ON INCREMENTAL CUTOFF

ABSTRACT

Among the main computational techniques currently applied to study proteins, classical molecular dynamics plays a important hole, specially its variation called replica exchange molecular dynamics or REMD, which provides efficient conformational sampling. Regular secondary structures elements of proteins are formed and maintained via stabilization by hydrogen bonds within helices and between strands of a β-sheet. Packing of these structural elements, allowed by flexible turns and loops connecting them, leads to the formation of a structure that, in the successful cases, represents the native, functional state of a protein. Ionic, dipole, van der Waals, hydrophobic interactions, and hydrogen bonding are fundamental to these events. Most of these forces are strong up to a distance of 4.0 Å. Hence, these are the distances involved in the formation of local structural nubs that can further propagate and form whole elements of secondary structure. The common practice while simulating is, however, to keep fixed the cutoff at values higher or equal to 8.0 Å. Here a novel replica exchange molecular dynamics approach based on running cutoffs (varying from 4.0 Å to 8.0 Å) to enhance protein structure prediction is presented. We first proved the method as a reproducible one, as well as following a Boltzmann distribution and sampling different structures of conventional REMD. The human villin headpiece protein (PDB ID: 1UNC) was used as case study. We tested 9 different simulation protocols, in triplicate, and proved the use of incremental cutoff as an effective approach to enhance the quality and speed of protein structure predictions via replica exchange molecular dynamics. Applying the method to the protein test set, although of limited size, CuT-REMD showed good performance against the *ab initio* methods, most of the time being either as the best prediction method or with close results to the best ones. This made it possible to also compare CuT-REMD with *de novo* methods. Despite the difficulties, CuT-REMD maintained a good performance even surpassing certain servers for all tested proteins. The results obtained are encouraging, with the emergence of new questions to be addressed in the future.

Keywords: Replica Exchange Molecular Dynamics, Running Cutoff, Protein Structure Prediction, Sampling.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

3D – Tridimensional

AG – Algoritmos Genéticos

AM – Algoritmos Meméticos

BB – Algoritmos de *Branch and Bound*

BC – *Balance Condition* ou Condição de Equilíbrio

CF – Campo de Força ou *Force Field*

CPC – Condições Periódicas de Contorno

CSA – *Conformational Space Annealing*

CG – Modelos *Coarse-Grained* ou reduzidos

CuT-REMD – Cutoff Temperature Replica Exchange Molecular Dynamics

Cu-MD – Cutoff Molecular Dynamics

DBC – *Detailed Balance Condition* ou Condição de Equilíbrio Detalhada

dDFIRE – *dipole Distance-scaled, Finite Ideal-gas Reference*

DFIRE – *Distance-scaled, Finite Ideal-gas Reference*

DOPE – *Discrete Optimized Protein Energy*

DM – Dinâmica Molecular

DR – *Disordered Regions* ou Regiões Desordenadas

DRES – *Dimensional Reduction Ensemble Similarity* ou redução dimensional de *ensembles*

EAF – *Exchange Attempt Frequency* ou frequência de tentativa de intercâmbio

EAR – *Exchange Acceptance Ratio* ou taxa de aceitação entre intercâmbios

EES – Elementos de Estruturas Secundárias

ES – Estrutura Secundária

ETR – *Exchange Trapping Ratio* ou taxa de aprisionamento entre intercâmbios

FACIN – Faculdade de Informática

FarmInf – Laboratório de FarmInformática

FM – *Free Modelling* ou Modelagem Livre

GB – Generalized Born

GDT – *Global Distance Test* ou Teste de Distância Global

K – Kelvin

LAD – Laboratório de Alto Desempenho

LABIO – Laboratório de Bioinformática, Modelagem e Simulação de Biossistemas

LINUS – Local Independent Nucleated Units of Structure

MC – Monte Carlo

MMC – Mecânica Molecular Clássica

MMCM – Método de Múltiplas Cadeias de Markov ou *Multiple Markov Chain Method*

ns – nanosegundos

PDB – *Protein Data Bank*

PME – Particle-Mesh Ewald

ps – picosegundos

PSP – *Protein Structure Prediction*

QA – *Quality Assessment* ou Avaliação de Qualidade

QCS – *Quality Control Score*

REMC – Replica Exchange Monte Carlo

REMD – Replica Exchange Molecular Dynamics

RMN – Ressonância Magnética Nuclear

RMSD – *Root-Mean-Square Deviation* ou desvio quadrático médio

RR – Resíduo-Resíduo

SCOP – *Structural Classification Of Proteins*

SB – *Swarm-based optimization algorithms*

ST – *Stocastic Tunneling* ou Tunelamento Estocástico

TBM – *Template-Based Modelling* ou Modelagem Baseada em Moldes

TE – *Tunneling Event* ou evento de tunelamento

TM – Taxa de Melhoria

TP – Têmpera Paralela ou *Parallel Tempering*

TS – *Tertiary Structure predictions* ou Predições de Estrutura Terciárias

LISTA DE SÍMBOLOS

SUMÁRIO

1. INTRODUÇÃO

Macromoléculas biológicas, como proteínas, são os componentes primários do maquinário celular. Conhecimento acerca da estrutura, dinâmica e função dessas moléculas pode melhorar significativamente o entendimento dos seres vivos. Esse entendimento leva a uma capacidade cada vez maior para lidar com fenômenos naturais em relação aos quais, a princípio, o ser humano não tem controle: doenças, envelhecimento, dor, etc. Embora muitos experimentos possam ser utilizados para determinar a função de moléculas biológicas, a análise funcional por si só não pode descrever o comportamento físico ou químico inerente a uma molécula. Assim, torna-se vantajoso estudar a estrutura e a dinâmica dessas moléculas, a fim de se obter uma melhor compreensão de sua função biológica. A estrutura tridimensional (3D) adotada por uma proteína em seu estado nativo é requisito para sua função.

Ao longo dos anos, milhões (cerca de 96 milhões em 14 de setembro de 2016) de proteínas não redundantes (que possuem apenas uma única entrada no banco de dados) tiveram sua sequência de aminoácidos descoberta (http://www.ncbi.nlm.nih.gov/genbank/). Entretanto, até 15 de setembro de 2016, apenas 122.583 tiveram sua estrutura 3D ou terciária revelada no *Protein Data Bank* (PDB) [BWF+00]. Destas, apenas 38.752 são sequências de proteínas distintas, que correspondem a 1.393 enovelamentos SCOP (Structural Classification Of Proteins) [BWF+00, CKML+16].

Técnicas experimentais para resolver estruturas de proteínas, como difração de raios X, ressonância magnética nuclear e microscopia eletrônica, são demoradas e caras, bem como limitadas a certas condições biológicas [Gü04]. Para reduzir esse grande hiato entre a capacidade de produzir novas sequências de proteínas e a limitada capacidade de resolver suas estruturas 3D, particularmente estruturas que se configurem como novos enovelamentos, tornou-se primordial o desenvolvimento e aplicação de abordagens computacionais alternativas para prever sua estrutura 3D a partir da estrutura primária ou sequência de aminoácidos. Esse problema, conhecido como o problema da predição de estrutura de proteínas ou *Protein Structure Prediction* (PSP), tem sido investigado há pouco mais de 60 anos. Sua relevância biológica, combinada com sua complexidade NP-Completa, qualifica-o como um dos grandes desafios da ciência moderna [CGP+98, DM12].

Uma série de métodos computacionais tem sido proposta, variando entre modelagem por homologia [MRSF+00], reconhecimento de padrões [ANZ95, Sö05, KWW+12], métodos *de novo* [SBRB99, ZAS05, LTR+16] e *ab initio* [CRSB05, JBS+06]. Entre eles, destacam-se os métodos *ab initio*, que utilizam apenas a estrutura primária da proteína, sem o uso de homólogas com estruturas conhecidas ou demais informações provenientes de bases de dados. Os métodos *de novo* e *ab initio* possuem como importante característica a capacidade de encontrar até mesmo novos enovelamentos [FFM+06].

Um dos métodos mais utilizados para estudar a dinâmica de proteínas é a Dinâmica Molecular (DM). No entanto, devido à sua superfície de energia altamente rugosa [FSW91] e ao fato das simulações por DM convencional funcionarem a uma temperatura constante, a amostragem fica comprometida, fazendo que haja a tendência das conformações estarem presas em mínimos locais, limitando a eficiência de amostragem do método. O método REMD (*Replica Exchange Molecular Dynamics*) foi projetado para solucionar esse problema utilizando um conjunto de réplicas independentes a diferentes temperaturas, mas permitindo o intercâmbio entre elas [Han97, SO99].

Simulações REMD têm sido cada vez mais aplicadas ao estudo da dinâmica de enovelamento e caracterização de estrutura de proteínas específicas [Sue03, SPHvdS05, LWLD07, XM08, LWW+08, BWD07, KSJ10]. Em 2.5.3, encontram-se mais informações sobre a aplicação de REMD para o problema PSP.

Diferentes técnicas de amostragem exploram diferentemente o espaço de configuração. É um bom sinal ter uma exploração mais ampla, mas para fins de predição de estrutura de proteínas, não é garantido que uma exploração mais ampla resultará na obtenção de melhores estruturas. Conforme mencionado acima, o tempo de computação também é crítico, e assim sendo, é esperado de novos métodos que não sobrecarreguem as simulações de DM ao passo em que aumentam sua capacidade de amostragem. O método exposto aqui busca não onerar as simulações.

Elementos de Estruturas Secundárias (EES) regulares de proteínas são formados e mantidos através de estabilização por pontes de hidrogênio dentro de hélices e entre fitas de uma folha β. O empacotamento desses elementos estruturais, permitido por voltas e laços flexíveis conectando-os, leva à formação de uma estrutura que, nos casos bem sucedidos, representa o estado nativo, funcional de uma proteína. Interações iônicas, dipolo-dipolo, de van der Waals e hidrofóbicas, além de ligações de hidrogênio, são fundamentais para esses eventos. A maioria dessas forças é mais forte até uma distância de 4,0 Å. Assim, de 0,0 Å até 4,0 Å são as distâncias envolvidas na formação de estruturas locais, que podem ainda se propagar e formar elementos inteiros de estrutura secundária. A prática comum ao se executar simulações por DM é, no entanto, manter um raio de corte fixo em valores maiores ou iguais a 8,0 Å.

Por essas razões, decidimos implementar um método que considere esses eventos enquanto prediz a estrutura 3D de uma proteína por métodos como simulações de dinâmica molecular. Breda e colaboradores [BSBNDS07] foram os primeiros a usar essa abordagem. Eles simularam um feixe de três hélices e em menos de 10 ns conseguiram obter a estrutura enovelada. No entanto, na configuração levógira. Esse resultado foi contrário ao esperado, embora a configuração levógira seja uma das duas (destrógira e levógira) configurações possíveis para um feixe de três hélices [SLD98].

O raio de corte utilizado para avaliar interações intermoleculares de átomos não ligados, em simulações moleculares, geralmente varia de 8,0 Å para cima. O pressuposto

é que, iniciando-se a partir de 8,0 Å, se promove o rápido colapso de toda a estrutura e, a menos que sejam utilizadas temperaturas mais altas para superar mínimos locais altamente estáveis, não se pode amostrar eficientemente o espaço conformacional em direção à estrutura nativa. Iniciando-se de um raio de corte menor, espera-se que o protocolo de simulação permita a iniciação de estruturas locais dentro de diferentes segmentos ao longo da cadeia polipeptídica. Esses conglomerados de estruturas, por sua vez, podem se agregar, desagregar, reagregar e, finalmente, se expandir para formarem EES de tamanho adequado na medida em que se aumenta o raio de corte gradualmente de 4,0 até 8,0 Å. A partir de 6,0 Å, é possível notar o início do empacotamento estável da estrutura terciária [BSBNDS07], e as temperaturas mais altas dos métodos REMD aparecem então como cruciais para se escapar de conformações indesejáveis presas em mínimos locais.

Nesta tese, apresenta-se o método *Cutoff Temperature Replica Exchange Molecular Dynamics* (CuT-REMD) para abordar o problema de PSP. CuT-REMD é baseado em simulações REMD que consideram todos os átomos (do Inglês, *all-atom simulations*) com solvente implícito e um raio de corte incremental. Para comparações de desempenho, também se aplica a abordagem que utiliza raios de corte incremental em simulações por MD simples (não REMD), de temperatura única, denominadas aqui como a abordagem *Cutoff Molecular Dynamics* (Cu-MD). Todas as simulações foram realizadas durante 50 ns cada, um pequeno tempo de simulação quando comparado com os trabalhos atuais [MJG+14, SKS+15, JW14a, PMD15]. Utilizou-se o subdomínio C-terminal da proteína *villin headpiece* de humanos (código PDB 1UNC) como estudo de caso.

Primeiramente, foi verificado se as distribuições de energia geradas como saída pelo método estavam em conformidade com a curva teórica esperada para métodos que seguem uma distribuição de Boltzmann. Sendo um método que visa a predição da estrutura 3D de proteínas, a necessidade de ser um método reprodutível é manifesta, e portanto, verificou-se como diferentes execuções de CuT-REMD flutuam umas em relação às outras ao se mover dentro da robusta superfície de energia de uma proteína. O impacto da aplicação de um raio de corte incremental na exploração do espaço de energia potencial é quantificado para se compreender até que ponto 5/10 ns de simulação com raios de corte mais curtos influenciam uma simulação REMD de 50 ns. O método foi testado utilizando seis protocolos diferentes: quatro e dois protocolos envolvendo, respectivamente, simulações CuT-REMD e Cu-MD. Os métodos convencionais de REMD e DM também foram aplicados, visando dar suporte às avaliações. Todas as simulações iniciaram de uma estrutura polipeptídica estendida, variando-se o tempo de permanência em cada raio de corte, além do tempo de simulação entre tentativas de intercâmbio ou *Exchange Attempt Frequency* (EAF). O tempo necessário para as simulações convergirem foi examinado, assim como sua capacidade exploratória. De modo geral, os resultados sustentam a proposição de que a utilização do esquema incremental de raio de corte apresentado por CuT-REMD melhora a qualidade e a rapidez da predição de estruturas tridimensionais via REMD, permitindo ex-

ploração conformacional mais ampla, maior difusão entre réplicas e resultados satisfatórios quanto à amostragem de estruturas nativas.

O protocolo de melhor desempenho no estudo de caso foi então aplicado a um conjunto teste de proteínas, heterogêneo quanto a classes de proteínas. Os resultados foram comparados em duas frentes: em relação a REMD convencional e aos métodos disponíveis na literatura. CuT-REMD mostrou melhor aptidão para predizer as estruturas contendo hélices, sejam elas da classe α ou $\alpha\beta$, sendo menos apto a predizer estruturas da classe β. Quanto à comparação com a literatura, CuT-REMD mostrou bom desempenho em relação aos métodos *ab initio*, colocando-se, na grande maioria das vezes, ou como o melhor método de predição ou com resultados próximos aos melhores métodos, dependendo da proteína estudada. Além disso, uma vez que os resultados de CuT-REMD comparados aos métodos *ab initio* foram satisfatórios, estendeu-se a comparação aos métodos *de novo*, e embora com mais dificuldade, CuT-REMD manteve bom desempenho, inclusive superando certos servidores em todas as ocasiões. Em suma, os resultados obtidos pelo estudo mostram-se encorajadores, abrindo espaço para novos desafios e novas pesquisas relacionadas.

1.1 Organização

Esta tese está organizada em nove capítulos, seguidos de três apêndices:

- O primeiro capítulo introduz o problema de pesquisa problema e a solução proposta.
- O segundo capítulo contém a fundamentação teórica necessária para o entendimento do trabalho. Nele, o conceito de proteínas é introduzido, juntamente com o problema da predição de suas estruturas 3D e os diferentes métodos utilizados para o tratamento desse problema, além de elucidações no que se refere aos métodos de simulação abordados com mais profundidade pela tese. O encontro bianual CASP também é abordado, seguido das medidas de avaliação da qualidade de modelos de proteínas. Por fim, faz-se a apresentação do *software* para simulação molecular de proteínas utilizado neste trabalho, o AMBER14.
- No terceiro capítulo, a tese é apresentada, elencando-se a motivação do trabalho, o objetivo geral e os objetivos específicos.
- A metodologia empregada na criação do método aqui apresentado além dos recursos utilizados formam o capítulo 4.
- O capítulo 5 descreve a primeira parte dos resultados e discussão, explicitando, entre outros pontos, a implementação realizada, os parâmetros das simulações e os *softwares* codificados e disponibilizados. Além disso, o capítulo traz, também, a descrição

das proteínas alvo de teste neste trabalho, juntamente com a especificação de cada tipo de análise realizada.

- No capítulo seguinte, de número 6, o qual representa a segunda parte dos resultados e discussão, o foco é o estudo de caso da proteína *villin headpiece* de humanos, de código PDB 1UNC.
- Em seguida, tem-se a terceira e última parte dos resultados e discussão, a qual está relacionada a um conjunto teste de proteínas.
- As conclusões compõem o capítulo 8, onde as principais contribuições desta tese são elencadas.
- No último capítulo, de número 9, são feitas as considerações finais, as perspectivas em relação à continuação da pesquisa são compartilhadas, além das limitações da abordagem desenvolvida.
- Por fim, seguem ainda quatro Apêndices com o objetivo de complementar o texto principal.

2. FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta os principais conceitos que envolvem esta tese de doutorado. Primeiramente, é abordado o conceito de proteínas, seguido do problema-alvo da abordagem deste estudo, o problema da predição estrutural de proteínas. Em seguida, são expostos os principais métodos computacionais de predição de proteínas, os conceitos referentes a simulações moleculares de proteínas e, por fim, as medidas de avaliação de qualidade a serem utilizadas no trabalho.

2.1 Proteínas e sua Composição

Proteínas são as macromoléculas biológicas mais abundantes, ocorrem em todas as células e em todas as partes das células. Todas as proteínas, sejam das linhagens mais antigas de bactérias ou das formas mais complexas de vida, são construídas a partir de um mesmo conjunto formado por 19 aminoácidos diferentes e um iminoácido (prolina) que se ligam em uma sequência linear [LNC08, Les08].

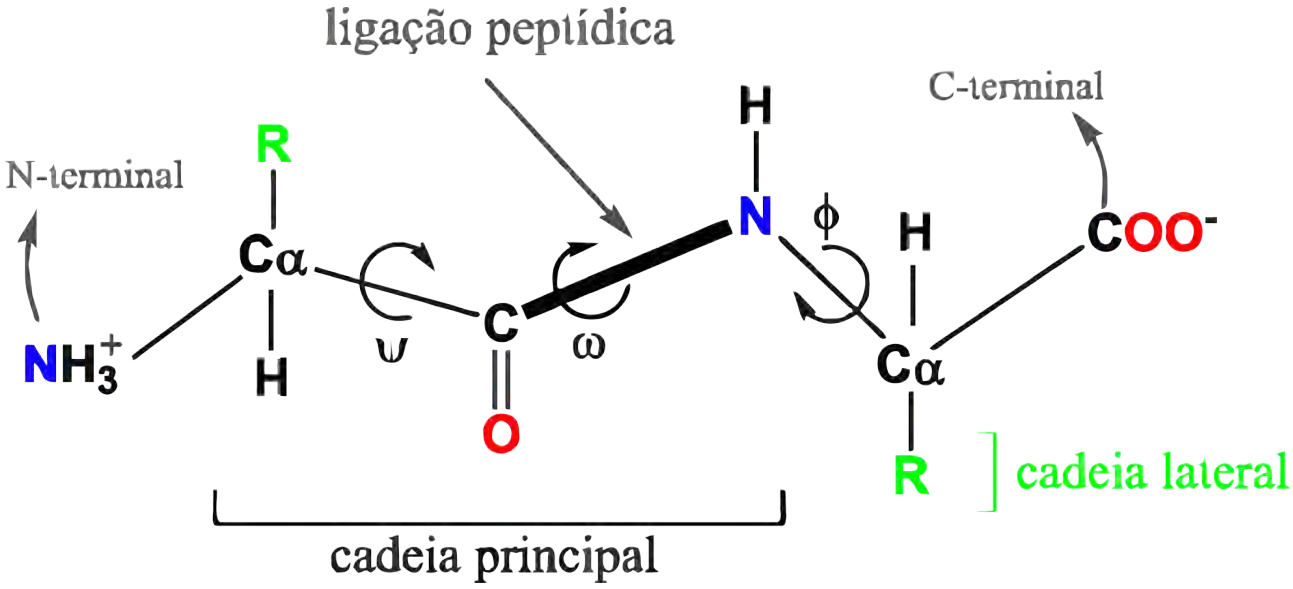

Figura 2.1 – Estrutura química de dois resíduos de aminoácidos, onde R representa as cadeias laterais. A estrutura dos aminoácidos tem uma característica comum: a presença de um grupamento carboxilato (COO-) e um grupamento amino (NH3+) ligados ao mesmo átomo de carbono (o carbono α). Os aminoácidos diferem entre si por suas cadeias laterais, também conhecidos como grupos R, que se ligam também a seus respectivos carbonos α.

Os grupos R variam em se tratando de estrutura e carga elétrica, além de tamanho, podendo contar com de 1 a 18 átomos [LNC08]. Um peptídeo é uma molécula composta por dois ou mais aminoácidos unidos por uma ligação peptídica (Figura 2.1) e possui três ângulos de torção em sua cadeia principal, chamados phi (ϕ), psi (ψ) e ômega (ω).

A ligação peptídica assume preferencialmente a configuração *trans* (ω=180 graus), na qual as cadeias laterais de aminoácidos adjacentes ficam em lados opostos da ligação peptídica.

Na configuração cis (ω=0 grau), as cadeias laterais de aminoácidos adjacentes situam-se de um mesmo lado da ligação peptídica. Essa e outras observações indicam que o esqueleto de uma proteína compõe-se de uma sequência de grupos peptídicos planares rígidos e ligados [VV06]. Assim sendo, o enovelamento da proteína ou o enovelamento do esqueleto polipeptídico depende dos ângulos de torção que essa cadeia pode assumir. A rotação somente é permitida nas ligações simples de todos os resíduos: N–Cα e Cα–C (exceto prolina).

O enovelamento de uma proteína é dado pelos ângulos diedrais ϕ (phi) e ψ (psi) dessas ligações e pelo ângulo ω (ômega) de rotação em torno da ligação peptídica [Les08]. Os ângulos ϕ, ψ e ω da cadeia principal representam de forma única a conformação de uma proteína. Das combinações entre os 20 tipos de resíduos de aminoácio/iminoácido, uma gama imensa de proteínas pode ser formada e, assim, diferentes organismos podem então fazer uso de variados produtos. Algumas proteínas realmente contêm resíduos que não os 20 acima referidos, todavia esses são produzidos por modificações químicas pós-traducional ou pela introdução de uma selenocisteína durante a tradução, como na glutationa peroxidasse [Les00]. Entre a gama de proteínas existentes, é possível citar alguns tipos, como por exemplo, enzimas, hormônios, anticorpos e fibras musculares. Proteínas são constituintes de muitas partes vitais dos seres vivos, como as proteínas da lente do olho, penas, teias de aranha, chifres de rinocerontes, proteínas do leite, antibióticos, venenos de cogumelo e uma infinidade de outras substâncias com distintas atividades biológicas [LNC08].

Sobre a estrutura das proteínas, existem quatro níveis definidos. A sequência linear dos aminoácidos que se associam por meio de ligações peptídicas formando a proteína é a sua estrutura primária. A Estrutura Secundária ou ES (Figura 2.2) é o primeiro nível de dobramento da proteína e é obtida pelo arranjo espacial de aminoácidos que formam padrões de estruturas regulares (ER) do tipo α hélice e fitas β.

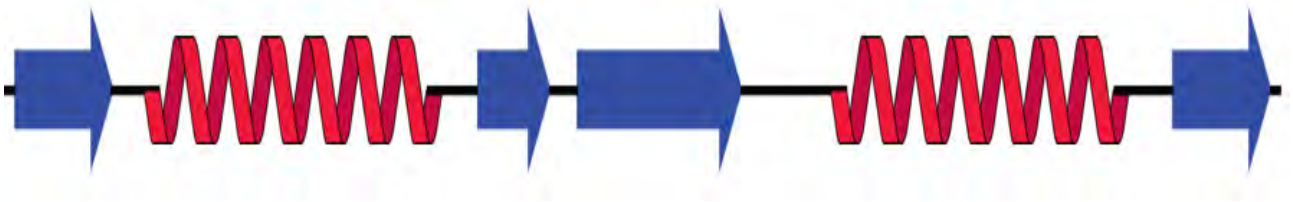

Figura 2.2 – Estrutura secundária de uma proteína. Hélices α e folhas β estão coloridas de vermelho e azul, respectivamente. Voltas e alças são as linhas retas conectando essas ES regulares. Figura obtida de [ZB07].

As regiões que conectam ES regulares são denominadas voltas e alças. Voltas são estruturas secundárias irregulares e, normalmente, possuem de dois a quatro resíduos

de aminoácidos. As alças possuem cinco ou mais resíduos de aminoácidos e são denominadas espirais desorganizadas (do inglês *random coils*).

A estrutura terciária (Figura 2.3) é formada pelo dobramento e empacotamento tridimensional das ES da proteína, chegando-se até uma conformação final única para a proteína. Quando a proteína tem mais de uma subunidade polipeptídica, a conformação espacial dessa proteína é chamada de estrutura quaternária (Figura 2.4) [ZB07].

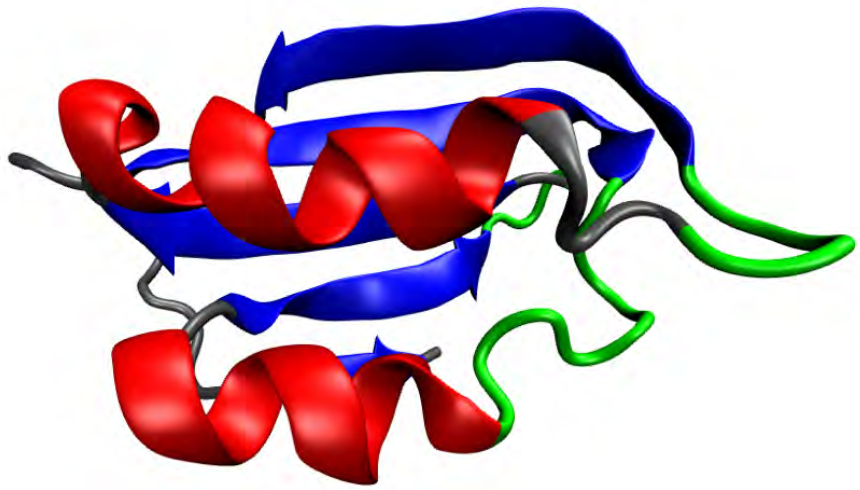

Figura 2.3 – Estrutura terciária da proteína acilfosfatase de *Eschaerichia coli* (código PDB 2GV1). Hélices α e a folha β, contendo cinco fitas, estão coloridas de vermelho e azul, respectivamente. As alças estão em cinza e as voltas em verde. Imagem criada pelo *software* VMD, representação do tipo *cartoon* [HDS96].

2.2 Problema: Predição de Estruturas 3D de Proteínas

O problema PSP é o problema da predição da estrutura 3D de uma proteína partindo-se do pressuposto de que já se conhece a sua estrutura primária ou sequência de aminoácidos. A estrutura terciária de uma proteína está diretamente ligada à sua função, pois pode permitir a identificação de domínios conhecidos, como sítios catalíticos, sítios de modificação alostérica e outros, além de contribuir para melhor entendimento relacionado a funções regulatórias, de transporte e armazenagem, controle de transcrição de genes e catálises em reações químicas [LRO07, Les08, RO09, WL03].

A determinação de estruturas proteicas de forma experimental é dispendiosa em duas frentes: (i) seja em relação ao tempo necessário ou (ii) ao custo inerente a técnicas como cristalografia, microscopia eletrônica, ressonância magnética nuclear (RMN) ou criomicroscopia eletrônica [Gü04]. Tendo em vista que a grande maioria dos fármacos atualmente no mercado atua interagindo com enzimas, o estudo da relação estrutura-função

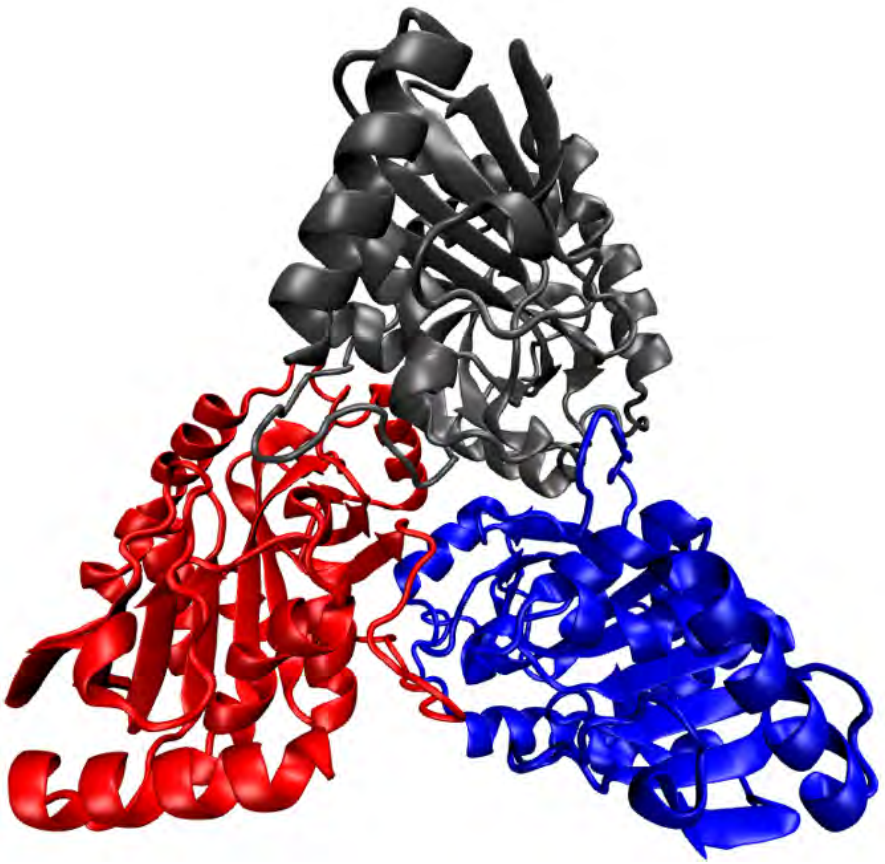

Figura 2.4 – Estrutura quaternária da proteína PNP de *Mycobacterium tuberculosis* (código PDB 1G2O). Formada pela interação de três subunidades diferentes, uma em azul, outra em cinza e outra em vermelho. Imagem criada utilizando o *software* VMD, representação do tipo *cartoon* [HDS96].

mostra-se vital para a criação de novas drogas, e a bioinformática possui o importante papel de acelerar o processo de evolução desse conhecimento [ZB07].

A abordagem aqui escolhida para a descoberta da estrutura 3D da proteína é dase pela busca da conformação de menor energia livre uma abordagem *ab initio* baseada na hipótese termodinâmica de Anfinsen (1973), segundo a qual a conformação nativa adotada por uma proteína é justamente aquela com a menor energia livre (Figura 2.5), o que representa o estado mais estável [Anf73]. Entretanto, a predição dessa estrutura tridimensional é nada trivial e até mesmo abordagens simplificadas têm complexidade NP - Completa [CGP+98].

2.2.1 O Paradoxo de Levinthal

A superfície de energia livre de grandes moléculas como proteínas é complexa. Existem milhares de graus de liberdade e uma grande quantidade de possíveis configurações. O número de conformações estruturais que uma proteína pode ter é enorme. Para uma cadeia com 100 aminoácidos, por exemplo, cada resíduo pode amostrar o espaço

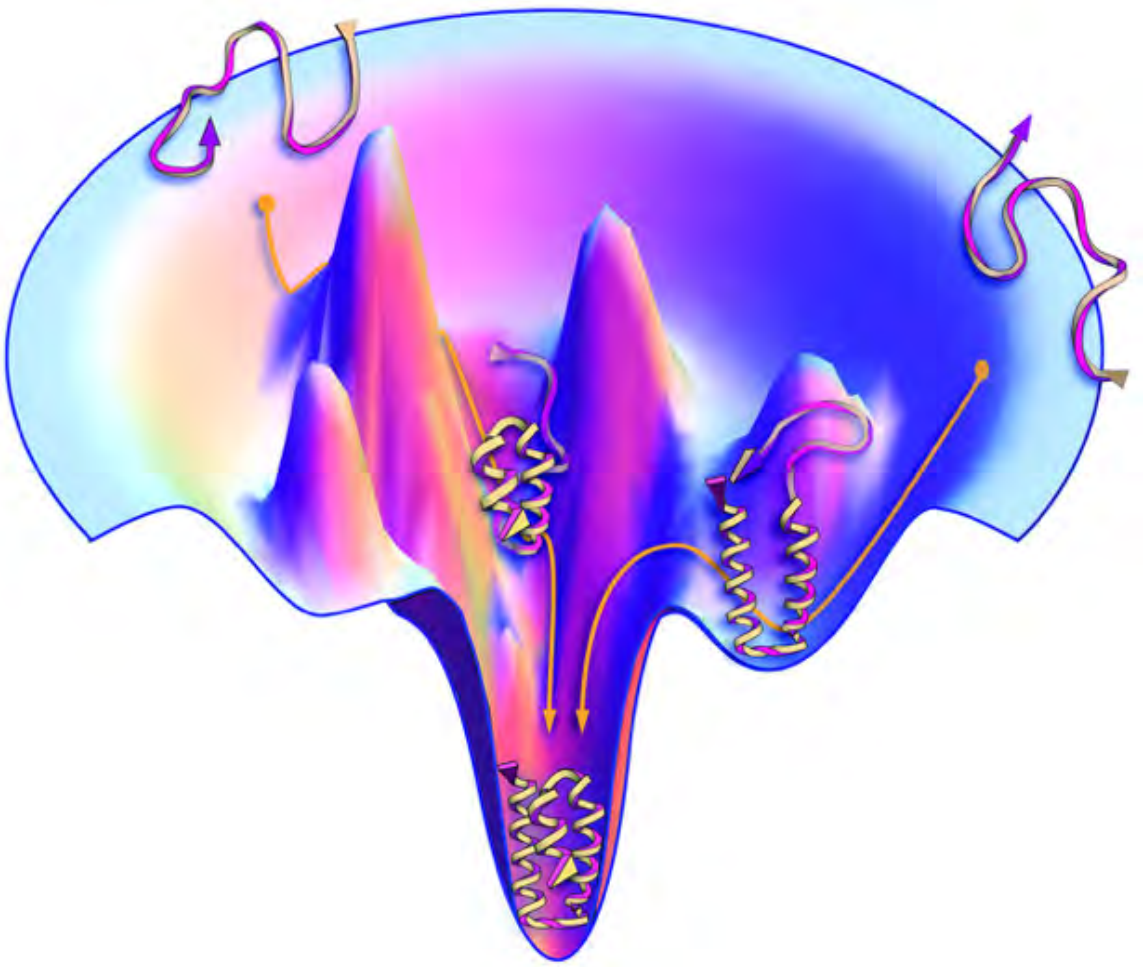

Figura 2.5 – Proteínas possuem um funil de distribuição de energia, com vários picos e vales relacionados a estruturas não enoveladas e poucos vales com energia baixa e estruturas enoveladas. Figura obtida de [DM12].

relativo a seus ângulo diedros (ϕ e ψ). Assim, para a proteína inteira, devem existir aproximadamente 2^{100} ou 10^{50} estados conformacionais disponíveis. Somente isso já caracteriza um problema intratável [Tra04].

Agora, considerando-se que a proteína precisa sequencialmente amostrar cada conformação a uma taxa similar à taxa experimental de transição (por volta de 1 pico segundo por transição), isso levaria a cerca de 10^{38} segundos para amostrar completamente seu espaço de configuração. Para se ter uma perspectiva, a idade do universo é de aproximadamente 10^{17} segundos.

Considerando as taxas de enovelamento encontradas experimentalmente, na ordem de segundos ou até mesmo mais rápidas, junto ao fato do existirem organismos vivos na Terra, acaba-se por perceber uma contradição nas premissas. Cyrus Levinthal originalmente introduziu essa contradição, em 1968, em uma tentativa de explicar que a natureza, diferentemente de procurar de forma aleatória por configurações, busca “caminhos de enovelamento” para encontrar o estado enovelado biologicamente necessário de uma proteína [Lev68].

De todo modo, o processo físico pelo qual um polipeptídeo se dobra em uma proteína funcional é uma questão antiga (revisado por C. D. Snow e colaboradores [SSRP05]) e um dos maiores desafios da bioinformática atual. Nas últimas cinco décadas, diferentes abordagens algorítmicas foram testadas e, embora progressos tenham ocorrido, o problema continua não solucionado até mesmo para proteínas de tamanho pequeno. Sabe-se, no entanto, que uma proteína iniciando de um estado de configuração não enovelado gradualmente se move na direção geral do estado enovelado devido ao gradiente local de sua superfície de energia potencial, e que, para descrever a superfície de energia de uma proteína, sejam as simulações determinísticas ou estocásticas, devem amostrar uma porção no mínimo equivalente de espaço de configuração que a natureza amostra sob a mesma escala de enovelamento.

2.3 Métodos Computacionais para Predição de Estruturas 3D de Proteínas

Os métodos computacionais para predição de estruturas de proteínas podem ser classificados em quatro grupos, segundo C. A. Floudas e colaboradores [FFM+06]:

1. modelagem comparativa por homologia [BBW+14, MRSF+00, LM14, BPBP12];

2. reconhecimento de padrões de enovelamento ou *fold recognition* via alinhavamento [BLE91, JTT92];

3. métodos *de novo* [RSMB04, SR95] e

4. métodos *ab initio* [Osg00].

2.3.1 Modelagem Comparativa por Homologia

A modelagem comparativa se baseia no princípio de que, se duas sequências de proteínas são relacionadas evolutivamente, elas possuem estruturas 3D similares [Flo07]. Para proteínas com razoável relação evolucionária, a modelagem por homologia é uma abordagem que gera modelos de alta precisão e, além disso, apresenta alto grau de confiabilidade, pois é possível estimar a qualidade da estrutura predita. Por outro lado, o método não permite a predição de novas formas de enovelamento, justamente por ser baseado em buscas por estruturas já existentes na base do PDB. Esse tipo de modelagem também não permite o estudo do processo de enovelamento de uma proteína [MRSF+00]. Entre os principais métodos desse grupo, encontram-se: SWISS-MODEL [BBW+14], MODELLER [MRSF+00], ReformAlign [LM14] e PyMOD [BPBP12].

2.3.2 Reconhecimento de Padrões

Reconhecimento de Padrões ou *Folding Recognition* [LC76] é o nome dado aos métodos motivados pela noção de que a estrutura é mais evolucionariamente preservada que a sequência. Se uma sequência de alta similaridade com estrutura conhecida não pode ser encontrada, uma nova proteína pode ainda ser estruturalmente similar a alguma proteína de estrutura já conhecida [DM12].

Nesse caso, as proteínas são ditas estruturalmente análogas. O reconhecimento de padrões visa à identificação de estruturas remotamente homólogas por meio de uma coleção de enovelamentos candidatos. Se essa identificação obtém sucesso, começa a etapa de alinhamento estrutural das sequências, assim como na modelagem por homologia. Quando não é possível identificar homologias pelo alinhamento par a par de sequências, utiliza-se a técnica de alinhavamento [JTT92]. Assim como na modelagem por homologia, nesse método só é possível predizer estruturas que possuam sequências idênticas ou semelhantes armazenadas no PDB. Dentre os principais métodos desse grupo, é possível destacar: GENTHREADER [Jon99], 123D [ANZ95], ORFEUS [GPW+03], PROSPECT (Protein structure prediction and evaluation computer toolkit) [XX00], BioShell-Threading [GKKG14], FFAS03 server [JRL+05], RaptorX server [KWW+12], Phyre server [KS09], HHpred [Sö05], LOOPP server [TGPE04], SPARKS-X [YFZZ11].

2.3.3 Métodos *de novo*

Métodos baseados em primeiros princípios, sejam eles com ou sem informações de banco de dados (nesta tese referenciados como *de novo* e *ab initio*), são abordagens que não se baseiam em estruturas 3D, e sim na termodinâmica estatística, mais especificamente na hipótese termodinâmica de Anfinsen [SBRB99]. Para saber qual a energia global livre da proteína, é utilizada uma função de energia potencial, a qual descreve a energia interna da proteína e suas interações com o meio. Esse tipo de modelagem tem como principal vantagem perante os métodos citados anteriormente o fato de que, utilizando-a, é possível predizer novas formas de enovelamento (se é que existem), devido ao fato de não ser baseado em proteínas com estruturas conhecidas [Flo07].

Nos métodos *de novo*, regras gerais relacionadas à estrutura de proteínas são extraídas de bases de dados de proteínas e utilizadas para construir estruturas 3D iniciais. Podem ser utilizadas, por exemplo, predições relativas às estruturas secundárias e predições de contato. São métodos que não comparam uma estrutura com a experimental, mas comparam fragmentos [FFM+06]. Como consequência, é possível observar que, quando novos enovelamentos emergem, estes são resultado da composição de *motifs* ou fragmentos de estruturas supersecundárias de proteínas com estrutura conhecida [Tra07].

Dentre os métodos que se enquadram nesse grupo, destacam-se: TASSER e I-TASSER [RKZ10, ZS04a], ROSETTA e ROSETTA@home [RSMB04, SBRB99], FRAG-

FOLD [Jon01], CABS-Fold [BJKK13], SIMFOLD [CFT03], PROFESY (PROFile Enumerating SYstem) [LKJK04], A3N (Artificial neural network N-gram-based method) [DNdS10a], CREF (Central-residue-fragment-based method) [DNdS10b, DNdS10b], PEP-FOLD [LTR+16], BHAGEERATH [JBS+06, NBBJ06] e QUARK [XZ12].

2.3.4 Métodos *ab initio*

São métodos baseados exclusivamente na termodinâmica estatística e na hipótese de Anfinsen [Anf73, Tra07], consideram a predição da estrutura tridimensional de uma proteína a partir apenas de sua sequência de aminoácidos ou estrutura primária. O único tipo de informação utilizada pelos métodos *ab initio* é relativo à parametrização dos campos de força (constantes usadas para descrever os chamados potenciais interatômicos ou funções matemáticas que descrevem um sistema de partículas de acordo com sua posição dos átomos). Esses campos de força são normalmente incorporados às abordagens computacionais internas de cada método, as quais vão desde algoritmos genéticos até tunelamento estocástico, entre outros. Dentre os principais campos de força desenvolvidos atualmente, citam-se: AMBER [CCB+95], CHARMM [BBO+83], GROMOS [CHB+05] e OPLS [JMTR96]. Mais detalhes são fornecidos em 2.5.1.

Uma das principais características dos métodos *ab initio* é o fato de serem capazes de predizer novos enovelamentos, uma vez que não são limitados a modelos provenientes do PDB. No entanto, é importante ter em mente que, em virtude de tal liberdade de atuação, os métodos *ab initio* precisam considerar um enorme número de conformações. Como já destacado anteriormente, devido ao grande número de graus de liberdade em uma cadeia polipeptídica não enovelada, ao se optar por obter a conformação de menor energia, se está lidando com um problema NP-Completo [CGP+98, Fra93, HI97, Lev68, NMK94].

Uma vez que o método aqui proposto se enquadra nesse grupo, é importante salientar com maior ênfase os métodos disponíveis na literatura. Em se tratando de métodos *ab initio* aplicados à predição de estruturas de proteínas, uma recente revisão [DeSBL14] deixa claros os métodos disponíveis, juntamente com informações referentes às abordagens computacionais internas a estes, responsáveis por guiar a maneira pela qual os métodos encontram a estrutura nativa dos polipeptídios. A Tabela 2.1 traz informações a esse respeito. Conforme já elencado anteriormente, os pacotes de modelagem molecular implementam e disponibilizam várias funções de energia potencial. Normalmente, na área da predição de estruturas, as funções de energia são utilizadas como funções de *escore*, fora dos pacotes de simulação (por estes não serem construídos especificamente para esse fim). Assim sendo, a maneira como essas funções serão utilizadas fica a cargo de cada método proposto, dando origem então a varias alternativas, cada uma com suas peculiaridades. Alguns métodos podem ser destacados, como é o caso de LINUS (Local Independent Nucleated Units of Structure) [SR95, SR02] e ASTROFOLD [KF03]. No âmbito do assunto

Tabela 2.1 – Tabela adaptada de [DeSBL14]. Métodos *ab initio* para predição de estruturas 3D de proteínas e seus métodos computacionais internos. Algoritmos Genéticos/Evolucionários (AG), Algorítmos Meméticos (AM), Algoritmos de Branch and Bound (BB), Conformational Space Annealing (CSA), Monte Carlo (MC), Tunelamento Estocástico ou Stocastic Tunneling (ST), Swarm-based optimization algorithms (SB), Replica Exchange Monte Carlo (REMC) e Têmpera Paralela (TP).

Métodos *ab initio*	MC	BB	CSA	AG	REMC	ST	TP	SB	AM
Abagyan [AT94]	Sim								
Astrofold [KF03]	Sim	Sim	Sim						
Bahamisch et al. [BAS09]								Sim	
Brasil et al. [BDdS13]				Sim					
Custódio et al. [CBD10]				Sim					
Dandekar e Argos [DA92]				Sim					
Derreumaux [Der99]	Sim								
Fonseca et al. [FPW10]								Sim	
Gibbs et al. [GCS01]	Sim								
Grand and Merz [LGMJ93]				Sim					
Herges et al. [HMW02]						Sim			
Hoque [HACD05]				Sim					
Linus [SR95, SR02]	Sim								
Mello et al. [MBFP12]	Sim								
Pedersen e Moult [PM97]				Sim					
Pokarowski [PKS03]					Sim				
Schug et al. [SHVW05]							Sim		
Smith [Smi05]									Sim
Sun [Sun95]				Sim					
Thachuk [TSH07]					Sim				
Unger e Moult [UM93]				Sim					

específico desta tese, percebe-se, após análise criteriosa da Tabela 2.1, que apenas um método [SHVW05] utiliza-se de *Parallel Tempering* ou Têmpera Paralela (TP). No entanto, este limita-se à TP estocástica (MC), não envolvendo DM. Verificou-se ainda, por meio da recente revisão, a escassez de trabalhos que relacionam a DM em geral com o problema PSP.

Destaca-se ainda, a presença de grupos de pesquisa brasileiros, como os dos professores Barroso [BDdS13], Dardenne [CBD10] e Pascutti [MBFP12]. Uma vez conhecidos os principais grupos de pesquisa em Dinâmica Molecular, no entanto, percebe-se que alguns artigos ficaram de fora da citada revisão. Isso posto, optou-se por implantar um protocolo de mapeamento sistemático da literatura, a fim de encontrar trabalhos relacionados à tese que não foram levados em consideração pela revisão em questão. O protocolo de mapeamento está exposto no Apêndice A e os resultados de interesse deste trabalho estão incorporados nas sub-subseções 2.5.1,2.5.2 e 2.5.3, as quais trazem, além de maior ex-

planação sobre específicos métodos relacionados, a devida atualização quanto à aplicação destes no âmbito do problema PSP.

2.4 CASP: Critical Assessment of Structure Prediction

Seja qual for o grupo em que os métodos se enquadram, a comunidade mundial de pesquisadores em predição estrutural se reúne a cada dois anos, desde 1994, para um encontro crítico da área, onde os diferentes métodos são analisados de forma cega. Esse encontro crítico é chamado CASP (*Critical Assessment of Structure Prediction*). No CASP, um grupo de experimentalistas é selecionado para fornecer novos alvos para os métodos de predição. Os experimentalistas resolvem as estruturas pelos métodos experimentais e disponibilizam apenas a sequência de aminoácidos ou estrutura primária aos participantes do encontro. Entre seus 20 anos de existência, várias mudanças ocorreram, seja nos métodos de avaliação ou nas categorias de participação.

Chegou-se então, em 2016, ao CASP12, em que três são as principais modalidades de participação para predição de estruturas terciárias ou, seguindo a nomenclatura do próprio CASP, Predições de Estrutura Terciárias ou *Tertiary structure predictions* (TS): (i) Modelagem Livre ou *Free Modelling* (FM) e (ii) Modelagem Baseada em Template ou *Template-Based Modelling* (TBM) e (iii) Refinamento. Tanto (i) quanto (ii) são divididos em subcategorias humano e servidor. Além da categoria TS, há também estudos relacionados à detecção de contatos resíduo-resíduo (RR), identificação de regiões desordenadas ou *disordered region*s (DR) e avaliação de qualidade de modelos em geral (sem o conhecimento da estrutura experimental), do Inglês: *Quality Assessment* (QA).

A categoria em que este trabalho se enquadra é FM, no entanto, em suas últimas edições, o CASP tem disponibilizado apenas sequencias de tamanho maior que 60 aminoácidos e, por esse motivo, os resultados dessa tese não serão avaliados levando em consideração proteínas alvo do CASP. Esse é, no entanto, um dos objetivos futuros deste trabalho.

2.5 Métodos de Simulação Molecular e suas Aplicações ao Problema PSP

O desenvolvimento dos computadores digitais na década de 50, com a supercomputação e sua aplicação na resolução de problemas científicos, introduziu o que alguns chamaram de “terceira metodologia” para a pesquisa científica: a simulação computacional [SK93]. Esse método, de caráter complementar e muitas vezes alternativo às formas convencionais de fazer ciência, experimental e teórica, teve um forte impacto em praticamente todos os campos da ciência (para exemplos, ver [IOP96, SK93].

O objetivo da simulação computacional em geral é resolver modelos teóricos em sua total complexidade, mediante as equações envolvidas e fazendo uso intensivo (e extensivo) dos computadores. Na área da física, a simulação computacional foi introduzida como uma ferramenta para o tratamento de sistemas de muitos corpos no início dos anos 50, com o trabalho pioneiro de N. Metropolis e colaboradores [MRR+53]. Mais tarde, resultados obtidos na mecânica estatística clássica, particularmente no estudo de líquidos, deram credibilidade à simulação computacional, estendendo seu uso rapidamente.

Hoje, graças ao rápido desenvolvimento da tecnologia de computadores, cuja velocidade cresce aproximadamente a um fator de 2 a cada 18 meses, a simulação computacional tem se firmado como uma ferramenta de cálculo essencial para ambos: experimentalistas e teóricos. Mediante um bom modelo computacional, não somente se pode reproduzir experimentos de laboratório, mas, além disso, graças à possibilidade de variação de parâmetros, é possível provar (ou desqualificar) modelos teóricos existentes até mesmo em intervalos de parâmetros inatingíveis experimentalmente, pelo menos por agora, assim resolvendo conflitos entre explicação teórica e observação. Outro papel-chave está relacionado à visualização dos resultados: não só é possível obter dados que podem ser comparados com os experimentos, mas também é possível obter-se um modelo gráfico do processo em questão.

Os dois métodos de simulação molecular de proteínas mais utilizados atualmente são a Dinâmica Molecular [AT89, Hee86, MGK77, VGB90], a qual possui caráter determinístico, e Monte Carlo, que possui caráter probabilístico [Fei85]. Ambos podem ser considerados métodos para a geração de diferentes configurações de um sistema de partículas, ou seja, pontos no espaço de fase compatível com as condições externas. O método REMD combina Dinâmica Molecular e Monte Carlo e surge como uma alternativa atrativa por conta de sua eficiência [Nym08]. Com intuito de alcançar resultados satisfatórios em relação ao estado da arte na área de atenção específica desta tese, optou-se pela utilização de um protocolo estruturado para a execução da pesquisa bibliográfica por trabalhos relacionados. O protocolo foi utilizado ainda para solidificar o conhecimento inerente ao tema de pesquisa e, ao mesmo tempo, identificar lacunas a serem abordadas pela tese. O protocolo de mapeamento sistemático, criado com base em [PPLB07], está disposto no Apêndice A, e seus resultados compõem 2.5.1, 2.5.2 e 2.5.3.

2.5.1 Dinâmica Molecular

A Dinâmica Molecular (DM) é uma das técnicas mais versáteis para o estudo de macromoléculas biológicas no que diz respeito à simulação computacional ou técnicas *in silico*. Por definição, a DM é uma abordagem computacional na qual conceitos advindos das conhecidas equações de Newton são aplicados para a resolução de representações ato-

místicas de um sistema molecular sujeito às condições periódicas apropriadas à geometria e simetria do sistema [VGB90].

Assim sendo, a metodologia da DM é fundamentada nos princípios da Mecânica Clássica e pode fornecer uma visão microscópica do comportamento dinâmico de átomos individuais que constituem um sistema como uma proteína, tornando possível obter-se informações desses átomos individuais em função do tempo [ABG06].

Uma vez que se trata de sistemas moleculares, a fim de se evitar ambiguidade com o nome Monte Carlo (MC), as referências à mecânica clássica serão feitas como Mecânica Molecular Clássica (MMC). O fato da DM ser baseada em MMC é de grande importância, tendo em vista sua simplicidade em comparação com os métodos quânticos, os quais embora mais precisos possuem custo computacional extremamente alto. Na MMC, não se tem a informação da parte eletrônica como no método quântico, e é possível ter-se uma simulação atomística de sistemas orgânicos envolvendo centenas de milhares (ou milhões) de átomos [KK99].

Os algoritmos utilizados nos programas de DM consistem da solução numérica de equações de movimento ao longo do tempo, tendo como resultado uma trajetória ou sequência de fotos ou *snapshots* (coordenadas e momentos conjugados em função do tempo) do sistema em questão.

Em 1977, McCammon e colaboradores realizaram a primeira simulação de DM envolvendo proteínas. Essa simulação foi realizada *in vácuo*, e o tempo de simulação foi de 8,8 x 10^{-12} s [MGK77]. A partir de então, a técnica de DM vem se aperfeiçoando e, como consequência, os sistemas a serem simulados tornam-se cada vez mais realísticos. Se for traçado um paralelo entre a evolução da DM em relação especificamente à Ciência da Computação, fica claro que o avanço nas arquiteturas dos computadores, com a disponibilização de máquinas cada vez mais robustas, foi, vem sendo e continuará a ser de suma importância para que os avanços na área da química, *i.e* aprimoramento de parâmetros dos campos de força (ver 2.5.1), tenham real possibilidade de ocasionar avanços em termos de resultados de pesquisa. Atualmente, é possível a realização de simulações mais longas, chegando a 10^{-9} e 10^{-8} s.

A DM tornou-se ferramenta importante e vastamente utilizada por profissionais de áreas como a química, física, biofísica e biologia, auxiliando na modelagem de minúcias microscópicas relativas ao comportamento dinâmico de uma gama de diferentes sistemas incluindo gases, líquidos, sólidos, superfícies e aglomerados [TM99].

Além de predição de estruturas proteicas, a DM é empregada em diversas áreas, como o refinamento de estruturas cristalográficas, otimização de parâmetros geométricos, avaliação da interação ligante-receptor, entre outras. O *software* AMBER14 [CCID$^+$05] contém parte dos programas utilizados para realizar todas as simulações de DM desta tese.

Campos de Força

A descrição mais simples de mecânica molecular é considerar a aproximação de Bohr & Oppenheimer. A aproximação de Bohr & Oppenheimer considera a movimentação dos núcleos como sendo mais lenta que a movimentação dos elétrons, sendo possível então separar a informação nuclear e eletrônica em duas partes, calculando-as separadamente.

Dessa aproximação (da mecânica quântica), constata-se que, em se tratando de MMC, a energia total do sistema depende exclusivamente da posição dos átomos do sistema, não se computando explicitamente os efeitos eletrônicos. A energia total desse sistema é dada via um potencial (nuclear) dependente das posições (ou conformação), mais conhecido pela denominação de campo de força (CF) ou *force field* [KSB+99].

O CF é uma peça fundamental no decorrer de uma simulação. Seja qual for o método que se estiver usando para varrer o espaço de energia que o campo de força possibilitará ser acessado, esse campo de força deve ser adequado ao tipo de sistema que se está simulando. Os componentes dos CF são, normalmente, compostos por termos harmônicos (comprimentos, ângulos de ligação) e uma função periódica contínua no intervalo completo de ângulos possíveis (de 0 a 360 graus) para diedros [Fie07].

Para a interação entre os átomos não-ligados são utilizadas as interações de van der Waals e eletrostáticas. As interações de van der Waals são modeladas, no AMBER, pelo do potencial 6-12 de Lennard-Jones e as interações eletrostáticas pelo termo de Coulomb. A soma dos vários termos de energia descreve a função de energia potencial que permite calcular a energia potencial total do sistema com base em sua estrutura tridimensional. A Equação 2.1 a seguir demonstra uma função de energia potencial P(r) típica:

$$P(r) = \sum P_l + \sum P_\theta + \sum P\phi + \sum P_{vdW} + \sum P_{elet} \tag{2.1}$$

sendo que P_l é a energia de estiramento da ligação em relação a seu valor de equilíbrio (ou ideal), P_θ é a energia de deformação do ângulo de ligação em relação a seu valor de equilíbrio, P_ϕ é a energia devido à torção em torno de uma ligação, P_{vdW} representa a energia das interações de van der Waals e P_{elet} representa as energias de atração/repulsão eletrostática entre duas cargas. Nos campos de força de classe I, os termos dos átomos ligados possuem a forma da equação 2.2:

$$P_l = P_\theta = kx^2 \tag{2.2}$$

onde x pode assumir valores de distância (l) ou ângulo de ligação (θ). Já se tratando de uma torção, a forma de seu potencial é dada pela Equação 2.3:

$$P_\phi = \frac{P_n}{2}(1 + \cos(n\phi - y) \tag{2.3}$$

onde P_n é a barreira de energia para a torção, n é o número de máximos (ou mínimos) de energia em uma torção completa, ϕ é o ângulo diedro, e y é o ângulo de fase (defasagem no ângulo diedro que pode gerar um ponto de mínimo ou de máximo na posição $\phi = 0$) [VGB90]. O parâmetro dependerá do tipo de torção considerada e, geralmente, não excede o valor 3, sendo que alguns CF adicionam ainda um quarto potencial harmônico a fim de evitar certas oscilações, o chamado "potencial torcional impróprio". Para mais informações sobre as demais classes de CFs, ver [PJW03].

Uma das representações funcionais dos termos de van der Waals e eletrostático refere-se respectivamente aos potenciais de Lennard-Jones (conhecido também como 6-12) e de Coulomb. O cálculo das forças relativas a interações de átomos não ligados é um processo próximo do limite assimptótico de complexidade O(n) [TD11]. Para dois átomos e i e j, tem-se a Equação 2.4:

$$P_{vdW} = 4\varepsilon \left[\left(\frac{\sigma}{r_{ij}} \right)^{12} - \left(\frac{\sigma}{r_{ij}} \right)^{6} \right] \tag{2.4}$$

onde o parâmetro ε governa a força da interação e define uma escala de distância onde o potencial interpartícula entre i e j é zero, o que ocorre quando $r_{ij} = \sigma$. O termo elevado a 12^a potência é dominante a distâncias pequenas e modela a repulsão entre i e j, ocorrida devido à sobreposição de orbitais eletrônicos. A atração fica por conta do termo elevado à 6^a potência.

$$P_{elet} = \frac{q_i q_j}{4\pi\varepsilon_0\varepsilon_r r_{ij}} \tag{2.5}$$

No caso das interações eletrostáticas (Equação 2.5), q_i e q_j correspondem à magnitude das cargas pontuais de cada átomo, r_{ij} à distância entre as cargas, ε_0 à permissividade do espaço livre e ε_r à constante dielétrica relativa do meio. Os campos de força mais utilizados atualmente são AMBER [CCB+95], CHARMM [BBO+83], GROMOS [CHB+05] e OPLS [JMTR96], entre outros.

Raio de Corte

As simulações moleculares ainda estão longe (e devem permanecer assim neste aspecto) dos sistemas reais no que diz respeito à quantidade de partículas. Um sistema de DM, se comparado a um sistema macroscópico, possui número de átomos insignificante. Desse modo, em um sistema macroscópico, apenas uma pequena quantidade de átomos está localizada próximo das paredes da superfície. Já em uma DM, o número relativo de partículas na superfície do sistema é muito maior, e o comportamento dos sistemas de DM é dominado por efeitos da superfície. A maneira mais comum de minimizar esses efei-

tos, quando se usa solvente explícito, é a utilização de Condições Periódicas de Contorno (CPC).

Usar CPC implica colocar os átomos do sistema em uma caixa a qual pode ter variadas formas geométricas, normalmente toroidal, ou seja, uma caixa com dimensões específicas em x, y, z e replicada em todas as direções (chamadas de imagens). Ao passo que a DM se desenvolve, quando um átomo se move na caixa original, caso ele adentre um espaço que seria de superfície (fora dos seus limites), esse átomo adentra uma das caixas imagem, conservando o sistema. Os efeitos de superfície são então eliminados. Uma vez utilizando CPCs, cada partícula na caixa de simulação estará interagindo não apenas com outras partículas de sua caixa, mas também com suas imagens. Aparentemente, o número de pares interagindo cresce enormemente. Contudo, esse inconveniente pode ser superado ao utilizar-se um potencial que possua um alcance finito, ou seja, em que a interação entre duas partículas dispostas a uma distância maior do que certo valor seja ignorada. Essa distância é referenciada pelo nome de raio de corte ou *cut-off* [Beu11]. O raio de corte é necessário para evitar que uma partícula enxergue mais de uma imagem ao mesmo tempo.

Há muito tempo tem sido estudada a influência do tamanho do raio de corte no cálculo de interações de longo alcance em simulações por dinâmica molecular [BAD05, DKAA+11, NdSO99, Sai94, SS92a, SS92b, SS92c]. S. Piana e colaboradores [PLLD+12] demonstram que a energia livre de enovelamento é relativamente insensível a raios de corte além de 9,0 Å. Já Yuan e colaboradores [YCK12] avaliaram 45 definições de contato variando as distâncias de raio de corte e descobriram que “em geral, se pode distinguir melhor entre enovelamentos quando os contatos são definidos por pares de resíduos cujos átomos estão a 7,0 Å ou menos de distância entre si e que um raio de corte maior é muitas vezes vantajoso para a captura do arranjo espacial de estruturas secundárias”. Esta tese visa trazer contribuições para o entendimento da influência exercida pelo raio de corte em simulações.

Dinâmica Molecular aplicada ao problema PSP

DM vem sendo amplamente utilizada por pesquisadores na área de PSP, embora a maioria dos trabalhos utilize-a apenas para refinamento [Dal12, DBL13, Fer14, JWS08, KDN+04, MHS12, MBFP12, MNF14, PGW+12], como os trabalhos de Lee e colaboradores, em que a DM é aplicada no refinamento de 12 pequenas proteínas [LTBK01]. Entretanto, além da aplicação em refinamento, existem trabalhos em que a DM é aplicada a fim de se chegar a estruturas nativas [EHLSW02], foco desta tese.

Um dos grandes marcos na aplicação de DM a PSP foi a simulação de enovelamento de 36 resíduos que formam a região c-terminal da molécula de vilina humana ou *villin headpiece* [DK98]. Hegler e colaboradores apresentam um método híbrido que combina informações de bancos de dados e mecânica molecular, em que restrições conformacionais

são testadas [HLS+09]. Embora a maioria dos estudos seja limitada a peptídeos e proteínas realmente pequenas [Bro02, KFNH08, LB02, RGFP09, VRS03, YSG09, ZAH05], surgem ainda métodos com sucesso na obtenção de estruturas de alta-resolução, e *ab initio*. É o caso dos trabalhos do grupo do professor Simmerling [SSR02], de Pietra e Swope [PS03] e de trabalhos do grupo do professor Duan [CLXD03], nos quais predições de alta resolução (com RMSD menores ou iguais a 2,0 Å foram obtidas para o peptídeo gaiola de triptofanos, composto de 20 resíduos, utilizando-se diferentes versões dos campos de força do AMBER e modelos de solvatação Generalized Born (GB) [TC00]. A título de exemplificação, pode-se citar os trabalhos do grupo do professor Duan, em que a estrutura foi enovelada a menos de 0.5 Å em comparação à nativa [LWLD07].

Como se pode observar, a partir do trabalho de Duan e Kollman, em 1998 (citado anteriormente como marco da DM aplicada à PSP), diversos trabalhos perceberam a capacidade do método, gerando grande avanço na área, partindo-se de pequenos (*vilin head-piece*) [BBBP09, ZSSP02] até proteínas maiores desde 28, 47 até 60 resíduos, como o domínio B da proteína A (BdpA) [LWWD09, LWW+08]. Nessa época, percebia-se que o enovelamento *ab initio* possuía grande capacidade de predizer hélices, porém proteínas com múltiplas estruturas secundárias mostravam-se como um desafio. Não obstante, os resultados para pequenas proteínas, em termos de RMSD, tornava o campo encorajador, sugerindo que, com a melhoria dos campos de força, as simulações obteriam, com o passar dos anos e não em um futuro tão distante, uma gradativa melhoria no grau de acerto. Foi quando, em 2009, o trabalho de Ken Dill e colaboradores alcançou, por meio da utilização de solvente implícito em uma técnica *ab initio*, resultados médios com precisão compatível a técnicas baseadas em conhecimento [OWCD07, SSBOV+09]. Esse foi outro marco da aplicação de DM ao problema da predição tridimensional de estruturas de proteínas, fator notavelmente encorajador às pesquisas.

Como exemplo de trabalhos subsequentes na área, pode-se destacar o de Lindorff-Larsen e colaboradores, os quais alcançaram estados de estabilidade termodinâmica para 12 proteínas, por meio de dinâmica molecular extensiva, em solvente explícito [LLPDS11]. Importante destacar, no entanto, a limitação do trabalho, restrito a um conjunto de pequenas proteínas [RPE+12]. Outro exemplo a ser destacado é o esforço de Shaw e colaboradores na montagem de ANTON [SDD+08], supercomputador específico para DM que tornou possíveis simulações de enovelamento *ab initio* em escala detalhada [SMLL+10], resultando até mesmo em pesquisas para avaliar seus resultados, constatando a capacidade da DM de seguir distribuições de enovelamento teóricas e explicar uma gama de resultados experimentais [HBE13].

Atualmente, mesmo com os computadores mais poderosos, a maneira de efetuar as varreduras conformacionais ainda é muito limitada, o que sustenta mais uma vez a ideia de que, com a inclusão de melhores campos de força e melhores modelos para interações com água, ao passar dos anos, melhores resultados emergirão [DM12]. Mais informações

sobre DM aplicada ao problema PSP podem ser obtidas das revisões de Lee e colaboradores e Zhou e colaboradores [LDK01, ZDY+11].

2.5.2 Monte Carlo

Outro método computacional para otimização e, no presente caso, descoberta de um mínimo global em termos de energia dentre uma extensa gama de conformações existentes no espaço de configuração de um sistema composto por um polipeptídio inicialmente estendido é o método de Monte Carlo (MC). O domínio de uma função de energia pode ser dividida em regiões, e para cada região pode-se ter um mínimo local diferente: a Figura 2.6 ilustra uma hipotética função unidimensional, onde se apresentam três regiões, cada uma associada a um mínimo local A, B ou C. Pelo menos um caminho existe para cada ponto em uma região conectando-o com um mínimo local de tal forma que uma vez em direção a esse mínimo o valor da função não mais aumenta. Começando-se do ponto P_1, por exemplo, se vai chegar até A, enquanto começando de P_2, se vai chegar até B. Para encontrar o mínimo global A começando de P_2, é necessário subir até um máximo local antes de cair em A. Uma maneira de localizar o mínimo global nesse caso seria executar a função iniciando aleatoriamente de vários pontos diferentes, esperando que um desses pontos leve até uma região de mínimo global. Para problemas envolvendo um número pequeno de variáveis, essa pode ser uma maneira confiável de identificar o mínimo global; entretanto, o problema da predição de estruturas é excessivamente complexo, tornando o esquema ineficaz [ZB07].

O método de Monte Carlo permite que os movimentos sejam feitos em qualquer direção e especifica uma probabilidade para cada um desses movimentos. Por exemplo, definindo-se um estado 1 pela posição de todos os átomos do sistema, se vai ter uma energia E_1 relacionada a ele. Quando o sistema está em equilíbrio, a probabilidade relativa de um dado estado 1 ocorrer é dada pelo fator de *Boltzmann* $e^{-(-E_1)/\kappa t}$, onde k é a constante de *Boltzmann* e T é a temperatura absoluta em Kelvin (K). A partir disso, resolvendo-se comparar o estado 1 com um estado 2 considerando uma energia E_2, a relação de probabilidade seria dada pelo seguinte termo:

$$e^{-(E_2-E_1)/\kappa t} = e^{-\Delta E_{21}/\kappa t} \tag{2.6}$$

Partindo-se do estado 1, pode-se facilmente determinar se o novo estado 2 é mais provável ou não de ocorrer em equilíbrio. Se ΔE_{21} é negativo (estado 2 possui menor energia), o numerador terá um valor maior que 1 (definindo o estado 2 como estado mais provável), e o movimento para o estado será aceito. Se o estado 2 possui energia maior que 1 (o movimento está sendo para um valor de energia acima do atual), o numerador possuirá um valor entre 0 e 1 e, ao invés de simplesmente acontecer a rejeição do estado 2 pelo fato do movimento ser não favorável, há a escolha de um número aleatório em uma distribuição

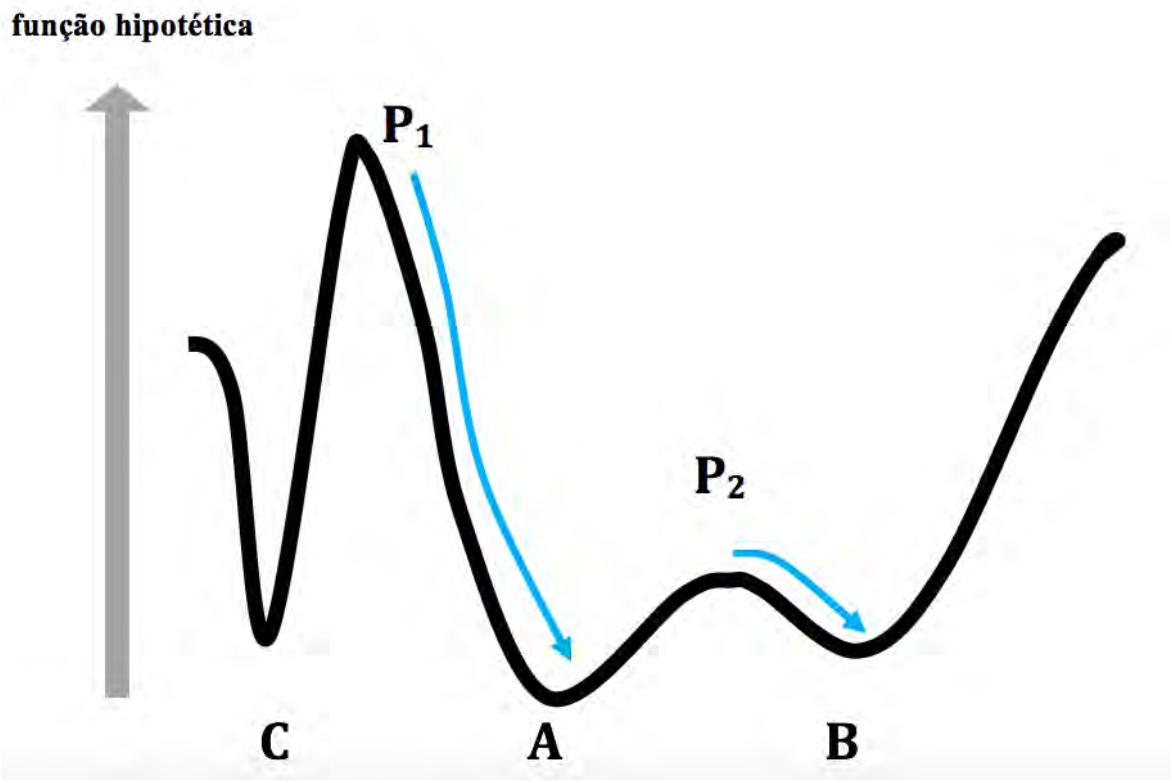

Figura 2.6 – Diagrama ilustrando o problema do mínimo global unidimensional, adaptado de [ZB07]. A função mostrada contém três mínimos: A, B e C, onde A é o mínimo global. O mínimo encontrado por uma otimização depende do ponto de início e da topologia da superfície. Se uma otimização é iniciada em P_1, chegará até A. Entretanto, se começar em P_2, o cálculo levará ao mínimo B.

uniforme no intervalo de [0,1], e, se esse número for menor que o número gerado pelo fator de Boltzmann (Equação 2.6), o movimento é aceito, sendo, caso contrário, rejeitado. Selecionando os movimentos dessa maneira, o método de Monte Carlo tem condições de, sob condições adequadas (não é o caso e será explicado melhor mais à frente), localizar a região do mínimo global energético, o qual seria o estado de melhor probabilidade [ZB07].

Monte Carlo aplicado ao problema PSP

Com o intuito de simplificar o problema em termos computacionais (comparado a DM), o método de Monte Carlo surgiu como alternativa bastante atrativa para a comunidade de PSP. Assim como a DM, MC é aplicado em variadas etapas da predição, servindo a diversos propósitos que vão desde refinamento [CCOS06, OS14], predição de estruturas secundárias [HVKS14, LAW+12, LSW+09], predição da conformação de cadeias laterais [NRB12] ou como principal técnica na busca por estruturas 3D nativas [AT94, CTTM03, CHLL03, GCS01, HPLS02, JBS+06, LPNdS12, LPNdS14, NBBJ06, PM97, ZLC+07].

2.5.3 REMD: Replica Exchange Molecular Dynamics

Em 1996, Hukushima e Nemoto desenvolveram um método, o qual nomearam Exchange Monte Carlo [HN96]. Similar à Têmpera Simulada ou *Simulated Tempering* [MP92], Replica Monte Carlo [SW86] ou métodos de *ensemble* expandido [LMSVV92], o algoritmo tem como objetivo a superação de barreiras de energia dentro do espaço de configuração, utilizando-se para isso de uma gama de diferentes temperaturas. Nos anos seguintes, os trabalhos de Hansmann [Han97] e Sugita e Okamoto [SO99] desenvolveram uma formulação do método Replica Exchange para Dinâmica Molecular ou Replica Exchange Molecular Dynamics (REMD), também atualmente conhecido como Método de Múltiplas Cadeias de Markov ou *Multiple Markov Chain Method* (MMCM). Desde então, o método vem sendo utilizado em diversos ramos da Bioinformática, desde estudos estrutura-função [MS15], DNA [MSLS14], RNA [BHR+14, RBC14], estudo da estabilidade de proteínas [HSD14], dinâmica de enovelamento [EG14, JSJ14, XYZ15] e predição de estruturas secundárias [ZS15]. A Figura 2.7 demonstra a visão geral de uma simulação REMD.

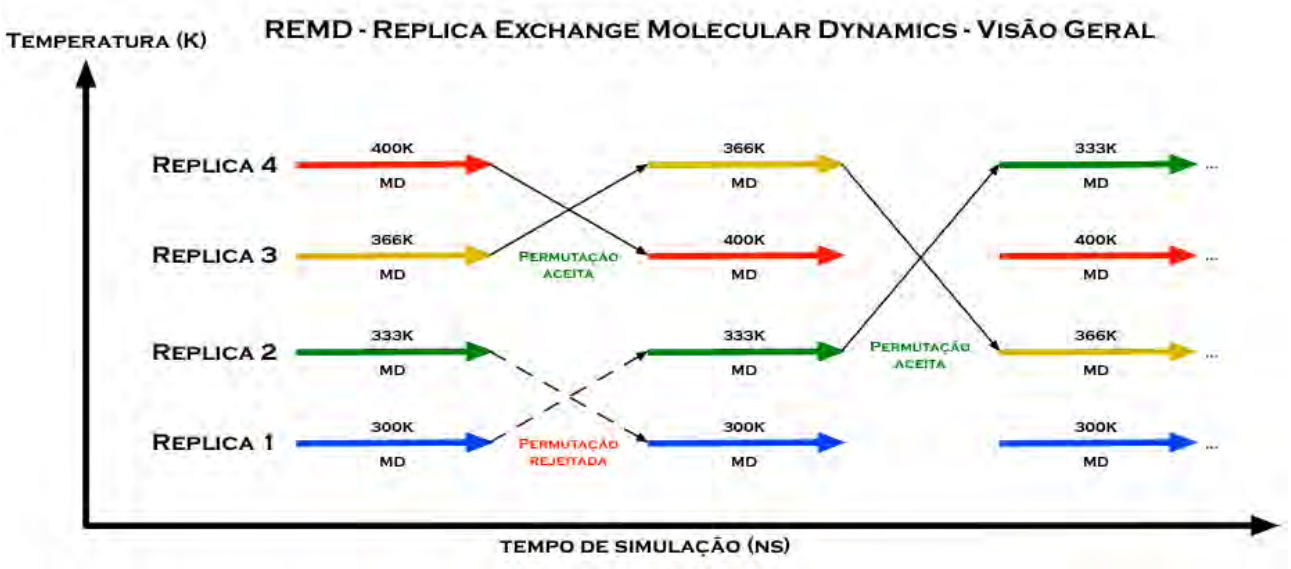

Figura 2.7 – Visão geral do método REMD convencional. Réplicas independentes são simuladas por DM e, a cada certo intervalo de tempo (ns), há a tentativa de intercâmbio entre as estruturas imersas em banhos térmicos de diferentes temperaturas.

Em uma simulação por REMD típica, um conjunto de simulações regulares de DM iniciam-se de forma independente, cada uma com sua configuração (estrutura), a certas temperaturas, no tempo 0.

Então, a uma frequência estipulada pelo usuário intitulada *EAF* ou *Exchange Attempt Frequency*, como por exemplo a cada 1 ps, uma tentativa de intercâmbio de conformações entre temperaturas adjacentes é executada.

Em relação a esse procedimento de intercâmbio, foi demonstrado que, para um sistema convergir no sentido de uma distribuição de equilíbrio, é necessário e suficiente

para um método acatar a chamada “condição de equilíbrio” ou *Balance Condition - BC*, onde BC simplesmente requer que uma distribuição de Boltzmann seja mantida [MD99].

Entretanto, a probabilidade de aceitação de REMD, computada pelo critério de Metropolis (ver 2.5.2, Equação 2.6), garante não apenas BC como também a chamada “condição de equilíbrio detalhada” ou *Detailed Balance Condition (DBC)* expressa na Equação 2.7, a qual declara que a probabilidade de aceitar uma tentativa de intercâmbio deve ser a mesma que a probabilidade de aceitar o movimento inverso [SO99].

$$P(n)^A P(m)^B \rho(n^A \rightarrow m^B) = P(m)^A P(n)^B \rho(m^B \rightarrow n^A) \tag{2.7}$$

Onde $\rho(n^A \rightarrow m^B)$ é a probabilidade de transição entre os estados A e B, e $P(n)^A$ é a população do estado A na temperatura n. Se uma tentativa de intercâmbio é bem sucedida, as temperaturas das réplicas envolvidas são trocadas e uma reescala uniforme das velocidades de todos os átomos nas réplicas, pela raiz quadrada da razão das duas temperaturas, é realizada [SO99].

De todo modo, depois de certo tempo de integração, um novo conjunto de configurações é obtido. Nesse momento, um intercâmbio é avaliado pelo critério de Metropolis. Essas duas etapas (DM seguida de MC) são repetidas até que se entenda que a simulação está terminada. O procedimento pode ainda ser entendido como um processo de Markov com dois operadores: (M) operador relacionado à geração de resultados por DM; e (P) operador relacionado à tentativa de intercâmbio entre duas configurações q_0 e q_t. Tratar-se-ia então de uma cadeia de Markov em que as propriedades termodinâmicas e conformacionais se adequariam à seguinte regra:

$$q_t = (PM)^t \, q_0 \tag{2.8}$$

Na prática, apenas intercâmbios entre temperaturas adjacentes são permitidas, na tentativa de aumentar a taxa de aceitação. Uma excessão a essa regra é o fato de REMD aceitar, normalmente, tentativas de intercâmbio entre a temperatura mais alta e a mais baixa.

Além de na predição de estruturas proteicas, a REMD é empregada em diversas áreas, como o refinamento de estruturas cristalográficas, otimização de parâmetros geométricos, avaliação da interação ligante-receptor, entre outras.

Em relação à eficiência de simulações REMD, por exemplo, Periole e Mark [PM07], em comparações com a DM convencional de um β-heptapeptídeo em solvente explícito, encontraram que “para determinar populações em baixas temperaturas (275–300 K), a técnica de REMD foi, no mínimo, oito vezes mais eficiente que DM, para este sistema”.

Zhang, Wu e Duan [ZLC+07] estudaram um peptídeo de 21 resíduos da classe beta, em solvente implícito e reportaram que: “Em comparação com DM convencional,

REMD pode significantemente melhorar a eficiência de busca em 14,3 $\pm$ 6,4, 35,1 $\pm$ 0,2 e 71,5 $\pm$ 20,4 vezes nas temperaturas aproximadas de $\sim$360, 300, e 275 K, respectivamente".

Sanbonmatsu e Garcia [SG01] estudaram um pentapeptídio em solvente explícito e constataram que o espaço de busca percorrido por REMD "é aproximadamente 5 vezes maior, durante o mesmo tempo", o que sugere um aumento mínimo na eficiência, utilizando REMD, de um fator de 5.

Rao e Caflisch [RC03] estudaram uma proteína de 20 resíduos, esta da classe beta. O tempo médio para enovelamento dela foi de 0,064–0,067 µs com REMD, e 0,085 µs com DM convencional. Seibert e colaboradores testaram longas simulações de um *beta-hairpin*, com modelo de água explícita [SPHvdS05]. A fase de equilíbrio foi obtida depois de centenas de nano segundos de simulação por réplica, já com DM, após 1–2 µs.

Ainda em termos de eficiência, a gaiola de triptofanos, ao ser simulada, demonstrou resultados similares: utilizando REMD, a estabilização da estrutura ocorreu em cerca de 100 ns de simulação, e com DM convencional foram necessários µs [PNG07]. Uma extensiva análise sobre a eficiência de simulações REMD pode ser obtida em [Nym08]. De todo modo, fica evidente que, utilizando REMD, as simulações provavelmente devem encontrar estados de equilíbrio mais rapidamente, se comparado à DM convencional.

REMD aplicado ao problema PSP

Entre as principais técnicas computacionais atualmente aplicadas ao estudo de proteínas, conforme já mencionado antes, REMD desempenha grande papel, uma vez que fornece amostragem conformacional eficiente. No entanto, tais abordagens são frequentemente limitadas à investigação dos caminhos de enovelamento das proteínas ou *protein folding* e não são aplicados à PSP. Dois exemplos de trabalhos desse tipo, os quais inclusive fazem parte dos métodos *ab initio* alvos de comparação com este estudo, são os de Seibert *et al.* e Suenage *et al.* [Sue03, SPHvdS05].

De todo modo, por meio do protocolo de mapeamento sistemático exposto no Apêndice A, foi possível a pesquisa e descoberta estruturada dos trabalhos presentes na literatura que endereçam, de algum modo, a obtenção de estruturas 3D aproximadas capazes de representar a estrutura nativa de proteínas. Destaca-se, em primeiro momento, o fato de novos campos de força estarem sendo desenvolvidos pela comunidade para dar suporte a simulações REMD objetivando melhores predições e melhor amostragem, como é o caso dos trabalhos de Zhou, Jiang e Wu, e Mou *et al.* [Zho04, JW14b, MJG+14].

Dando seguimento à explanação acerca dos métodos que utilizam REMD para PSP, existem abordagens que fazem uso de constrições - como os trabalhos de Gront *et al.* e Balaraman *et al.* [GKH05, BPJV11] -, restrições - como o trabalho de Raval *et al.* [RPE+12] - ou intercâmbios auxiliados por avaliação de hidrofobicidade, caso do trabalho de Liu e colaboradores [LHZB06]. Enquanto isso, Zacharias e colaboradores aplicaram com

êxito potenciais enviesados para predizer e refinar estruturas de proteínas [OZ14, KZ09b, KZ09a, KZ07, KZ10]. Já Ding e colaboradores utilizaram-se de um tipo específico de REMD baseado em DM discreta para predizer a estrutura de 6 pequenas proteínas [DTND08].

A combinação de REMD com dados semiconfiáveis [MPD15] ou com diferentes métodos, assim como *Umbrella Sampling* [JSJ14], e dinâmicas autoguiadas de Langevin (*self-guided Langevin dynamics*) [LO10], também vem sendo aplicada, bem como a combinação entre diferentes níveis de abstração [VS12]. Grupos como o do professor Ken Dill também têm explorado o problema utilizando REMD, com destaque para os trabalhos de Ozkan *et al.* [OWCD07] e seu estudo sobre o mecanismo de *zipping and assembly* em proteínas orientado a predição, e o trabalho de Perez *et al.* [PMD15] composto pela combinação de REMD com inferências Bayesianas derivadas de estruturas secundárias e informações adicionais (como por exemplo, o fato de proteínas possuírem núcleos hidrofóbicos).

Previamente limitado a miniproteínas ou pequenos fragmentos [UUAD08, HD06, FWT02, YP03], um trabalho recente de Shaw e colaboradores mostrou simulações atomísticas por DM serem bem sucedidas quando aplicadas a proteínas maiores, ainda que não sejam proteínas grandes. Em seu trabalho, Shaw e colaboradores [LLPDS11] realizaram simulações utilizando solvente explícito por longos períodos de tempo, utilizando a infraestrutura do supercomputador de propósito específico Anton [SDD+08]. Tal trabalho tornou possível o estudo de dinâmicas de enovelamento e a predição da estrutura 3D de 12 pequenas proteínas. Apesar das melhorias alcançadas em *hardware* e *software* nos últimos anos, o método de simulação REMD ainda é computacionalmente caro, especialmente para simulações de solvente explícito envolvendo todos os átomos.

As principais alternativas para os altos custos computacionais de simulações *all-atom* com solventes explícitos e atomísticas são o uso de abstrações (*coarse-grained models*) e a utilização de solvente implícitos [MSC+10, JSJ14, SKS+15]. Recentemente, Nguyen e colaboradores [NMH+14] provaram ser possível enovelar proteínas com diversas topologias e tamanhos variando de 10 a 92 aminácidos utilizando solvente implícito e REMD e, ainda que existam limitações relativas aos CFs a serem empregados, pesquisadores da área acreditam que o poder das simulações de enovelamento baseado em funções de energia deve continuar a crescer [PMSD16].

2.6 Software para Simulação Molecular de Proteínas: AMBER14

O AMBER [CCID+05, PCC+95] é um exemplo de pacote de programas de simulação molecular que permite aos usuários executar e analisar simulações de DM para proteínas, ácidos nucleicos e carboidratos. Basicamente, é composto por duas partes: (i) um conjunto de campos de força e (ii) um conjunto de programas de simulação. Em uma

simulação típica do AMBER, encontram-se três etapas: (i) preparação do sistema; (ii) simulação; e (iii) análise de trajetória. O AMBER fornece suporte à DM com solvente implícito e explícito [Nym08], lembrando que comumente os modelos de solvente implícito são consideravelmente menos onerosos computacionalmente. A implementação de modelos de solvente implícito é dada pelas aproximações de Poisson-Boltzmann e Generalized Born [OCB02, STHH90], enquanto os modelos de solvente explícitos são tratados pelo método chamado Particle-Mesh Ewald (PME) [DYP98]. Esta tese utilizou o AMBER em sua versão 14.0.

2.7 Medidas de Avaliação da Qualidade de Modelos

A fim de facilitar a leitura da tese, esta seção está dividida em duas subseções. A primeira apresenta apenas medidas aplicadas efetivamente neste trabalho e a segunda apresenta medidas que, embora não tenham sido aplicadas para fins de análise no trabalho, são disponibilizadas pelos *softwares* de apoio oriundos desta tese (suíte de *scripts* CuT-REMD e interface gráfica GTK-REMD).

As medidas a seguir podem ser ainda subclassificadas em dois tipos: relativas e absolutas. Entende-se por medida relativa aquela que necessita de uma estrutura de referência para ser calculada. Entende-se por medida absoluta aquela que não necessita de uma estrutura de referência para ser calculada. Por conseguinte, são medidas que podem estar inclusas em procedimentos automáticos para triagem de estruturas nativas.

2.7.1 Medidas Aplicadas

RMSD

O desvio quadrático médio, do Inglês: *Root-Mean-Square Deviation* (RMSD), é a medida da distância média entre os átomos de proteínas sobrepostas. É a medida mais comum no que se trata da comparação de estruturas de proteínas. A Equação 2.9 mostra como o cálculo de RMSD é feito.

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\delta_i^2} \tag{2.9}$$

Onde δ é a distância entre N pares de átomos equivalentes. Normalmente, esses pares são formados por átomos referentes aos carbonos alfas (Cαs) ou ao esqueleto da cadeia (C, N, O, Cβ). É comum também que, durante o cálculo de RMSD, sejam efetuadas rotações e translações em uma das proteínas, com o intuito de se obter a melhor sobreposição, a qual minimiza o RMSD. Dados dois conjuntos *v* e *w* de *n* pontos, o RMSD é definido

pela Equação 2.10 e o valor retornado é expresso em uma unidade de medida de distância, usualmente o Angström (Å), que equivale a 10^{-10} m.

$$RMSD(v, w) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} ||v_{ix} - w_{ix}||^2}$$
$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{n}\left((v_{ix} - w_{ix})^2 + \left(v_{iy} - w_{iy}\right) + (v_{iz} - w_{iz})^2\right)} \quad (2.10)$$

GDT

Outra medida para avaliar o quão similar uma topologia de proteína é da outra é o GDT. O nome do método vem do Inglês *Global Distance Test* ou Teste de Distância Global, e seu algoritmo leva em consideração diferentes valores para raio de corte [Zem03]. O GDT é calculado por meio da Equação 2.11:

$$GDTscore = (C1 + C2 + C3 + C4)/4N \quad (2.11)$$

Onde C1 é o número de resíduos sobrepostos levando-se em consideração um raio de corte igual a (*threshold*/4), C2 é o número de resíduos sobrepostos levando-se em consideração um raio de corte igual a (*threshold*/2), C3 é o número de resíduos sobrepostos levando-se em consideração um raio de corte igual a (*threshold*), C4 é o número de resíduos sobrepostos levando-se em consideração um raio de corte igual a (*threshold* * 2) e N é o número total de resíduos. O valor resultante do cálculo de GDT (TS ou com *threshold* = 4) possuirá valores variando de 0 até 1, onde valores de até 0,2 são tidos como a sobreposição aleatória de estruturas, e valores acima de 0,6 demonstram estruturas de alta similaridade. No presente trabalho, foi utilizado o limiar de 4,0 Å (GDT_TS ou GDT Total Score).

2.7.2 Medidas Disponibilizadas

A seguir são apresentadas as medidas disponíveis na "suíte de *softwares* CuT-REMD" (descrita em 5.2.2) e na interface gráfica GTK-REMD (5.2.3). Destaca-se que medidas RMSD e GDT (apresentadas em seção anterior) integram também o conjunto de medidas disponibilizadas, o qual conta não apenas com medidas relativas (RMSD, GDT e QCS) como também com absolutas (DOPE, G-Factor, ProbScore, DFIRE, dDFire, RWPlus, OPUS-PSP e GOAP).

QCS

O *Quality Control Score* (ou simplesmente QCS) é um método desenvolvido que visa contribuir com o CASP em termos de avaliação automática das estruturas previstas. Essa métrica é considerada particularmente útil para comparar predições ruins. De acordo com seus autores, a métrica QCS reflete as experiências pessoais de avaliação manual de proteínas e visa capturar características globais de modelos definidos pelo arranjo mútuo de EES. Um componente de contato inter-resíduo está incluso no QCS, a fim de quantificar a precisão da modelagem de detalhes atômicos. Em geral, o QCS está de acordo com a inspeção manual e se correlaciona bem com o GDT_TS. No entanto, QCS pode revelar modelos com uma melhor topologia global desapercebidos para GDT_TS. Essa métrica não é apenas adequada para selecionar candidatos para inspeções manuais em futuras competições CASP, mas também pode ser útil como um método independente e objetivo para avaliar a qualidade da previsão de estruturas, com ênfase na topologia global [CKP+11].

DOPE

Discrete Optimized Protein Energy ou DOPE é potencial estatístico dependente da distância entre átomos. Embora derivado de um conjunto de estruturas nativas, DOPE não depende de parâmetros de ajuste, é um parâmetro quantitativo que indica o potencial estatístico da energia de estruturas tridimensionais [SB93]. A pontuação é não normalizada em relação ao tamanho das proteínas e possui uma escala arbitrária, assim as pontuações de diferentes proteínas não podem ser comparadas diretamente. Quanto mais baixo o valor do DOPE *score*, melhor é, teoricamente, o modelo [SB93].

G-Factor

O G-Factor [LMMT93] provê uma medida do quão “normal” ou, alternativamente, quão “não usual” uma estrutura é, em termos de estéreo-química. O G-Factor é, essencialmente, uma informação probabilística baseada em distribuições de parâmetros estéreo-químicos. Quando aplicado a um resíduo, um G-Factor baixo indica que se está lidando com uma conformação de baixa probabilidade. Assim sendo, resíduos com ângulos ϕ e ψ em regiões não permitidas no mapa de Ramachandran terão um G-Factor baixo, senão negativo. Se uma proteína possui muitos resíduos com G-Factors baixos, isso sugere que algo deva estar a prejudicar a geometria desta.

ProbScore

Molprobity é um serviço *web* de validação de estrutura que fornece uma ampla avaliação de modelos tanto em nível global quanto local, para proteínas e ácidos nucleicos. Baseia-se na verificação de posicionamento de hidrogênios e na análise de contatos *all-*

atom, complementada por versões atualizadas de geometria covalente e ângulo de torção [CAH+10]. A métrica Probscore é uma métrica criada com base em três pontuações diferentes geradas por esse *software*: avaliação de colisões, avaliação de rotâmeros *outliers*,e *outliers* em termos de Ramachandran.

DFIRE e dDFire

DFIRE significa *Distance-scaled, Finite Ideal-gas Reference* [ZZ02]. Essa métrica utiliza a estrutura de referência para construir um potencial *all-atom* baseado em uma base de dados de estruturas de proteínas não homólogas.

O dDFIRE é o DFIRE que considera interações do tipo dipolo, utilizando uma função de energia baseada na orientação dos ângulos envolvidos em interações dipolo-dipolo. No dDFIRE, cada átomo polar é tratado como um dipolo, e a função de energia é extraída de estruturas de proteínas com base na distância entre dois átomos e os três ângulos envolvidos nas interações dipolo-dipolo. Segundo seus autores, a métrica é capaz de prover tratamento consistente para a possível interação "orientação-dependente" entre átomos polares e apolares assim como entre átomos polares não ligados por pontes de hidrogênio [YZ08].

RWplus

O potencial RW é um potencial atômico par a par dependente de distância baseado em "caminhos aleatórios" ou *"random walks"* de uma cadeia ideal [ZZ10]. Segundo seus autores, como essa cadeia ideal não possui interações específicas entre resíduos de aminoácidos de diferentes subunidades, mantendo continuidade na sequência proteica, ela mimetiza a elasticidade entrópica e a conectividade genérica de uma proteína, o que não é possível fazer, por exemplo, com DFIRE ou DOPE.

OPUS-PSP

O OPUS-PSP é um potencial estatístico *all-atom* derivado do empacotamento de cadeias laterais. Possui um conjunto básico de 19 blocos de corpo rígido extraídos das estruturas químicas de todos os 20 aminoácidos. O potencial é gerado a partir das estatísticas de empacotamento de pares desses blocos em uma base de dados de estruturas não redundantes. Em geral, o OPUS-PSP é um potencial aplicável para a modelagem de estruturas de proteínas, especialmente para tratar das conformações de cadeia lateral, uma das etapas mais difíceis na predição e refinamento de proteínas [LDM08].

GOAP

A métrica GOAP é um potencial *all atom* geral dependente de orientação. Depende da orientação relativa entre planos associados a cada átomo pesado em pares de interação. Essa métrica é uma generalização de abordagens anteriores de potenciais dependentes de orientação que consideram apenas átomos representativos ou blocos de cadeias laterais e átomos polares. GOAP pode ser decomposta em contribuições dependentes de distância e de ângulo. De acordo com os autores, “o GOAP integra naturalmente a orientação-dependência entre interações entre átomos polares, pontes de hidrogênio e cadeias laterais” [ZS11].

3. MOTIVAÇÃO E OBJETIVOS

3.1 Motivação

Em 2012, chegou-se aos 50 anos do “nascimento de um dos grandes desafios da ciência básica”, o problema da predição de estrutura de proteínas, conforme enfatizado pela revisão publicada por Dill e MacCallum na revista Science no dia 23 de novembro de 2012. A revisão ainda enfatiza os avanços consideráveis obtidos no entendimento do problema e destaca o considerável valor da pesquisa de métodos precisos para a PSP a partir de sequências [DM12]. O problema PSP surgiu na década de 60 e até hoje sua solução continua sendo uma das principais pendências da biologia molecular [DeSBL14, XYZ15, ZS15]. Limitações dos principais métodos de determinação experimental da estrutura 3D de proteínas, como cristalografia por difração de raios X e ressonância magnética nuclear, destacam a importância do emprego de métodos computacionais para a predição da estrutura 3D de proteínas [WAA+14]. A solução do problema PSP, ou avanços no seu tratamento, permitirá a obtenção de estruturas 3D de proteínas importantes com aplicações relevantes na indústria biofarmacêutica, além de permitir a compreensão de proteínas envolvidas em processos vitais, incluindo doenças como o câncer [DK01]. Tendo em vista as dificuldades encontradas pelas abordagens tradicionais (experimentos *in vitro* e *in vivo*) no tratamento de problemas referentes a sistemas biológicos, a utilização de simulação computacional torna-se uma atraente alternativa, pois torna possível, por exemplo, a execução de experimentos *in silico* menos custosos, tanto em termos financeiros quanto de duração. O número de participantes no CASP, a cada dois anos, é uma indicação do constante aumento no número de interessados na solução de tal problema.

3.2 Objetivo Geral

O objetivo geral deste trabalho foi a criação de uma nova abordagem de predição de estruturas tridimensionais de proteínas: CuT-REMD. Como entrada deve ser fornecida apenas a estrutura primária ou sequência de aminoácidos de uma proteína. CuT-REMD deve então ser capaz de, utilizando-se de simulações de DM envolvendo o intercâmbio de estruturas em diferentes temperaturas ou *Replica Exchange Molecular Dynamics* – REMD, sem informações provenientes de bases de dados (forma *ab initio*), chegar a estruturas tridimensionais potencialmente capazes de representar a estrutura nativa de proteínas. A complexidade computacional de REMD deve ser mantida.

3.3 Objetivos Específicos

- Desenvolver e/ou modificar códigos dentro do pacote de simulações AMBER14, tornando possível simulações baseadas em uma nova abordagem: CuT-REMD;
- Investigar a capacidade da manipulação de raios de corte conferir maior eficiência em acessar estruturas próximas à nativa, em simulações por DM e REMD;
- Investigar a capacidade da manipulação de raios de corte conferir maior eficiência de predizer estruturas próximas à nativa, em simulações por DM e REMD;
- Investigar se o tempo de simulação necessário para encontrar estruturas enoveladas (em comparação à literatura) pode ser diminuído utilizando raios de corte menores e simulações mais curtas;
- Avaliar o impacto de temperaturas elevadas nas simulações REMD;
- Avaliar a influência de raios de corte mais curtos em EARs e ETRs;
- Avaliar diferentes tempos de permanência em raios de corte curtos;
- Avaliar diferentes intervalos de tentativa de troca (EAFs) e sua relação com simulações REMD utilizando raio de corte incremental;
- Investigar CuT-REMD quanto à correta adoção de estruturas secundárias;
- Investigar ergodicidade e reprodutibilidade na amostragem da superfície de energia em simulações CuT-REMD e REMD convencional;
- Definir um protocolo geral de preparação das simulações, sendo capaz de estimar o número de temperaturas diferentes, raios de corte e intervalos de incremento de raio de corte e tentativa de permuta entre as simulações REMD e de DM;
- Executar experimentos tendo como alvo um conjunto teste de proteínas oriundos do estado da arte no âmbito da predição *ab initio* de proteínas e, obtendo resultados satisfatórios, estender a análise a métodos *de novo*;
- Desenvolver uma solução gráfica de uso facilitado para dar suporte à configuração das simulações desta tese e de REMD em geral;
- Desenvolver uma aplicação de suporte à análise automática das simulações do tipo CuT-REMD (que deve envolver bancos de dados); e
- Disponibilizar uma biblioteca de *scripts* para execução de simulações, geração de gráficos e análise de simulações por CuT-REMD e REMD.

4. METODOLOGIA

A metodologia empregada para a realização desta tese teve como base a hipótese de Anfinsen para a termodinâmica, a qual relaciona a estrutura nativa de uma proteína com seu estado de menor energia livre [Anf73]. Para isso, foi utilizada uma função de energia com termos baseados em leis físicas e químicas de interação entre aminoácidos, em uma representação que inclui todos os átomos.

Em sua forma fundamental, o problema da predição de estruturas pode ser separado em duas partes distintas. A primeira refere-se à busca precisa e eficiente de amostrar o vasto espaço conformacional de uma proteína. A segunda parte refere-se a como discriminar com precisão entre estruturas de proteínas na forma nativa e não nativa [PL96]. Dentro da primeira parte, tem-se ainda a acepção de que métodos *ab initio* requerem, genericamente, três elementos [CRBB03, Osg00]:

1. uma representação geométrica da cadeia proteica;
2. uma função de energia; e
3. uma técnica para amostragem da superfície de energia.

Sendo a abordagem aqui proposta uma abordagem *ab initio* puro, será apresentado a seguir como este pretende atender a cada um dos requisitos enumerados acima, e, logo em seguida, como pretende lidar com a segunda parte do problema.

4.1 Representação Geométrica

É a maneira como a estrutura da proteína ou polipeptídio é representada computacionalmente. Quanto à representação da proteína a ser simulada, existem diversos níveis de abstração. Representações de modelos reduzidos ou *coarse-grained* (CG) vêm sendo objeto de interesse de pesquisadores no estudo teórico de simulações da estrutura e da dinâmica de proteínas [Cle08, CM06, Kol04, Toz05]. A primeira razão para tal é a de envolver esforços computacionais muito menores se comparado com simulações atomísticas de cadeias polipeptídicas, o que facilita a aceleração de simulações tanto de dinâmica quanto de enovelamento e termodinâmica de proteínas em quatro ordens de magnitude [LKS05]).

A representação mais detalhada possível inclui todos os átomos da proteína (*all atom*) e também as moléculas do solvente que a circunda, normalmente água. Quando todos os átomos das moleculas de água são representados individualmente, chamamos a simulação de uma simulação com solvente explícito, entretanto, calcular todas as interalções entre todos essas molécular requer custo computacional, uma razão pela qual tra-

balhos que envolvem água explícita terem grandes limitações no tamanho das proteínas [SPHvdS05, BBO+83].

Como alternativa ao uso de solvente explícito, existem vários modelos em que o solvente é modelado por campos de força que tratam as moléculas de água como átomos unificados, são os chamados solventes implícitos. Variados tipos de informações desde função a topologias podem ser obtidos sem a utilização de solvente explícito. A relação custo computacional e precisão tem se mostrado gratificante em simulações solvatadas implicitamente. Com solvente implícito é possível preservar as características principais de uma estrutura e ainda assim reduzir o tempo computacional necessário para as simulações. Essa foi a principal razão pela qual, neste trabalho, foi escolhido trabalhar-se com solvente implícito.

O pacote AMBER tem sido muito utilizado com processamento por placas gráficas. Desenvolvedores do AMBER14 atuaram em conjunto com os desenvolvedores de uma corporação desenvolvedora de placas, a NVIDIA. Um ponto importante a ser destacado é que, no entanto, o *software* não permite a edição de códigos referentes ao método REMD para GB em placas gráficas [GWX+12], o que fez necessário que as simulações desta tese fossem executadas exclusivamente em CPU. Uma vez que as placas gráficas possuem a capacidade de conferir às simulações um aporte de desempenho consideravelmente alto, essa é uma dificuldade a ser discutida em detalhes no futuro [SFGP+13].

4.2 Função de Energia

Outro ponto importante na descrição de um método é a função de energia a ser utilizada. É através dela que as conformações serão analisadas em termos de energia potencial e, levando-se em conta a hipótese de Anfinsen, é um dos fatores para diferenciar estruturas potencialmente perto ou longe do estado nativo. Na literatura, são encontradas duas categorias nas quais as funções se enquadram [ZS04b]: (i) potenciais baseados na Mecânica Molecular e (ii) funções estatísticas derivadas de estruturas reais.

A primeira categoria é fisicamente baseada em parâmetros obtidos normalmente de dados quânticos calculados em vácuo para pequenas moléculas. A segunda categoria é derivada empiricamente de estruturas experimentais do PDB [GHK00, HS99, KS95, LK00, MDK+99, Sip95]. Ambas as categorias representam forças que culminam na determinação das conformações macromoleculares e envolvem dois termos principais: relativo a átomos ligados (*bonded*) e relativo a átomos não ligados (*non-bonded*). Os termos *bonded* levam em consideração ligações, ângulos e torções. Já os termos *non-bonded* consideram ligações iônicas, interações hidrofóbicas e forças van der Waals, além de ligações dipolo-dipolo e de hidrogênio.

A principal vantagem de se utilizar funções de energia baseadas em conhecimento é o fato de se poder modelar o comportamento observado em estruturas conhecidas, mesmo que não exista bom entendimento físico sobre esse comportamento. Por outro lado, a desvantagem fica por conta da obtenção de novos comportamentos, o que pode não ser obtido. Existe uma gama considerável de funções de energia disponíveis na literatura. Dentre as principais estão: AMBER [CCB+95], CHARMM [BBO+83, MBN+98], GROMOS [CHB+05] e ECEPP [MMBS75].

O *software* utilizado nesta tese para executar simulações, o AMBER, foi desenvolvido para se adequar a vários tipos de campos de força. Suas parametrizações tradicionais usam cargas parciais fixas, centradas nos átomos, e são desenhadas especificamente para o tipo de sistema a ser simulado. Diversos grupos noticiaram que os conjuntos de parâmetros dos campos de força ff99 e ff94 não retornavam o devido equilíbrio energético entre regiões de hélice e estendidas. O campo de força ff94 possuía tratamento incorreto de parâmetros do esqueleto de glicinas. Por outro lado, o campo de força ff14SB, até o momento em que se decidia qual campo de força utilizar neste trabalho; era o recomendado pelos desenvolvedores do AMBER para proteínas e ácidos nucleicos e, assim sendo, foi escolhido como campo de força utilizado pelas simulações da tese. O ff14SB é uma continuação do antigo ff99SB [HAO+06].

4.3 Técnica de Amostragem da Superfície de Energia

Dentre os métodos *ab initio* tem-se, como técnicas de amostragem da superfície de energia, abordagens que envolvem DM, MC, AG e busca exaustiva/semi-exaustiva, dentre outras. Nesta tese, o método REMD convencional (método que combina Dinâmica Molecular e Monte Carlo e que otimiza a varredura do espaço, conforme exposto em 2.5.3) foi modificado a fim de se obter uma abordagem original a ponto de atender melhor ao problema da predição 3D de estruturas proteicas, e recebeu o nome de CuT-REMD.

O cerne da abordagem CuT-REMD são alterações em parâmetros envolvidos nos cálculos de energias, mais especificamente nos parâmetros referentes aos chamados raios de corte, iniciando-se as simulações com raios de corte pequenos (dando ênfase a enovelamentos locais) e gradativamente expandindo-os, com a finalidade de realçar as interações entre átomos mais distantes entre si, e consequentemente favorecer a compactação de estruturas secundárias. A nova abordagem foi criada da observação de pontos presentes no processo biológico da formação de proteínas reais.

Em meio biológico, a síntese de proteínas ocorre de forma gradativa. Durante o processo de tradução do mRNA, cada aminoácido é adicionado sequencialmente até que toda a estrutura primária esteja completa.Como exposto por Levinthal [Lev68], a proteína em formação busca “caminhos de enovelamento” para encontrar o estado enovelado bio-

logicamente necessário para sua função. Embora tais “caminhos” não sejam plenamente conhecidos, sabe-se que à medida que os aminoácidos são anexados à estrutura primária, surgem enovelamentos locais, iniciando a formação de estrutura secundária antes mesmo da tradução completa do peptídeo. Assim sendo, têm-se EES regulares, os quais, por sua vez, são formados e mantidos através de estabilização por ligações de hidrogênio dentro de hélices e entre fitas de folhas β. O empacotamento gradual desses elementos estruturais, permitido por voltas e alças flexíveis conectando-os, contribui para se chegar à estrutura funcional, nativa.

Interações iônicas, dipolo-dipolo, de van der Waals e hidrofóbicas, além de ligações de hidrogênio, são fundamentais para esses eventos. Como se pode ver nas equações 2.2 e 2.3, as interações moleculares são inversamente proporcionais as distâncias entre os átomos. Em grtsl, essas forças são maiores a uma distância de 4,0 Å. A fim de se priorizar o enovelamento, o efeito local pode ser descrito de forma aproximada com a redução do raio de corte durante a simulação de DM. Ao passo que a simulação se desenvolve, o raio de corte é aumentado, sendo mais efetiva a captura do arranjo espacial de estruturas secundárias. Esse é o cerne desta tese.

4.4 Captura e Apresentação da Estrutura mais Próxima da Nativa

Outro ponto importante em uma abordagem que visa à predição de estruturas é, seja qual for o método de exploração da superfície de energia, uma maneira de selecionar, dentre uma grande quantidade de estruturas geradas, aquela que representará a estrutura nativa.

O tempo total de simulação de cada REMD descrita neste trabalho foi limitado em 50 ns e, por esse motivo, é possível verificar que, mesmo no final da simulação, existem ainda flutuações em nível estrutural, pois o método segue buscando novos poços de energia para visitar. Assim sendo, entende-se que não seria uma abordagem adequada a captura da última estrutura da simulação como sendo o retorno do método preditivo. Isso posto, mostrou-se factível a ideia da estipulação de um protocolo de captura de estruturas baseado nos *ensembles* gerados.

Como o conjunto de trajetórias gerado forma uma grande quantidade de dados, percebeu-se a necessidade de um método de filtragem das estruturas, a fim de que as análises posteriores fossem feitas em um volume menor de dados. Foram feitas pesquisas a fim de encontrar, na literatura, *softwares* bem adaptados para o específico problema; entretanto, em se tratando de simulações REMD, estes não foram encontrados - apenas protocolos de clusterização envolvendo somente informações estruturais; e, assim sendo, optou-se pelo desenvolvimento de uma abordagem própria que considerasse também informações como as diferentes temperaturas em que os sistemas foram simulados.

A fim de criar um novo protocolo para a captura e apresentação de estruturas próximas à nativa, inicialmente, foram avaliadas as simulações geradas pelos protocolos de simulação propostos por esta tese em comparação aos protocolos de simulação baseados em métodos convencionais. As Figuras 6.9 e 6.10 (expostas em capítulo posterior referente a resultados) demonstram estudos iniciais executados em ao relação ao protocolo para captura e apresentação de estruturas próximas à nativa. Por meio dos gráficos (os quais compreendem não apenas CuT-REMD, como também REMD convencional, Cu-MD e MD), é possível a observação de certo padrão quanto às temperaturas e aos melhores GDTs/RMSDs obtidos nas simulações.

Para explorar melhor tais resultados a fim de entender melhor tal padrão, após a demultiplexação das trajetórias geradas por REMD, foi contabilizada a quantidade de estruturas em cada intervalo de GDT-TS, para cada temperatura (Figura 4.1). Por meio da figura, é possível notar a ínfima contribuição das estruturas obtidas em altas temperaturas (em termos de qualidade de estruturas).

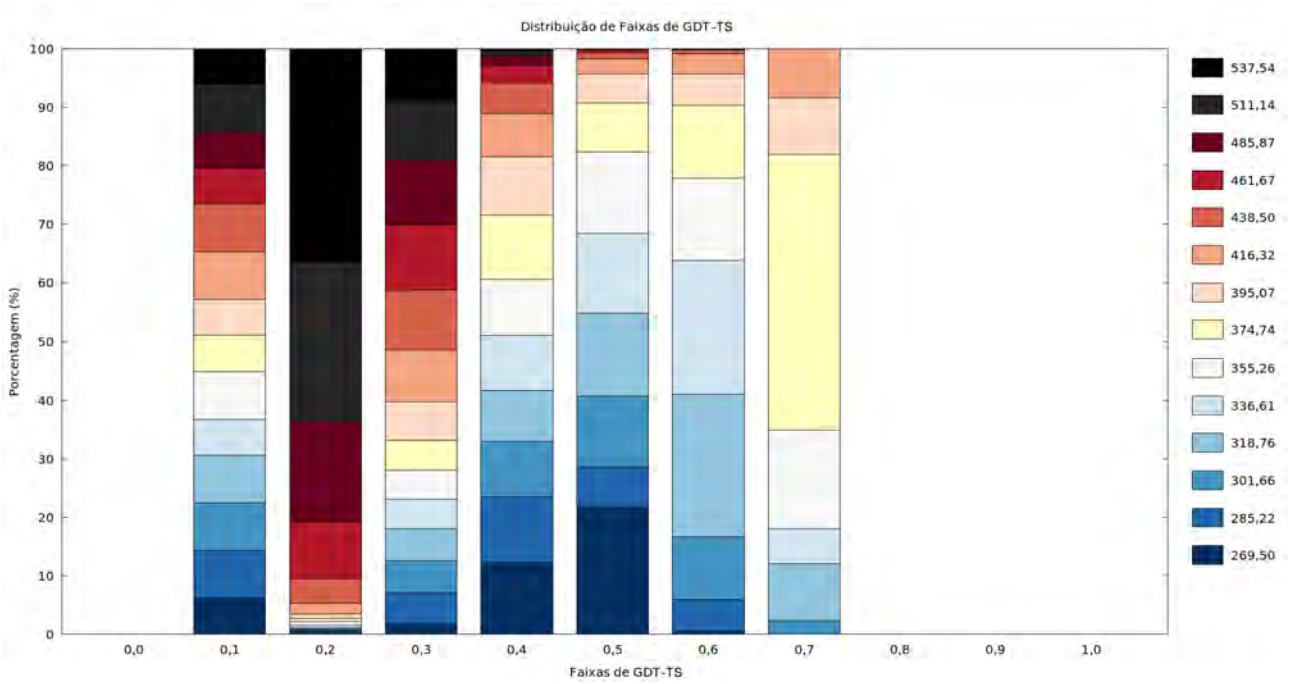

Figura 4.1 – Contabilização da quantidade de estruturas em cada intervalo de GDT-TS, para cada temperatura (em K). Resultados com base em trajetórias obtidas de simulações REMD convencional, para a proteína de código PDB 1UNC. Temperaturas muito altas pouco contribuem na obtenção de estruturas de GDT-TS mais alto.

Tal comportamento demonstra ser factível não levar em consideração, no momento da análise, estruturas provenientes de simulações a temperaturas mais elevadas. Com propósito de complementação, iniciaram-se esforços envolvendo medidas absolutas de avaliação da qualidade de proteínas, no intuito de utilizá-las como filtro adicional (além da quantidade de temperaturas) para se obter menor quantidade de estruturas a serem analisadas. De tal estudo e devido à complexidade envolvida neste ponto do trabalho, resultou

o trabalho de mestrado (ainda em desenvolvimento) realizado pelo aluno Rafael C. O. Macedo, vinculado ao Programa de Pós-Graduação em Ciência da Computação da Faculdade de Informática (FACIN) da PUCRS.

Isso posto, uma vez que tal trabalho encontra-se ainda em desenvolvimento, optou-se para esta tese por um protocolo simples de clusterização e obtenção de estruturas representativas comumente utilizado na literatura. Apenas trajetórias (demultiplexadas - que possuem apenas estruturas simuladas na mesma temperatura) referentes às quatro temperaturas mais baixas foram utilizadas como entrada para o protocolo de agrupamento. Os *clusters* foram calculados via cpptraj [PCC+95] usando o algoritmo de agrupamento hierárquico aglomerativo por ligação média [STTC07], com valor ε padrão de 2,0 [LLPDS11, DGJ+99]. Quando da clusterização, o algoritmo foi configurado para levar em consideração apenas resíduos de aminoácidos dentro das estruturas regulares presentes na estrutura RMN de referência. Trata-se de um procedimento comum que visa evitar que as voltas desordenadas e os resíduos terminais prejudiquem os *clusters* [PMD15].

Para avaliar o desempenho de CuT-REMD contra os métodos convencionais quanto à sua capacidade de prever estruturas nativas, foram calculadas as métricas *Best5Pop* e *BestStruc* [PMD15]. *Best5Pop* examina os cinco *clusters* mais populosos, computando o RMSD de seus centróides contra a estrutura experimental de referência e retorna a estrutura centróide com o menor RMSD. *BestStruc* retorna a estrutura prevista com o menor RMSD visitado em toda a simulação, incluindo trajetórias excluídas do protocolo de clusterização.

4.5 Recursos Utilizados

Pesquisas iniciais foram feitas com o objetivo de descobrir quais recursos seriam necessários para que o trabalho evoluísse da maneira pretendida. A última versão do AMBER não comporta alterações referentes à GB para processamento via placas gráficas, de forma que foi necessária, para simulações com solvente implícito, a utilização apenas de processamento paralelo, mais oneroso em termos de tempo computacional. Tendo em vista o custo computacional e a quantidade de experimentos elencados como parte desta tese, foi necessário estender as simulações à infraestrutura de *clusters* disponível no Laboratório de Alto Desempenho (LAD) da PUCRS, anexo ao prédio da FACIN. Outro recurso necessário foi o *software* proprietário AMBER, cujo custo para a comunidade acadêmica foi de cerca de U$ 500,00, em 2013. Todos os recursos necessários listados acima (com exceção da infraestrutura de *clusters*) foram disponibilizados pela estrutura dos laboratórios LABIO (Laboratório de Bioinformática, Modelagem e Simulação de Biossistemas) e FarmInf (Laboratório de FarmInformática).

5. RESULTADOS E DISCUSSÃO - PARTE 1: CUT-REMD

Neste capítulo, será apresentada a abordagem CuT-REMD. Primeiramente, será feita uma introdução às principais características da abordagem, seguida da explanação sobre sua implementação, a qual engloba a parametrização das simulações, alterações no código fonte do AMBER14 e dois entregáveis desta tese: (i) uma suíte de *scripts* e (ii) uma interface gráfica, em conjunto com a arquitetura geral da abordagem CuT-REMD (subseção 5.2.4).

Na sequência, serão apresentados os detalhes das simulações, seja para a proteína estudo de caso (*villin headpiece*) ou para o conjunto teste de proteínas.

A fim de facilitar o entendimento do leitor, todas as análises presentes nesta tese se encontram condensadas na seção 5.4.

5.1 Introduzindo CuT-REMD

A abordagem *Cutoff Temperature Replica Exchange Molecular Dynamics* ou CuT-REMD baseia-se no pressuposto de que a estrutura nativa de uma proteína é atingida por uma sequência de eventos que começa com o agrupamento de núcleos locais de EES dentro de segmentos distintos ao longo da cadeia polipeptídica. Desse modo, as distâncias na faixa responsável pela estabilização de pontes de hidrogênio (2,2 Å a 4,0 Å) [Jef97] aparecem como passíveis de relevância para iniciar a amostragem de conformações da cadeia polipeptídica, o que futuramente levará ao estado nativo, em raios de corte mais elevados. Para imitar esse efeito em uma simulação por DM, controlam-se os parâmetros usados pela função potencial para calcular trajetórias. Assim, para promover a formação de núcleos de EES locais, começam-se as simulações com raio de corte reduzido, e à medida que a simulação progride, esse raio de corte é gradualmente incrementado, aumentando assim as probabilidades da formação dos núcleos de EES para capturar o arranjo espacial de estruturas secundárias. Esse conceito é básico para este estudo e está exposto graficamente por meio da Figura 5.1.

5.2 Implementação

Uma gama de *scripts* (utilizando em sua maioria as linguagens *Python* e *Batch*) foram desenvolvidos para tornar possível a execução e análise da abordagem. Este capítulo está dividido em três seções. Inicialmente, serão apresentados os parâmetros base das simulações CuT-REMD. Será apresentada a descrição das alterações efetuadas no código

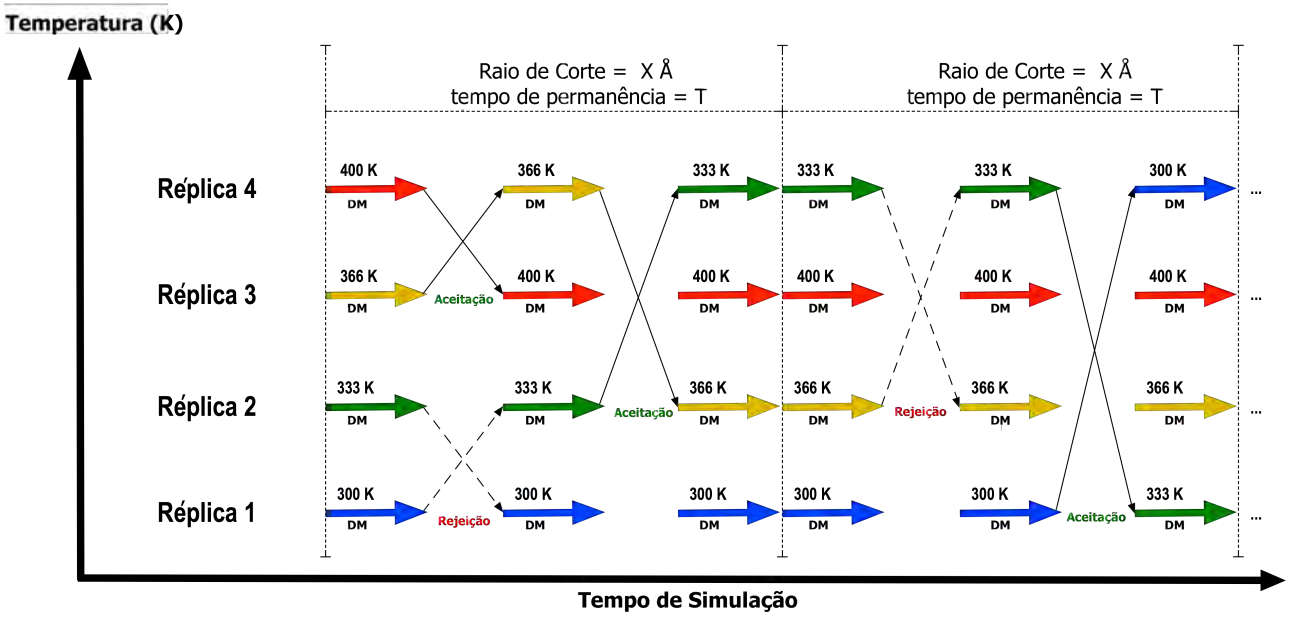

Figura 5.1 – Visão geral da abordagem CuT-REMD proposta. As réplicas são simuladas por DM independentes. À medida que a simulação prossegue, tentativas de intercâmbio entre estruturas imersas em diferentes banhos térmicos são executadas. Em cada tempo de permanência T, existe uma modificação do raio de corte. Neste exemplo, iniciando a partir do valor x Å e sendo gradualmente incrementado por y Å. Neste trabalho, x iniciou-se em 4,0 Å e y foi fixado em 1,0 Å arbitrariamente.

fonte do AMBER para tornar possíveis simulações com raios de corte mais curtos que o usual. Em seguida, a suíte de *scripts* criada para geração, execução e análise de simulações CuT-REMD é apresentada. Na sequência, tem-se a apresentação da ferramenta gráfica criada para auxílio na criação, execução e análise de simulações não apenas CuT-REMD, como também Cu-MD, REMD convencional e DM convencional. A arquitetura geral da solução é apresentada por fim.

Respeitadas as licenças relativas aos *softwares* proprietários utilizados por este trabalho, os quais em parte alguma estão reproduzidos ou sendo distribuídos por qualquer meio, todos os *softwares* gerados por esta pesquisa e apresentados nesta tese são *softwares* livres. É possível redistribuí-los e/ou modificá-los sob os termos da Licença Pública Geral GNU publicada pela Free Software Foundation, desde que estejam de acordo com a versão 2 ou superior da Licença.

Tais *softwares* serão distribuídos na esperança de que sejam úteis, porém sem nenhuma garantia. Inclusive sem a garantia implícita de adequação para determinados propósitos, uma vez que não foram exaustivamente testados (teste de *software*) com o intuito de terem robustez suficiente para funcionar 100% na heterogeneidade de plataformas disponíveis. Deve-se consultar a Licença Pública Geral GNU para obter mais detalhes (https://www.gnu.org/licenses/). Em adição, o autor se coloca à disposição para colaborações.

5.2.1 Parametrização CuT-REMD e Alterações no Código Fonte do AMBER14

A Figura 5.2 retrata um exemplo de entrada (arquivo .mdin) utilizado por simulações CuT-REMD. É importante lembrar que tal exemplo refere-se a apenas 1 ns de simulação. Mais informações quanto aos parâmetros utilizados podem ser encontrados no Apêndice B.

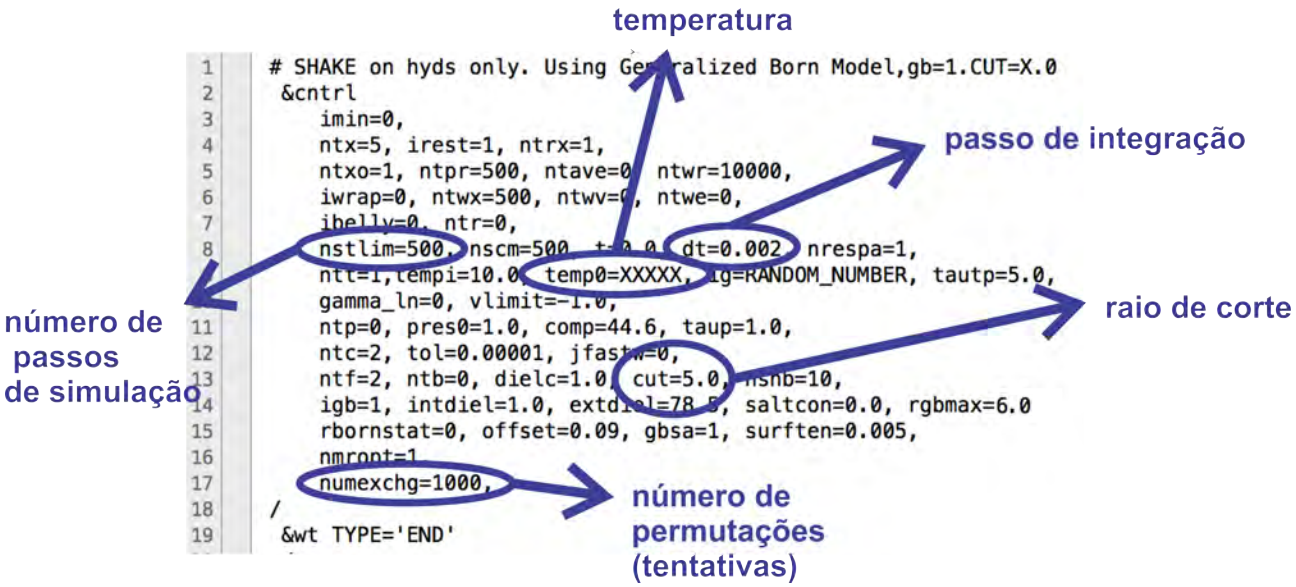

Figura 5.2 – Demonstração dos parâmetros de entrada de uma simulação CuT-REMD: neste exemplo, o parâmetro *cut* simboliza o raio de corte escolhido para a simulação, *nstlim* determina o número de passos de dinâmica entre cada tentativa de intercâmbio, e nesse caso são 500 passos, os quais levam dt = 0.002 ou 2 fento segundos para integração, o que reflete em 1 picosegundo de simulação entre cada intervalo de intercâmbio. O número de tentativas de intercâmbio é dado pelo parâmetro *numexchg*, e *ig* é a semente aleatória. O tempo total de simulação (ns) é dado pela multiplicação: *nstlim* x *dt* x *numexchg*.

A parametrização apresentada pela figura 5.2 só é possível de ser executada em ambientes que tenham o código fonte do AMBER14 modificado. Para fazê-lo, deve-se seguir o exposto:

Para tornar possível a utilização de raios de corte tão baixos, foi necessária a alteração do código fonte do AMBER. Para atingir tal etapa, foi editado o arquivo “src/pmemd/src/mdin_ctrl_dat.F90” conforme orientação abaixo:

- Alterar linha:
  if (gb_cutoff .lt. 8.05) then

  Para:
  if (gb_cutoff .lt. 0.00) then

- Alterar linha:
  if (cut < 999.0d0) then

Para:
if (cut < 0.0d0) then

- Alterar linhas:
  write(mdout, '(a)') '
  Require cut > 999.0d0.'

  Para:
  write(mdout, '(a)') '
  Require cut > 0.0d0.'

```
  if (es_cutoff .le. 0.d0 .or. vdw_cutoff .le. 0.d0) then
    write(mdout, '(a,a)') error_hdr, &
          'Cutoffs (cut, es_cutoff, vdw_cutoff) must be positive!'
    inerr = 1
  end if

else if (igb .eq. 1 .or. igb .eq. 2 .or. igb .eq. 5 .or. &
         igb .eq. 7 .or. igb .eq. 8) then

  if (gb_cutoff .lt. 0.00) then
    write(mdout, '(a,a)') error_hdr, &
      'Cut for Generalized Born simulation too small!'
    inerr = 1
  end if

  if (lj1264 .ne. 0) then
    write(mdout, '(2a)') error_hdr, &
      'Generalized Born is incompatible with the 12-6-4 potential'
    inerr = 1
  end if

end if
```

Figura 5.3 – Exemplo de alteração efetuada no código fonte do AMBER. Parte dos requisitos para que CuT-REMD possa ser aplicado. Arquivo mdin_ctrl_dat.F90.

Destaca-se ainda o fato de que, atualmente, as simulações CuT-REMD somente podem ser executadas em CPU, não sendo possível executá-las em GPU, uma vez que *procedures* essenciais à execução CuT-REMD possuem restrições quanto a alterações em seus códigos fonte. Qualquer alteração em código confere a necessidade de recompilação do programa, não apenas em sua versão sequencial, como também em sua versão paralela.

5.2.2 Suíte de *scripts* CuT-REMD

A suíte de *softwares* CuT-REMD é composta pelos seguintes *scripts* e está disponível *on-line* em https://github.com/paes/CuT-REMD:

- *analyze_all_temps.x*

  Ao passo que uma simulação REMD se desenvolve, quando da aceitação de intercâmbio entre duas réplicas, ao invés de haver a troca de conformações entre uma simulação sendo executada a uma temperatura A e uma simulação sendo executada a uma temperatura B, o código do AMBER troca somente um valor, a temperatura. Assim sendo, tem-se ganho computacional e, como consequência, geram-se trajetórias multiplexadas, ou seja, que possuem estruturas provenientes de simulações em banhos térmicos diferentes.

  As trajetórias de simulações REMD necessitam, portanto, passar por uma demultiplexação: processo para transformar as trajetórias em trajetórias compostas por estruturas geradas sob a mesma temperatura. É isso que Analyze_all_temps.x faz: transforma as trajetórias multiplexadas geradas em trajetórias demultiplexadas. Utiliza o módulo *cpptraj* do AMBER.

- *boxplot_gdt.sh*
  Script para geração de gráficos GDT, com todas as temperaturas e de toda trajetória.

- *boxplot_rmsd.sh*
  Script para geração de gráficos RMSD, com todas as temperaturas e de toda trajetória.

- *Calcula_Ranges_GDT.sh*
  Script para geração de gráficos GDT dividido em faixas de 0.1.

- *Calcula_Ranges_RMSD.sh*
  Script para geração de gráficos RMSD dividido em faixas 1.0 Angstroms.

- *clusterize.x*
  Script para automatização da clusterização. Uma vez que trabalha com muitos cálculos, se utiliza do módulo cpptraj compilado com Open MP.

- *computa_LN_distrib.sh*
  Computa a distribuição de energia potencial de cada par de temperaturas adjacentes. Utiliza como entrada os arquivos gerados pelo *script* energydistribution.py.

- *compute_folded.py*
  Script para contabilização de estruturas tidas como enoveladas.

- *compute_energies.x*
  Script responsável pelo cálculo de energias das simulações e geração de gráficos. Quando executado em simulações REMD, gera também gráficos relativos à probabilidade de distribuição canônica gerada pelas simulações.

- *compute_gdt.py*
  Avalia arquivos .gdt.dat (réplicas) e retorna um arquivo com: *Temperature(T) LowestValue(LV) LowerQuartile(Q1) MiddleQuartile(Q2) HigherQuartile(Q3)* e *HighestValue(HV)*.

- *compute_rmsd.py*
  Avalia arquivos .rmsd.dat (réplicas) e retorna um arquivo com: *Temperature(T) LowestValue(LV) LowerQuartile(Q1) MiddleQuartile(Q2) HigherQuartile(Q3)* e *HighestValue(HV)*.

- *convergence_plot_1UNC_avg.sh*
  Geração de gráficos da média de convergência entre três execuções.

- *convergence_plot_1UNC.sh*
  Geração de gráficos de convergência. Utiliza saídas do *software* ENCORE.

- *correlation.py*
  Script criado para avaliar o coeficiente de correlação entre curvas, mais especificamente entre coeficiente angular ou *slope* teórico de uma distribuição de Boltzmann e o *slope* retornado pelas simulações executadas na tese.

- *create_plots_relative_metrics.x*
  Cria gráficos de RMSD/GDT-TS comparando CuT-REMD, REMD, Cu-MD e DM.

- *dssp_go_1unc.py*
  Calcula a adequação de resíduos quanto à estrutura secundária, com base em estrutura pdb de referência. Utiliza o *cpptraj* e *software* DSSP.

- *dssp_go.x*
  Utiliza-se do *software* DSSP para computar, para a trajetória completa, o DSSP de cada *snapshot*.

- *dssp_plot.sh*
  Gera gráficos referentes a DSSP.

- *EF.py*
  Script gerado para contabilizar a quantidade de Eventos de Tunelamento ou *Tunneling Events* da simulação.

- *folded_plot.sh*
  Gera gráficos referentes a estruturas tidas como enoveladas ou *folded*.

- *gdt_from_traj.x*
  Script criado para calcular GDT-TS entre as trajetórias geradas e a estrutura de referência. Utiliza o *software* ClusCo [JK13].

- *gdt_min_avg.sh*
  Gera gráficos da média de GDT-TS mínimos obtida em três execuções diferentes.

- *gdt_min.sh*
  Script para computar GDT-TS mínimos atingidos pelas simulações.

- *gdt_ranges.py*
  Script utilizado para gerar matrizes relativas aos intervalos de GDT-TS.

- *generate_ncdf_total.x*
  Script para transformar trajetórias do AMBER em trajetórias binárias (ncdf).

- *generate_pdb_total.x*
  Script criado para unificar as trajetórias de mesma temperatura executadas com raios de corte diferentes, respeitando a sequência temporal em que foram geradas. Utiliza o módulo *cpptraj* do AMBER.

- *get_from_tra.x*
  Script para capturar modelo de número específico dentre um arquivo pdb multimodelos. Usualmente utilizada para, de posse do arquivo pdb da estrutura de referência, capturar o primeiro modelo.

- *graph_gdt_ranges_new.x*
  Gera gráficos de intervalos de GDT-TS.

- *map_allinone_Encore.sh*
  Gera mapas 6x6 provenientes do ENCORE.

- *radgyr-fromTtraj_md.x*
  Contabiliza o raio de giro de simulações por DM.

- *radgyr_fromTtraj.x*
  Contabiliza o raio de giro de simulações REMD.

- *Quality_Sample.x*
  Script responsável por cálculos de métricas absolutas (DOPE, G-Factor, ProbScore, DFIRE, dDFire, RWPlus, OPUS-PSP e GOAP).

- *remove_heatoms.py*
  Script utilizado para remover átomos de hidrogênio de arquivos pdb.

- *rmsd_from_traj.x*
  Script criado para calcular rmsd entre as trajetórias geradas e a estrutura de referência. Utiliza o módulo *cpptraj* do AMBER.
- *run_EF.sh*
  Executa cálculo de EF baseado em arquivos .log.
- *top5.sh*
  Script para calcular *BestClus* e *BestStru*.
- *verify_distribution.sh*
  Script para computar distribuição de energia.
- *verify_errors.sh*
  Script para contabilizar o erro entre coeficientes angulares teóricos e os obtidos pelas simulações.

5.2.3 A Interface Gráfica GTK-REMD

Uma vez que as simulações CuT-REMD possuem características específicas de configuração e fluxo de dados, não existem, na literatura, interfaces gráficas bem adaptadas para prover devido apoio àquele que considere a utilização da abordagem. Por esse motivo, optou-se pela criação de uma interface gráfica própria: GTK-REMD. Tal interface possibilita a configuração de simulações não apenas CuT-REMD, como também REMD convencional, Cu-MD e DM convencional, o que a torna uma ferramenta de uso geral, porém limitada a simulações feitas no AMBER [CCID+05]. A plataforma foi escrita em linguagem *Python* e pode ser utilizada em qualquer sistema, desde que este apresente instalados os seguintes pacotes/*softwares*:

1. GTK2.0;
2. Python com Numpy;
3. AmberTools [CBB+14];
4. ClusCo [JK13];
5. Procheck [LMMT93];
6. Molprobity [CAH+10]; e
7. Modeller [SB93].

Onde 1. e 2. são utilizados diretamente pelos módulos internos de GTK-REMD e os demais (3. a 7.), para cálculos envolvendo métricas absolutas e relativas. A interface gráfica GTK-REMD é divida em duas abas:

1. Configuração de Simulações e
2. Análise de Configurações.

As Figuras 5.4 e 5.5 apresentam essas duas abas. GTK-REMD foi escrito na linguagem Python e utiliza-se do conjunto de ferramentas GTK+, o qual provê interface amigável.

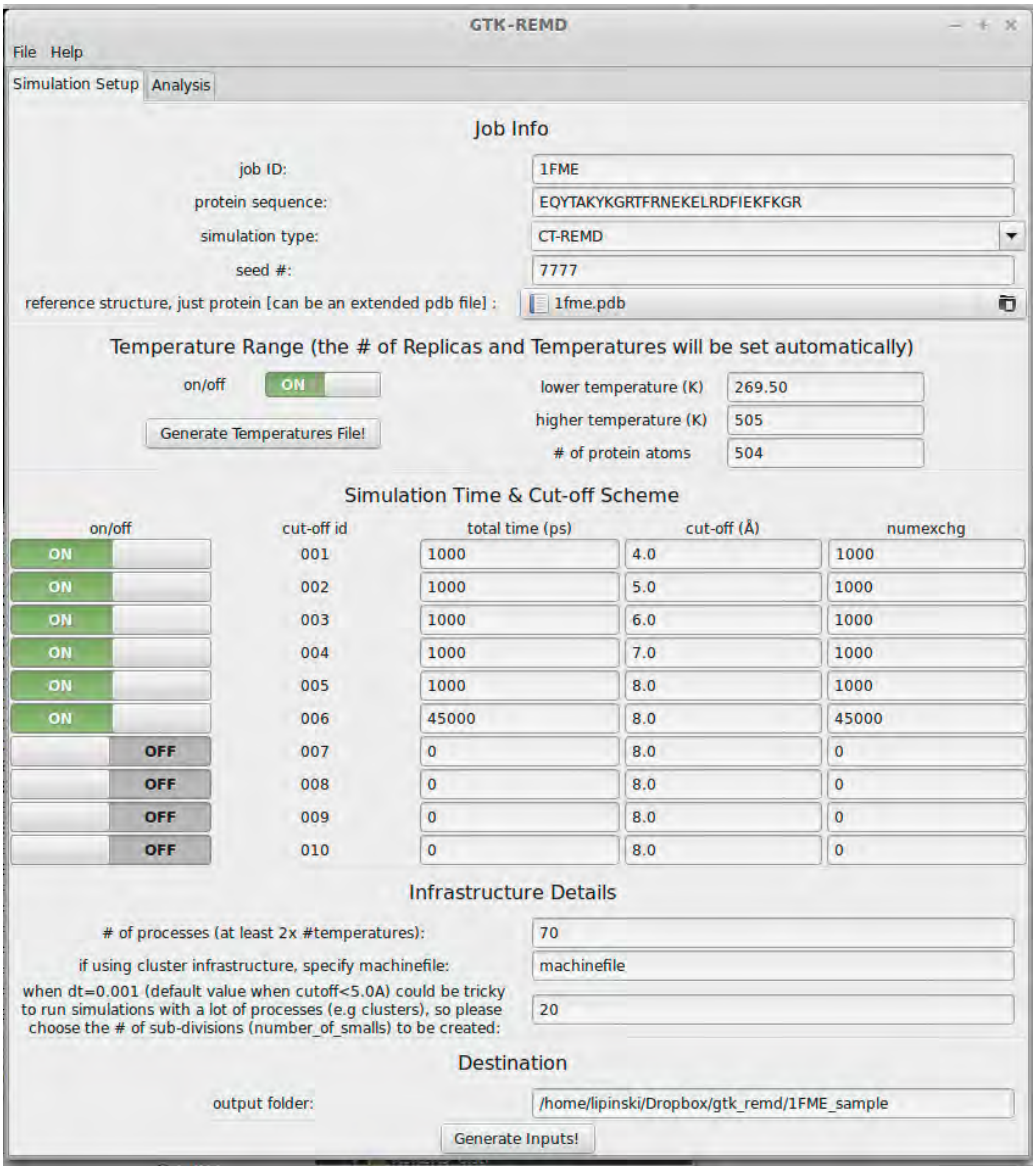

Figura 5.4 – GTK-REMD: Aba de configuração de simulações

Aba de Configuração de Simulações

A aba Configuração de Simulações ou *"Simulation Setup"* (Figura 5.4) foi criada no intuito de facilitar a execução de simulações, uma vez que possibilita a configuração e geração de arquivos de entrada destas. Por meio dela, é possível configurar não apenas simulações do tipo CuT-REMD como também simulações convencionais REMD, além de

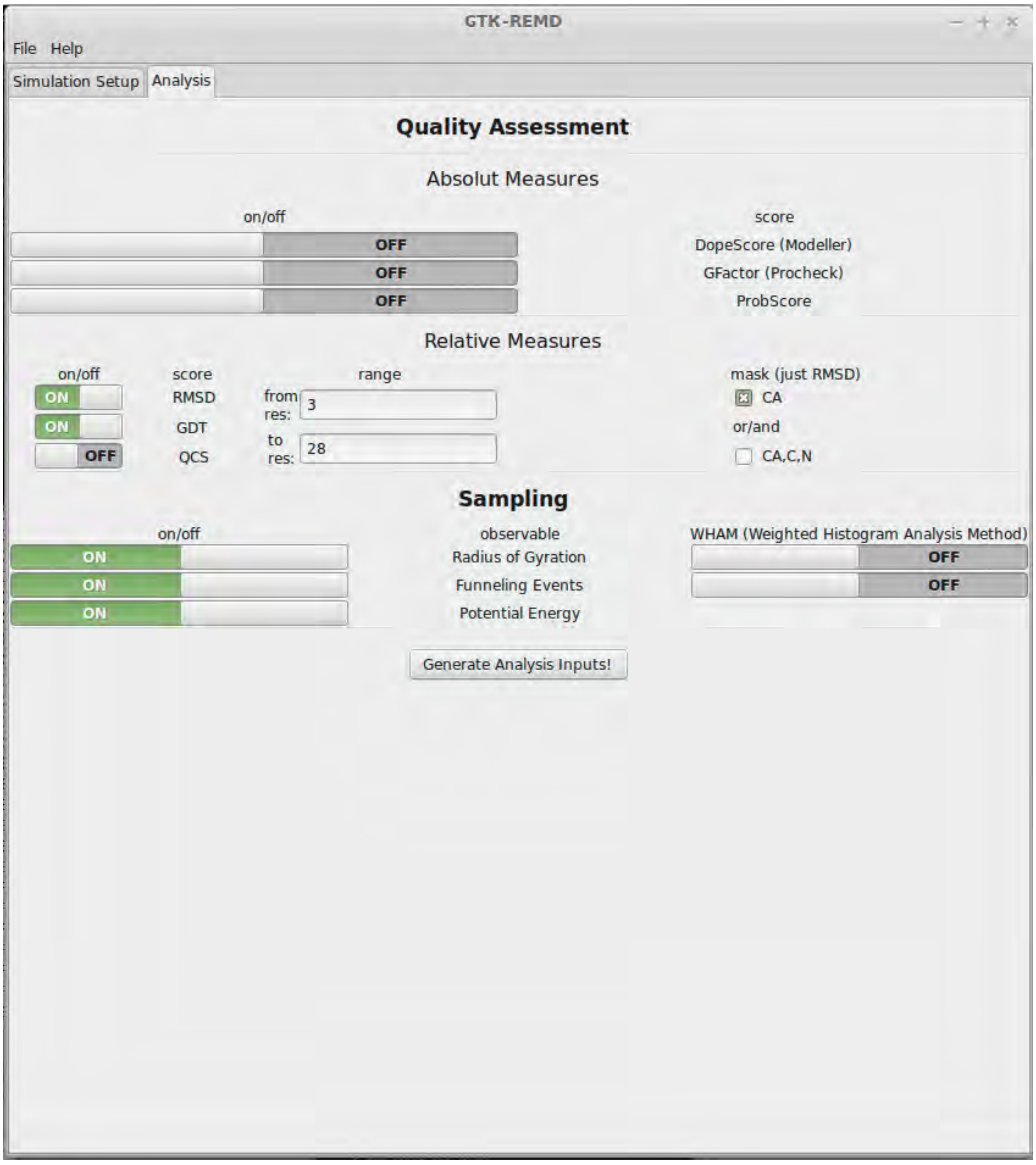

Figura 5.5 – GTK-REMD: Aba de análises

simulações Cu-MD e simulações por DM convencional. As entradas requisitadas por GTK-REMD para gerar os arquivos base das simulações são:

1. identificador;
2. sequência de aminoácidos da proteína;
3. tipo de simulação (CuT-REMD, REMD, Cu-MD ou DM);
4. semente aleatória;
5. estrutura de referência;
6. intervalo de temperaturas; e
7. quantidade de átomos na proteína.

O sistema é bem adaptado para automatizar também a estipulação da quantidade e valores de temperaturas a serem utilizadas, baseando-se na quantidade de átomos do sistema, e em extremos de temperatura estipulados pelo usuário. Para isso, a ferramenta GTK-REMD se vale de integração *on-line* com o *webserver* de Patriksson e van der Spoel [PvdS08]. No caso de simulações Cu-MD ou por DM convencional, são preparados X sistemas independentes, onde X é o número de temperaturas de interesse.

Fica a cargo do usuário estipular, para cada simulação, o tempo de simulação, quantidade de tentativas de intercâmbio e raios de corte. Sendo facilitada a criação de simulações que sejam formadas por pequenas sequências de execução (caso de CuT-REMD). GTK-REMD está, na presente versão, também adaptado para simulações envolvendo infraestrutura de *cluster* e execuções MPI.

As principais saídas geradas por GTK-REMD são:

- Arquivos .mdin
  Arquivos contendo os parâmetros da simulação. Em geral, em simulações CuT-REMD, existe uma sequência de arquivos .mdin, os quais são, posteriormente, executados de forma sequencial.

- Arquivo *temperatures.dat*
  Arquivo contendo uma lista de temperaturas nas quais as simulações ocorrerão.

- *fila.sh*
  Script para dar início à execução da simulação.

Aba de Análises

A segunda aba presente na interface gráfica GTK-REMD é a de análises ou *"Analysis"* (Figura 5.4). Nela, é possível preparar uma gama de análises a serem feitas em momento posterior ao da execução das dinâmicas.

Tais análises estão divididas em três grupos:

1. Análises de métricas absolutas:

   - DopeScore (Modeller);
   - GFactor (Procheck); e
   - ProbScore.

2. Análises de métricas relativas:

   - GDT_TS;
   - RMSD; e

- QCS (em desenvolvimento).

3. Análises quanto à amostragem:

   - Raio de Giro;
   - Eventos de Tunelamento; e
   - Energia Potencial.

5.2.4 Arquitetura Geral CuT-REMD

Para finalizar esta seção, apresenta-se a arquitetura geral da abordagem (Figura 5.6). Tal arquitetura baseia-se em três etapas A, B e C, onde A e B simbolizam a preparação da simulação e de suas análises básicas e C simboliza a execução da simulação. Destaca-se ainda que, caso seja de interesse do usuário, este poderá utilizar-se da suíte de *scripts* CuT-REMD, a qual provê suporte para análises mais detalhadas que as disponíveis em B. A suíte CuT-REMD está disponível *on-line* em http://www.github.com/paes/cut-remd.

5.3 Detalhes das Simulações

5.3.1 Proteína *villin headpiece* de Código PDB 1UNC

Com o intuito de validar CuT-REMD, foi realizado um estudo de caso com a estrutura *villin headpiece* de humanos, de código PDB: 1UNC [VVVT$^+$04]. Essa é uma pequena proteína contendo 35 resíduos de aminoácidos (estrutura primária: LSIEDFTQAFGMTPAAFSALPRWKQQNLKKEKGLF), nos quais vários EES (três hélices) são ligados entre si por um núcleo hidrofóbico bem empacotado (composto por três resíduos de fenilalanina e outros resíduos hidrofóbicos). A *villin headpiece* é uma das menores proteínas nativas em que se encontram características de proteínas muito maiores, caracterizando assim seu estudo como valioso para realçar o conhecimento sobre predição de estruturas 3D e enovelamento de proteínas [VVVT$^+$04].

Nove diferentes protocolos foram testados, incluindo duas simulações com REMD convencional e uma com DM convencional, para comparações. Os experimentos foram realizados no *cluster* Cerrado, disponibilizado pelo Laboratório de Alto Desempenho (LAD) da PUCRS, totalizando mais de 1.600 horas de tempo de CPU. Os protocolos de simulação estão resumidos na Tabela 5.1. Cada ID representa diferentes protocolos de simulação executados em triplicata, variando pelo valor semente ou *seed number*. Para simulações com diferentes IDs, no entanto, o valor semente permaneceu fixo. Destaca-se ainda que, mesmo para os métodos convencionais, os protocolos utilizados, embora gerais, foram gerados especificamente para este trabalho, o que engloba a estipulação de todos os parâmetros de

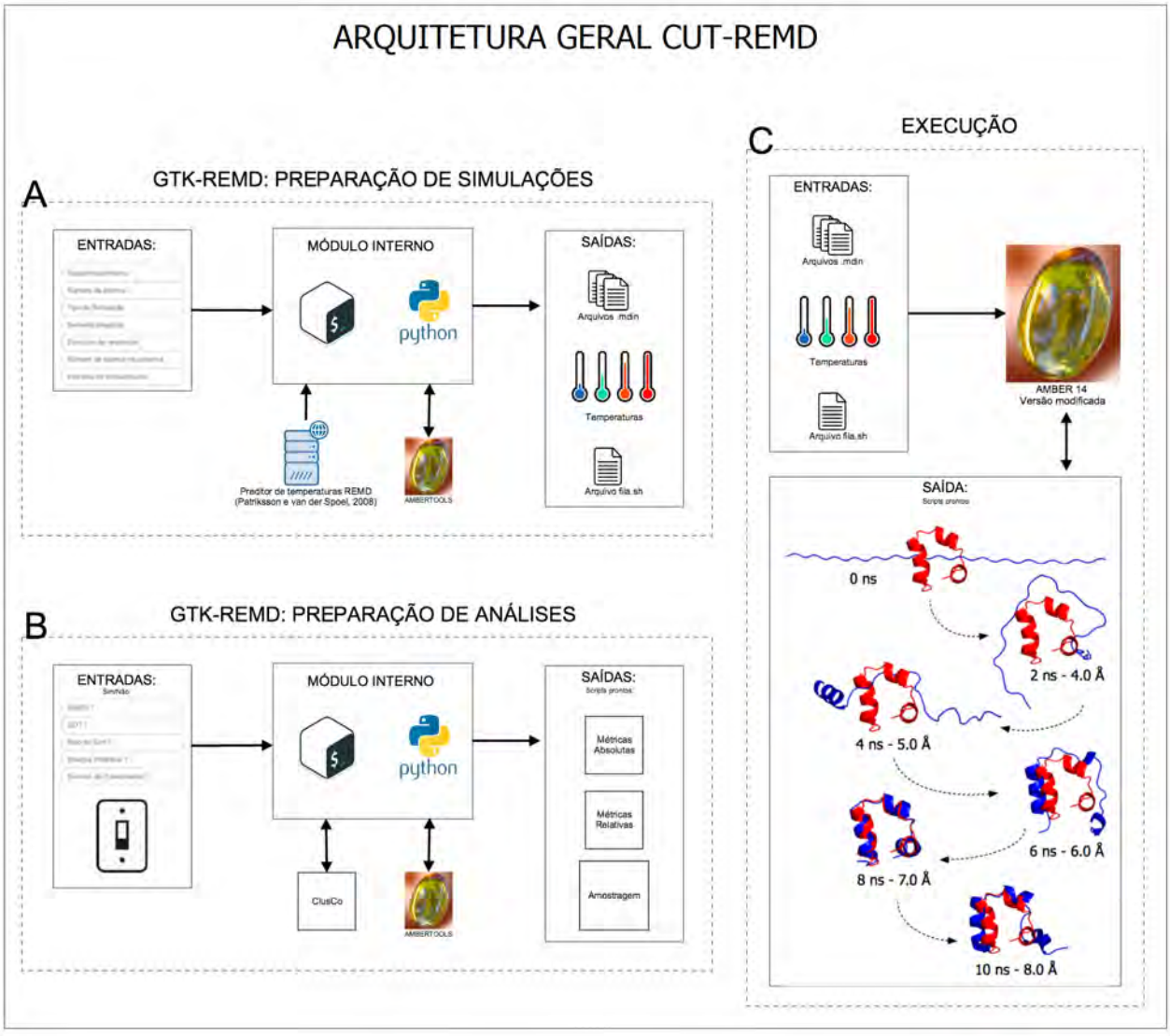

Figura 5.6 – Arquitetura geral da abordagem CuT-REMD. A) representa a etapa de preparação da simulação, B) representa a preparação de suas análises básicas e C) representa a etapa de execução da simulação. A suíte de *scripts* CuT-REMD não está representada e deve servir de suporte para análises mais detalhadas.

entrada de tais simulações, não havendo reutilização de protocolos da literatura. Assim sendo, os resultados retornados por via das simulações aplicando protocolos utilizando métodos convencionais, ainda que não sejam o foco do trabalho, configuram-se também como entregáveis desta tese.

Todos os sistemas iniciaram a partir de uma estrutura totalmente estendida de 1UNC, a qual foi construída com o módulo tleap do AMBER14 [CBB+14]. Após a etapa de minimização de energia, cada um dos nove protocolos foi executado em triplicata durante 50 ns (5.1, Etapas 1 a 6), em *ensemble* NT. As simulações se utilizaram da versão interna modificada do módulo PMEMD contido no AMBER14 [CBB+14]. Os protocolos A, B e G empregaram 1 ns de tempo de permanência em raios de corte mais curtos, enquanto B, D e H empregaram 2 ns. Protocolos com REMD convencional (E e F) e DM convencional (I)

Tabela 5.1 – Sumário dos protocolos de simulação. Para simulações do tipo REMD (A, B, C, D, E e F), EAF na Etapa 6 foi setado em 1 ps^{-1}(*), 0,025 ps^{-1}(†) e 0,020 ps^{-1}(§).

ID	Abordagem	Raio de Corte (Å) e tempo de permanência (ns)											
		Etapa 1		Etapa 2		Etapa 3		Etapa 4		Etapa 5		Etapa 6	
A	CuT-REMD	4,0	1	5,0	1	6,0	1	7,0	1	8,0	1	8,0	45*
B	CuT-REMD	4,0	1	5,0	1	6,0	1	7,0	1	8,0	1	8,0	45§
C	CuT-REMD	4,0	2	5,0	2	6,0	2	7,0	2	8,0	2	8,0	40*
D	CuT-REMD	4,0	2	5,0	2	6,0	2	7,0	2	8,0	2	8,0	40†
E	T-REMD	8,0	1	8,0	1	8,0	1	8,0	1	8,0	1	8,0	45*
F	T-REMD	8,0	1	8,0	1	8,0	1	8,0	1	8,0	1	8,0	45§
G	Cu-MD	4,0	1	5,0	1	6,0	1	7,0	1	8,0	1	8,0	45
H	Cu-MD	4,0	2	5,0	2	6,0	2	7,0	2	8,0	2	8,0	40
I	MD	8,0	1	8,0	1	8,0	1	8,0	1	8,0	1	8,0	45

também foram simulados, com raio de corte fixo em 8,0 Å. Na Etapa 6, os protocolos A, C e D utilizaram EAF de 1 ps^{-1}, B e F de 0,020 ps^{-1} e D de 0,025 ps^{-1}.

Todas as simulações de DM foram realizadas utilizando o campo de força ff14SB presente no AMBER, e o modelo generalized Born pareado de Hawkins, Cramer e Truhlar [HCT95, HCT96].

Constrições foram calculadas através do programa makeCHIR_RST e adicionadas às simulações para evitar rotações indesejáveis que pudessem levar a quiralidades não físicas em altas temperaturas. O algoritmo SHAKE [RCB77] foi aplicado para restringir ligações envolvendo átomos de hidrogênio. Realizaram-se simulações utilizando 14 temperaturas diferentes, variando de 269,50 K a 537,54 K. Esse número de réplicas/temperaturas foi escolhido com base no número de átomos e graus de liberdade do sistema e foi calculado através de um servidor de predição de temperaturas criado por Patriksson e van der Spoel [PvdS08].

Na Tabela 5.1, o raio de corte está relacionado com parâmetro *cut* utilizado para truncar pares não ligados (em uma base átomo a átomo) quando computando a energia potencial. O parâmetro *cut* estabelece a distância máxima para os termos eletrostáticos, de van der Waals e "fora da diagonal" da interação generalized Born [CBB+14]. A distância máxima entre os pares de átomos considerada quando da soma par a par ao calcular os raios efetivos de Born (rgbmax) e o tempo de integração foram ajustados, respectivamente, para 6,0 Å e 1 fs em raios de corte < 6,0 Å e 10,0 Å e 2 fs nos demais.

A partir de 4,0 Å, o raio de corte foi gradualmente aumentado em 1,0 Å (valor escolhido arbitrariamente) a cada 1 ou 2 ns (dependendo do protocolo) até atingir 8,0 Å, onde permaneceu até o final das simulações. Conforme discutido na literatura [SMR08, SER10, RC03, ZWD05, RP03], o EAF influencia a amostragem. Portanto, na Etapa 6 de todas as simulações, diferentes EAFs foram testados, a popular e sugerida frequência moderada de

1 ps^{-1} [PM07] e duas frequências relativamente mais baixas de 0,025 ps^{-1} e 0,020 ps^{-1}), para verificar o comportamento do sistema. Um total de 700.000 (50.000 para cada temperatura/réplica) *snapshots* foram utilizados para análise de cada uma das simulações.

5.3.2 Conjunto Teste de Proteínas

O restante das simulações foi efetuado com o intuito de testar CuT-REMD com proteínas de diferentes classes, respeitando, no entanto, um limite de 40 resíduos para o tamanho das cadeias polipeptídicas. Com base nos trabalhos encontrados na literatura acerca do foco desta tese (oriundos do mapeamento sistemático presente no Apêndice A), as seguintes proteínas foram escolhidas:

Tabela 5.2 – Conjunto teste de proteínas

Código PDB	SEQ	# Res	Classe	Referência
1L2Y	NLYIQWLKDGGPSSGRPPPS	20	α	[NFA02]
1RIJ	ALQELLGQWLKDGGPSSGRPPPS	23	α	[LLA+04]
1VII	MLSDEDFKAVFGMTRSAFANLPLWKQQ NLKKEKGLF	36	α	[MMK97]
1UAO	GYDPETGTWG	10	β	[HYSM04]
1LE1	SWTWENGKWTWKX	13	β	[CSS01]
1E0L	SEWTEYKTADGKTYYYNNRTLESTWE	26	β	[MGCO00]
1FME	EQYTAKYKGRTFRNEKELRDFIEKFKGR	28	$\alpha\beta$	[SM01]
1PSV	KPYTARIKGRTFSNEKELRDFLETFTGR	28	$\alpha\beta$	[DSM97]
2WXC	GSQNNDALSPAIRRLLAEWNLDASAIKGT GVGGRLTREDVEKHLAKA	30	$\alpha\beta$	[NSR+09]

Uma vez analisando-se os resultados obtidos em 5.3.1, o protocolo D foi escolhido para representar CuT-REMD. Para fins de comparação, executou-se também, para o conjunto teste de proteínas da Tabela 2, simulações por REMD convencional, seguindo o protocolo de simulação F (Tabela 5.1).

5.4 Análises

5.4.1 Verificação acerca da Distribuição de Boltzmann

Como uma nova abordagem, é importante analisar as simulações para avaliar se as saídas geradas seguem uma distribuição de Boltzmann, o que concede ergodicidade

(capacidade do sistema de atingir qualquer região do espaço de configuração iniciando de qualquer outra região). Em um sistema que satisfaz DBC, é imposto que cada permutação elementar precisa equilibrar-se com um procedimento inverso correspondente. Portanto, em teoria, CuT-REMD mantém a probabilidade de intercâmbio expressa pela Equação 2.6, preservando assim DBC. Consequentemente, espera-se que os *ensembles* gerados por CuT-REMD sigam uma distribuição de Boltzmann. Para verificar isso, calculou-se a distribuição probabilística da energia potencial do sistema considerando pares de temperaturas adjacentes, distribuições as quais se sobrepõem parcialmente. A relação deve obedecer à Equação 5.1 [GHP06, RP03]:

$$I_n \left[\frac{P(E,\tau_2)}{P(E,\tau_1)} = \left(\frac{1}{\tau_1 \rho c_\delta} \right) - \left(\frac{1}{\tau_2 \rho c_\delta} \right) E + constant \right] \tag{5.1}$$

O raio de corte afeta fortemente a energia do sistema. Como estão sendo utilizados pontos de corte mais curtos que o padrão, é possível que o equilíbrio térmico seja inibido, corrompendo assim a amostragem conformacional. Como o método T-REMD convencional é amplamente testado e usado, foram realizadas comparações entre suas distribuições de energia potencial e as de CuT-REMD para aferir sobre o comportamento da abordagem proposta nesta tese. Variações no raio de corte, uma vez afetando o cálculo de energia potencial, também afetam a faixa de energias acessíveis ao sistema. Portanto, seria impróprio, em vias de comparação, agregar distribuições de energias potenciais originárias de simulações com raio de corte diferente. Consequentemente, as distribuições de probabilidade foram computadas coletando apenas energias originárias de simulações com o mesmo raio de corte. Aplicou-se Equação 5.1 usando intervalos de 1 kcal/mol para calcular P(E) de simulações CuT-REMD e REMD convencional para testar o equilíbrio canônico.

Para fornecer uma medida quantitativa da diferença entre as simulações contra o *slope* teórico esperado de uma distribuição de Boltzmann, primeiro ajustou-se uma reta aos pontos definidos por pares de temperaturas adjacentes e calculou-se seu coeficiente angular. Foram calculados os coeficientes de correlação médios entre os *slopes* obtidos e o esperado da Equação 5.1.

5.4.2 Sobreposição de Energia Potencial

Assumindo duas simulações REMD distintas executadas na mesma faixa de temperaturas, é possível quantificar a similaridade entre elas calculando a sobreposição (*overlap*) entre cada par de distribuições de energia à mesma temperatura, permitindo assim verificar se duas entradas de simulação diferentes (*inputs* - por exemplo, diferentes protocolos de simulação) conduzem a diferentes amostras da superfície de energia ou não. Além disso, a mesma medida pode ser utilizada para avaliar a reprodutibilidade da abordagem (mesmos parâmetros de entrada porém execuções diferentes). Tal sobreposição de ener-

gia ou *Energy Overlap* é calculada pelo coeficiente de Bhattacharyya [Bha43] apresentado abaixo (Equação 5.2). Um valor de *overlap* de 1 significa duas distribuições idênticas e um valor de *overlap* de 0 significa distribuições completamente distintas.

$$Overlap = \int_{-\infty}^{+\infty} \sqrt{P(E)}_{sim1} \sqrt{P(E)}_{sim2} dE \tag{5.2}$$

Neste trabalho, primeiro calculou-se o *overlap* entre simulações que variaram apenas pela semente aleatória, a fim de verificar a reprodutibilidade da abordagem **(Seção 3.2)**. Em seguida, calculou-se o *overlap* entre diferentes protocolos de simulações para avaliar a diversidade do espaço de energia acessado (Seção 3.3).

5.4.3 Taxas de Aceitação de Monte Carlo

A taxa de aceitação entre intercâmbios ou *Exchange Acceptance Ratio* (EAR) é expressa como a proporção entre o número de movimentos aceitos e o número total de tentativas de intercâmbio. Essa proporção pode ser utilizada para ajustar a faixa de temperatura. Essa relação é geralmente calculada entre réplicas vizinhas, entre as quais são permitidas as trocas e, enquanto alguns trabalhos ignoram essa restrição visando acelerar a amostragem [Cal05, BSVI07, CS11], outros otimizam EAR em tempo real [NH07].

Para simulações REMD, uma prática comum é escolher a faixa de temperatura de modo que a temperatura mais baixa seja inferior à temperatura de interesse e a mais alta esteja acima da temperatura de enovelamento, seguindo uma distribuição exponencial de temperaturas [TTH06]. Para simulações aplicadas ao problema PSP, no entanto, essa informação nem sempre está disponível e o EAR pode desempenhar um papel importante na determinação da necessidade de mais réplicas/temperaturas para atingir uma amostragem adequada.

Outro parâmetro importante a ser analisado é a taxa de aprisionamento entre intercâmbios ou *Exchange Trapping Ratio* (ETR), conceito introduzido por Sindhikara e colaboradores [SMR08]. ETR quantifica a fração de trocas que ocorrem quando a nova temperatura é a mesma que foi duas trocas antes ($T_n = T_{n-2}$). Se uma réplica oscila continuamente entre duas temperaturas vizinhas, o sistema torna-se localmente preso e não abrangerá o espaço de fase adequadamente, necessitando de mais tempo de simulação para atravessar barreiras de energia. Neste estudo, calculou-se ETR para todas as temperaturas individualmente a fim de entender contribuições específicas.

5.4.4 Eventos de Tunelamento

Eventos de tunelamento ou *Tunneling Events* (TEs) é o número de vezes que a simulação vai desde a temperatura mais baixa até a mais alta e de volta para a mais

baixa. O número de TEs em um sistema denota sua velocidade de difusão no espaço de temperatura e é consequentemente um indicativo da eficiência de amostragem configuracional [BN92, MSO03, AG08]. O "Tempo de TE" é o tempo médio que uma réplica leva para mover-se de uma temperatura mais baixa T_1 até a temperatura mais alta T_N e voltar. Para um tempo de simulação fixo, menores quantidades de Tempo de TE significam melhor amostragem. Calculou-se o Tempo de TE médio para todos os protocolos de simulação.

5.4.5 Verificação de Convergência

Avaliar a convergência é um passo fundamental nas análises de simulações por MD, especialmente quando se espera uma amostragem adequada que siga uma distribuição de Boltzmann. Uma abordagem viável seria monitorar o grau de convergência das diferentes conformações visitadas [Mob12]. Uma vez que CuT-REMD tem por objetivo prever a estrutura 3D de proteínas rapidamente (ou seja, por simulações não superiores a 50 ns), as análises aqui descritas levaram em conta uma estrutura fixa de tempo de simulação. O pacote python ENCORE [TPB+15] foi utilizado para quantificar a diferença entre a trajetória completa e uma janela temporal de tamanho incremental para calcular a rapidez com que os diferentes protocolos testados convergem. Calculou-se a similaridade entre *ensembles* aplicando o método de redução dimensional de *ensembles* ou *Dimensional Reduction Ensemble Similarity* (DRES), o qual utiliza uma matriz de distâncias par a par de RMSD como entrada para projetar o *ensemble* conformacional de alta dimensionalidade em um espaço de baixa dimensão. Cada *ensemble* tem sua distribuição de probabilidade calculada, seguida pelo cálculo da divergência de Jensen-Shannon entre os *ensembles*. A divergência de Jensen-Shannon utilizada em DRES pode assumir valores entre zero e ln(2)$\sim$0,69 e, quanto menor for seu valor, menor é a contribuição entrópica (ganho de informação). Para mais detalhes sobre DRES, veja [LLFB09].

5.4.6 Formação de EES e Estruturas Terciárias Enoveladas

Os EES foram determinados utilizando o programa DSSP99. Para 1UNC, foram consideradas as suas três hélices α de tamanhos 6 (H1: Ile3 a Gln8), 6 (H2: Pro14 a Ala 19) e 10 (H3: Arg22 a Glu31), totalizando 22 resíduos em estruturas secundárias regulares. Um EES foi atribuído como correto se, para cada *snapshot* entre as simulações em triplicata de cada protocolo (Tabela 5.1), pelo menos 80% de seus resíduos estivessem em concordância com os EES na estrutura de referência (primeiro modelo na estrutura de RMN com código PDB 1UNC). As estruturas simuladas foram consideradas como "enoveladas" ou *folded* quando o RMSD entre C_αs, para os resíduos Ile3 a Gly33, estava dentro de 3,5 Å a partir da estrutura RMN de referência e continha a atribuição correta de EES.

5.4.7 Avaliação da Qualidade de Modelos

Para avaliar a habilidade da abordagem proposta por este estudo em amostrar conformações próximas ao estado nativo, utilizou-se RMSD e GDT-TS como medidas de similaridade estrutural entre estruturas preditas e as estrutura determinadas experimentalmente. Essas análises utilizaram uma referência comum (o primeiro modelo na estrutura de RMN experimental).

Para cálculos de GDT-TS, todos os resíduos foram considerados, para todas as proteínas testadas. Quanto a RMSD, a Tabela 5.3 a seguir apresenta os resíduos (intervalos) utilizados para os cálculos de RMSD deste trabalho, tanto para a proteína 1UNC (estudo de caso) quanto para as proteínas parte do conjunto de teste.

Tabela 5.3 – Intervalos de resíduos considerados para o cálculo de RMSD, para todas as proteínas testadas

Código PDB	Intervalo
1L2Y	3-18
1RIJ	2-22
1VII	3-32
1UNC	3-33
1UAO	1-10
1LE1	1-12
1E0L	1-26
1FME	2-28
1PSV	2-27
2WXC	10-28,36-47

5.4.8 Resíduos Considerados na Clusterização de Estruturas

Conforme descrito anteriormente (seção 4.4), na etapa referente à clusterização de estruturas, o cálculo de RMSD todos contra todos é feito considerando-se apenas os resíduos que fazem parte de estruturas secundárias na estrutura experimental.

A Tabela 5.4 demonstra os intervalos entre resíduos utilizado para os cálculos deste trabalho.

Tabela 5.4 – Resíduos de aminoácidos considerados pelo algoritmo de clusterização, para cada proteína testada. Apenas resíduos que fazem parte das estruturas secundárias presentes na estrutura de referência de RMN são levados em consideração.

Código PDB	Intervalo
1L2Y	2-8,11-14
1RIJ	2-11,14-16
1VII	4-8,15-18,23-32
1UNC	3-8,14-19,22-31
1UAO	1-10
1LE1	2-5,8-11
1E0L	4-7,13-17,22-24
1FME	8-12,19-23,29-30
1PSV	2-3,11-12,15-23
2WXC	10-19,23-25,37-46

6. RESULTADOS E DISCUSSÃO - PARTE 2: ESTUDO DE CASO DA PROTEÍNA *VILLIN HEADPIECE* DE CÓDIGO PDB 1UNC

Este capítulo apresenta a segunda parte dos resultados e discussão desta tese, composto pelo estudo de caso para a proteína *villin headpiece*, de código PDB 1UNC. Inicialmente, é avaliada a adequação da nova abordagem em relação à distribuição de Boltzmann, seguindo-se as análises referentes à reprodutibilidade da abordagem e à diversividade na amostragem da superfície de energia, se comparada ao método REMD convencional. As taxas de aceitação de movimentos de Monte Carlo são também avaliadas, assim como a eficiência na amostragem estrutural e a convergência das simulações. Por fim, apresentam-se os resultados referentes à descoberta de estruturas próximas à nativa, e ainda a verificação quanto à correta adequação de EES e à quantidade de estruturas enoveladas amostradas.

6.1 CuT-REMD Segue uma Distribuição de Boltzmann

As Figuras 6.1 e 6.2 apresentam gráficos de pontos para a Equação 5.1, aplicado a cada temperatura adjacente, comparando CuT-REMD e REMD convencional para a proteína de código PDB 1UNC. Na Figura 6.1, apresenta-se a comparação dos protocolos A, C e E, os quais diferem no tamanho de corte e tempo de permanência em pontos de corte mais curtos, mas mantêm o mesmo EAF na Etapa 6 do protocolo (Tabela 5.1). As comparações de B, D e F podem ser encontradas na Figura 6.2. Ambas representam uma de três simulações para cada protocolo.

Os experimentos de simulação, executados em triplicata, retornaram resultados semelhantes para diferentes números de semente aleatória, exceto para a Etapa 3 no protocolo E (REMD convencional), no qual uma das três simulações (Figura 6.1, Etapa 3) retornou um comportamento inverso ao esperado para o último par de temperaturas (511,14 K e 537,54 K). A verificação dessa área de sobreposição de energia particular mostrou que o sistema atingiu maiores probabilidades de atingir energias mais baixas a temperaturas mais elevadas do que a temperaturas mais baixas. Esse é um comportamento inesperado, o que leva a erros maiores ao validar o coeficiente angular das curvas geradas contra o coeficiente angular ideal para uma distribuição de Boltzmann. No entanto, analisando as energias totais acessadas pelas simulações, observa-se o comportamento típico, isto é, as energias mais baixas são normalmente encontradas em temperaturas mais baixas.

No início das simulações CuT-REMD, devido aos raios de corte mais curtos (4,0 Å e 5,0 Å), o sistema mostrou uma propensão de alargamento da área de sobreposição entre energias potenciais de temperaturas adjacentes, resultando em um padrão de distribuição

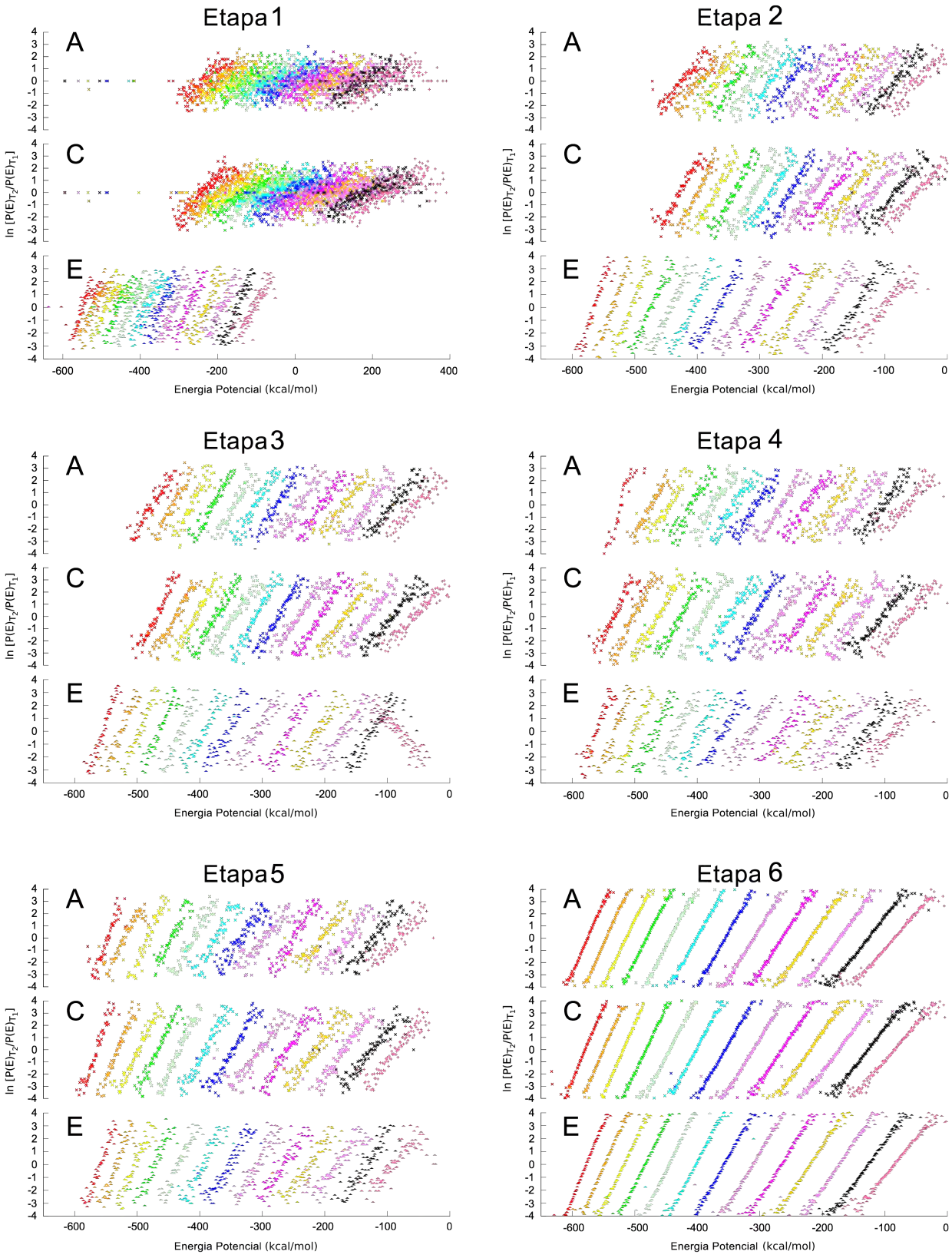

Figura 6.1 – Logaritmo natural da razão entre as distribuições de energia potencial de temperaturas adjacentes. Comparação das simulações CuT-REMD (A e C) contra simulações REMD convencional (E), para as Etapas 1 a 6 (ver Tabela 5.1). Todos os valores de P(E) foram computados utilizando-se uma janela de 1 kcal/mol.

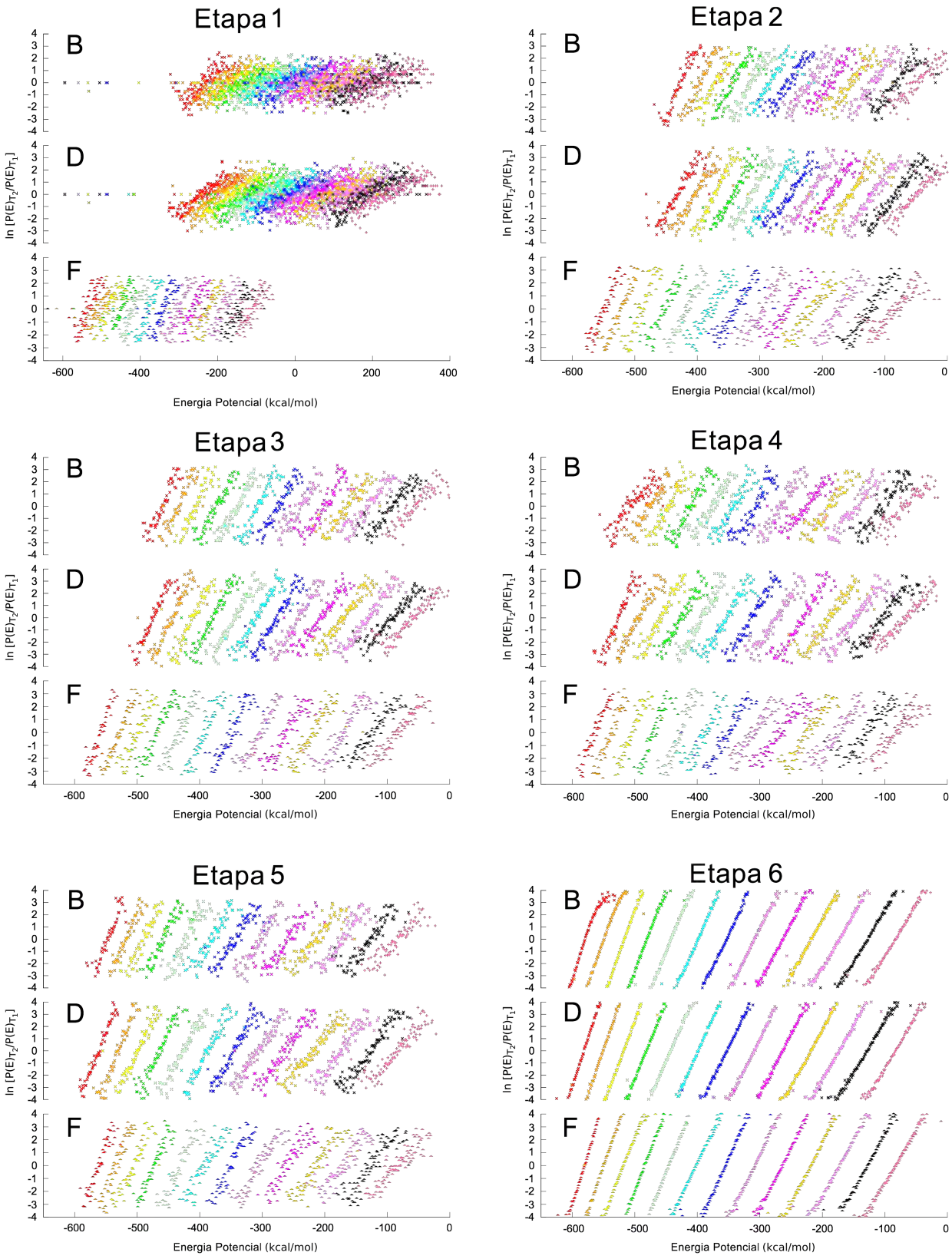

Figura 6.2 – Logaritmo natural da razão entre as distribuições de energia potencial de temperaturas adjacentes. Comparação das simulações CuT-REMD (B e D) contra simulações REMD convencional (F), para as Etapas 1 a 6 (ver Tabela 5.1). Todos os valores de P(E) foram computados utilizando-se uma janela de 1 kcal/mol.

menos inclinado (Figura 6.1, Etapas 1 e 2, A e C e Figura 6.2, Etapas 1 e 2, B e D). Esses resultados sugerem uma exploração mais ampla do espaço de energia [Shi13], em conformidade com o objetivo de favorecer a formação de EES nas fases iniciais da simulação. Como esperado, as energias potenciais mais baixas são maiores quando utilizando raios de corte mais curtos. O tamanho do raio de corte nas Etapas 1 e 2 limita as energias mais baixas a valores muito mais elevados do que aqueles encontrados quando a estrutura é formada por completo.

Os coeficientes de correlação médios entre os coeficientes angulares das curvas obtidas pelas simulações por CuT-REMD e REMD convencional em comparação ao coeficiente angular esperado (curvas teóricas) para as Etapas 1 a 6 estão resumidos na Tabela 6.1. As Tabelas C.1 até C.6, no Apêndice C exibem os resultados para todos os pares de temperaturas.

Tabela 6.1 – Coeficiente de correlação entre as curvas obtidas e a curva teórica para simulações de CuT-REMD e REMD convencional. Média de todos os pares de temperaturas para as Etapas 1 a 6. Na Tabela 5.1, podem ser obtidos detalhes sobre os IDs dos protocolos.

ID	Etapa 1	Etapa 2	Etapa 3	Etapa 4	Etapa 5	Etapa 6
A	0,465	0,752	0,915	0,91	0,911	0,985
B	0,477	0,781	0,903	0,902	0,905	0,986
C	0,615	0,845	0,937	0,936	0,938	0,983
D	0,610	0,798	0,938	0,885	0,94	0,987
E	0,481	0,715	0,754	0,766	0,764	0,799
F	0,341	0,803	0,919	0,914	0,913	0,986

Para todos os pares de temperaturas adjacentes, a pior média de *fitness* entre as curvas foi encontrada no início das simulações, independentemente do protocolo. Tal comportamento era esperado devido ao tempo de simulação muito curto empregado (1 ns ou 2 ns). Para o raio de corte de 4,0 Å, CuT-REMD C e D apresentaram melhor desempenho do que REMD convencional (E e F) atingindo uma relação linear ascendente forte (correlação entre 0,5 e 0,7) [Hop16], enquanto CuT-REMD A e B apresentaram desempenho semelhante ao REMD convencional E e F, os quais obtiveram uma relação linear ascendente moderada (correlação entre 0,3 e 0,5) [Hop16]. À medida que a simulação progride e o raio de corte é aumentado gradualmente, os resultados melhoram, assim como o *fitting* com as curvas teóricas, obtendo-se relações lineares ascendentes muito fortes (correlação entre 0,7 e 0,9) [Hop16] para REMD E e relações lineares ascendentes quase perfeitas (correlação superior a 0,9) [Hop16] para todos os outros protocolos. Tais resultados sugerem não apenas que os raios de corte de 5,0 Å, 6,0 Å e 7,0 Å levam à linearidade satisfatória, como também que a aplicação dos protocolos CuT-REMD C e D deve melhorar a capacidade de simulação quando utilizando raio de corte de 8,0 Å (Etapas 5 e 6).

Em conjunto, esses resultados sugerem que o aumento progressivo do raio de corte implementado por CuT-REMD mantém uma distribuição de Boltzmann. Para o protocolo REMD E, no entanto, mesmo no final de uma simulação de 50 ns, foi perceptível que o sistema não conseguiu um acordo satisfatório com o declive teórico (Tabela 6.1, protocolo E, Etapa 6).

6.2 Verificação de Reprodutibilidade

É importante para uma abordagem visando à predição de estruturas 3D de proteínas ser reprodutível. Para verificar isso, foram analisadas simulações em triplicata com os protocolos A, B, C, D, E e F (ver Tabela 5.1), calculando-se a sobreposição de todas as energias potenciais. Cada temperatura foi levada em conta separadamente, resultando na análise de 50.000 pontos de energia por temperatura. A Figura 6.3 ilustra esses resultados.

Figura 6.3 – Sobreposição de energia potencial entre triplicatas do mesmo protocolo em função da temperatura das réplicas. Verificação de reprodutibilidade para CuT-REMD e REMD convencional. Na Tabela reftabela-detalhes, podem ser obtidos detalhes sobre os IDs dos protocolos no eixo esquerdo. Valores mais altos simbolizam maior reprodutibilidade.

Para todos os protocolos, o uso de diferentes números de semente aleatória não alterou a exploração espacial de energia dos sistemas. O sistema acessou praticamente o mesmo espaço de energia, com uma sobreposição de quase unidade quando se avaliando a distribuição probabilística de energias. Isso mostra que CuT-REMD é uma abordagem reprodutível. Encontraram-se resultados semelhantes para simulações REMD convencionais.

Para temperaturas diferentes, observou-se uma ligeira sobreposição menor a temperaturas mais baixas, embora esta tenha permanecido também perto da unidade.

6.3 Diversidade na Amostragem do Espaço de Energia

Para verificar se a abordagem proposta por este estudo influenciou a amostragem de diferentes regiões do espaço de energia, aplicou-se a Equação 5.2 para comparar o espaço de energia explorado pelas simulações de CuT-REMD A e C e B e D, respectivamente contra REMD convencional E e F. Os resultados são apresentados na Figura 6.4. Em comparação com o REMD convencional E e F, o espaço de energia coberto pelos protocolos CuT-REMD A e C e B e D são notavelmente diferentes, especialmente quando se comparam C e D com E e F. Os resultados sugerem que, com os raios de corte em funcionamento, quanto mais a simulação se mantiver em raios de corte mais curtos, mais diferente é o espaço de energia coberto. Embora os raios de corte de aumento progressivo fossem aplicados somente nos primeiros nanossegundos das simulações, é evidente (Figura 6.4) o impacto que essa abordagem causa na exploração do espaço de energia até o final da simulação.

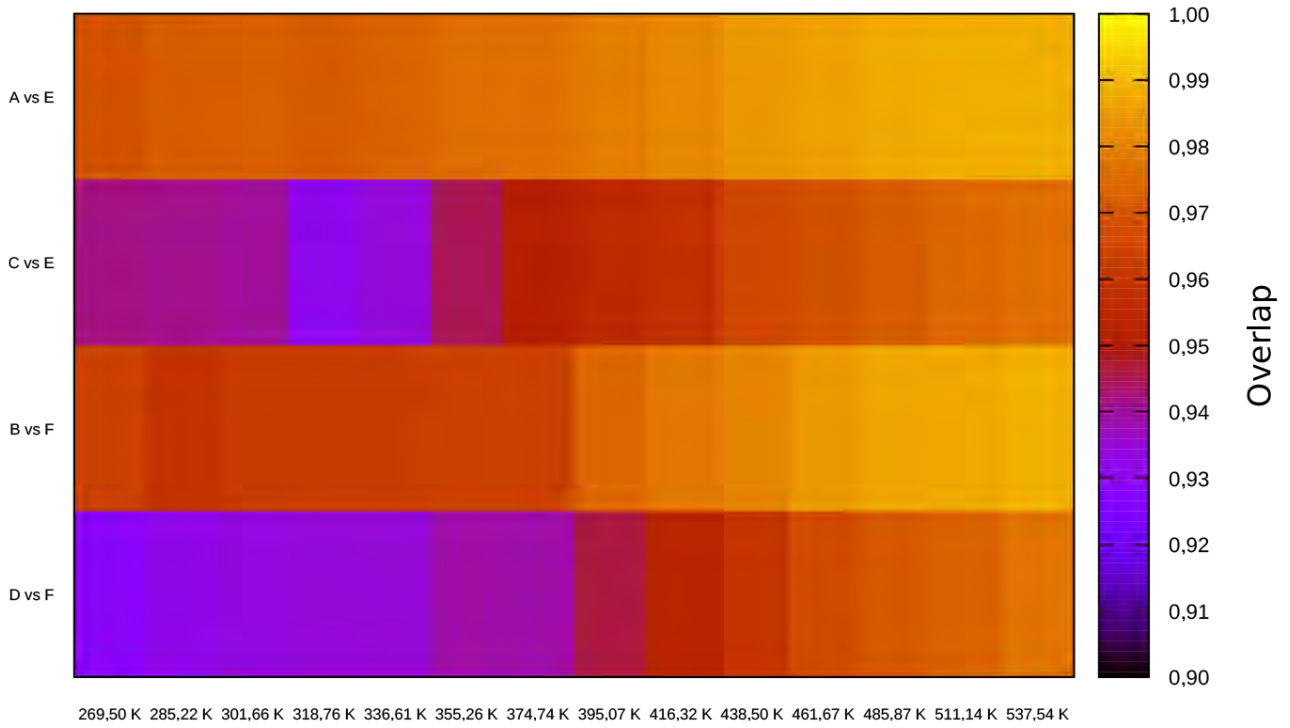

Figura 6.4 – Sobreposição de energia potencial entre triplicatas de protocolos diferentes. Comparação das superfícies de energia exploradas pelos diferentes protocolos CuT-REMD contra os protocolos REMD convencionais. Na Tabela reftabela-detalhes, podem ser obtidos detalhes sobre os IDs dos protocolos no eixo esquerdo. Valores mais altos simbolizam menor diversidade.

6.4 Aceitação de Movimentos de Monte Carlo

Para que um sistema utilize eficientemente os recursos REMD, é essencial inspecionar o número de réplicas, o qual aumenta em função da raiz quadrada do tamanho do sistema e da faixa de temperaturas [SO99]. Para permitir que as réplicas oscilem satisfatoriamente entre diferentes temperaturas, é necessário ter sobreposição suficiente entre as distribuições de energia potencial de temperaturas adjacentes [RCdP05, PPC05, Kof02, KK05]. Para verificar isso, foram feitos os cálculos de EAR e ETR para os protocolos de A a F.

Os resultados de EAR são mostrados na Figura 6.5. Para todas as simulações, temperaturas mais elevadas retornaram maiores taxas de aceitação. Para REMD convencional, a taxa de aceitação permaneceu constante durante toda a simulação, mesmo utilizando diferentes EAFs (Etapa 6 dos protocolos E e F).

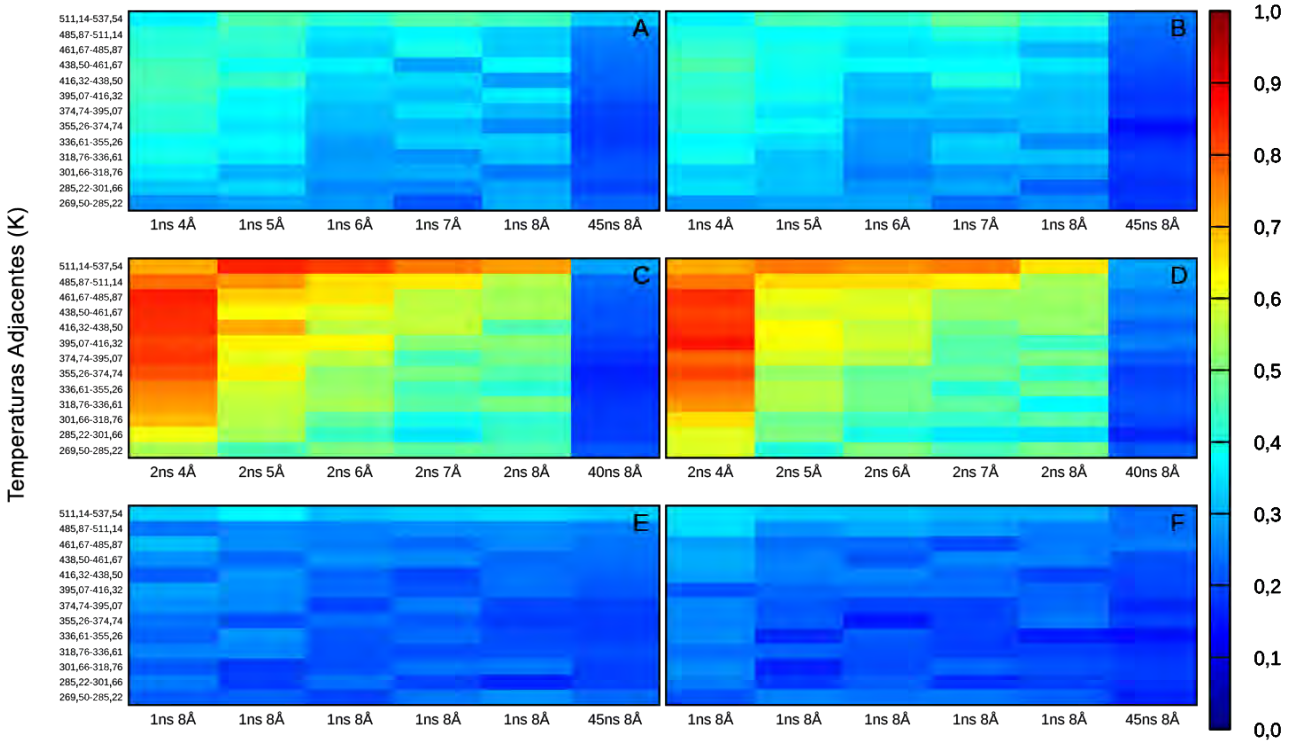

Figura 6.5 – Taxa de aceitação de intercâmbios ou *Exchange Acceptance Ratio* (EAR) para todas as temperaturas adjacentes. Resultados para os protocolos A, B, C e D de CuT-REMD, além de E e F de REMD convencional. Na Tabela reftabela-detalhes, podem ser obtidos detalhes sobre os IDs dos protocolos.

As simulações por REMD convencional retornaram uma EAR média de cerca de 0,23. Para as simulações CuT-REMD, no entanto, o tempo de permanência em cada raio de corte mais curto influenciou fortemente EAR: quanto mais tempo (1 ns ou 2 ns) a simulação permaneceu em raios de corte mais curtos, maior foi a aceitação de movimentos entre réplicas adjacentes. À medida que a simulação progride e o raio de corte é incrementado, EAR começa a diminuir, mantendo o padrão de valores mais altos em pares de

temperaturas mais elevadas. Isso pode ser observado nos protocolos A, B, C e D 6.5. Na Etapa 6, onde o raio de corte foi fixado em 8,0 Å e a simulação estendida por 40-45 ns, as flutuações de EAR não são perceptíveis, repetindo assim o comportamento observado em REMD convencional, mesmo com diferentes EAFs. Isso está de acordo com trabalhos anteriores [SMR08].

Comparando-se as simulações A e B (1 ns de tempo de permanência) com seus equivalentes C e D (2 ns de tempo de permanência), esta última retornou EAR 93 %, 63 %, 77 %, 53 % e 56 % superiores, respectivamente para as Etapas 1 a 5. Assim, propõe-se que gastar mais tempo em raios de corte mais curtos retorna maior sobreposição entre réplicas. Tais resultados sugerem que CuT-REMD pode ser aplicado usando o mesmo intervalo de temperatura, porém reduzindo o número de réplicas, levando assim a ganho computacional.

Em relação a ETR (Figura 6.6), a abordagem proposta neste estudo mostrou-se novamente eficaz, uma vez que a estratégia de aumento gradual de raio de corte diminuiu consideravelmente ETR. Esse resultado sugere que, em raios de corte mais curtos, o sistema é mais livre para se mover entre as temperaturas, podendo amostrar um espaço conformacional mais amplo. Isso é consistente com os resultados das Seções 6.1 e 6.3. Adicionalmente, os EAFs inferiores (testados na Etapa 6) apresentaram menores taxas de aprisionamento, corroborando as descobertas de Sindhikara *et al.* [SMR08].

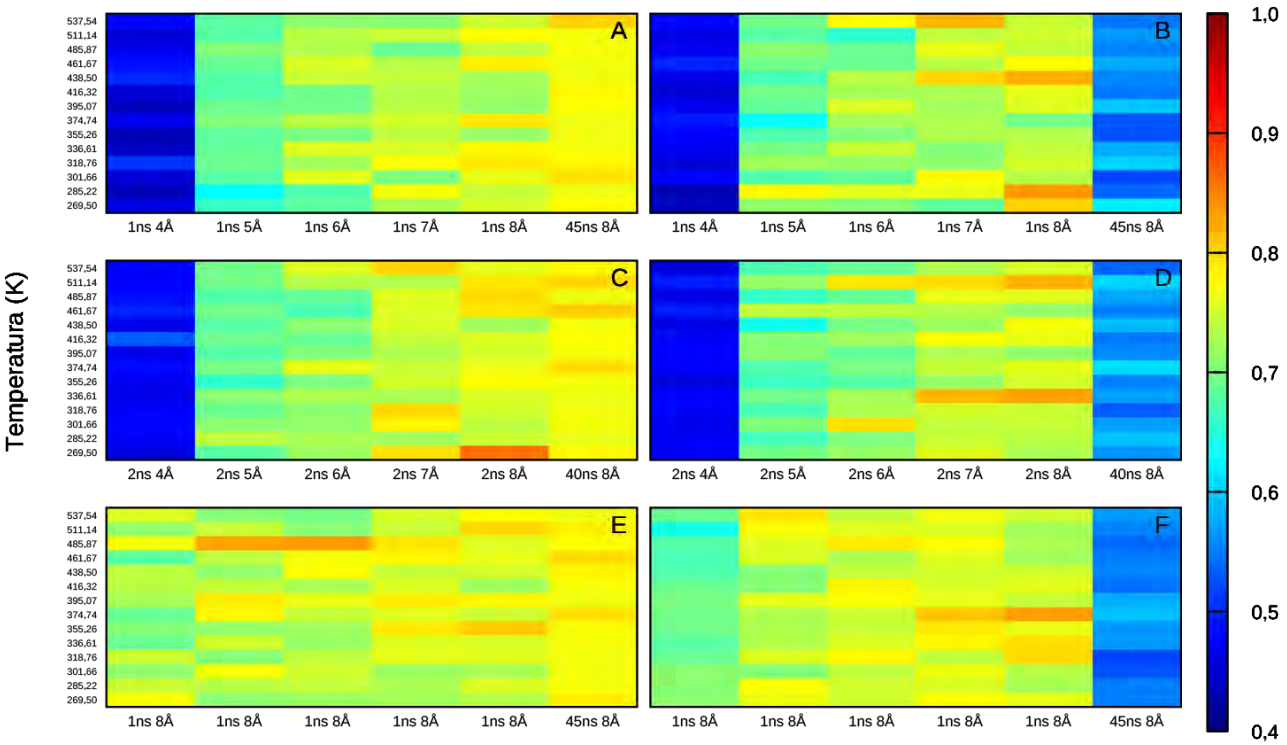

Figura 6.6 – Taxa de Aprisionamento entre Intercâmbios ou *Exchange Trapping Ratio* (ETR) para cada temperatura individual. Resultados para os protocolos CuT-REMD A, B, C e D e REMD convencional E e F.

6.5 Análise de Espaço de Conformações

6.5.1 Eficiência de Amostragem

A eficiência de amostragem varia entre protocolos. Para averiguá-los, foram então calculados os tempo de TE para os protocolos A, B, C, D, E e F (Table 3).

Tabela 6.2 – Tempo médio para completar um Evento de Tunelamento ou *Tunneling Event* (TE), para todos os protocolos de simulação.

ID	Tempo de TE médio (ns)
A	2,5
B	2,5
C	1,5
D	1,4
E	5,8
F	13,9

Analisou-se o impacto do tempo de permanência em raios de corte mais curtos e o impacto de diferentes EAFs na Etapa 6. A partir dessa análise, foi possível observar diferenças claras na capacidade de amostragem de CuT-REMD em comparação com as simulações REMD convencionais. Quanto mais tempo a simulação permaneceu em raios de corte mais curtos, menor o tempo necessário para completar um TE, caracterizando assim melhor amostragem. Como consequência dessa otimização, pode-se reduzir o número de réplicas nas simulações (como examinado em trabalho anterior por Nadler e Hansmann [NH07]), o que, por sua vez, aumentaria consideravelmente o ganho computacional.

Da análise de EAF, verificou-se ainda que o parâmetro não afetou significativamente a amostragem de simulações CuT-REMD, porém influenciou fortemente simulações padrão REMD, uma vez que o valor EAF moderado de 1 ps^{-1} fornece melhor amostragem do que os inferiores (0,025 ps^{-1} e 0,020 ps^{-1}), conforme sugerido em outras fontes [SMR08, SER10].

6.5.2 Convergência do Espaço Conformacional

A Figura 6.7 mostra a taxa de convergência para todos os protocolos. Analisaram-se os instantes, em cada simulação, em que a divergência de Jensen-Shannon se tornou < 0,01 (linhas tracejadas verticais na Figura 6.7). Isso é importante porque, acima desse limiar, o *ensemble* não gera informação significativa (não são visitadas novas conformações). É notável a partir do gráfico que os protocolos CuT-REMD B e C conduzem à mais rápida convergência entre os nove protocolos testados. Classificando-os pela taxa de convergência DRES, os mais rápidos seguiriam a ordem C, B, G, D e E, o que significa que quatro dos

cinco protocolos mais rápidos são protocolos CuT-REMD. Por outro lado, se fossem listados os protocolos mais lentos, a ordem seria F, I, A, H e E, o que significa que três dos cinco protocolos mais lentos são convencionais. Embora isso seja encorajador, observa-se que as diferenças entre os resultados de convergência dos diferentes protocolos foram consideravelmente baixas. Mais importante, entretanto, é o fato de que as novas abordagens aqui apresentadas mostraram o padrão de convergência esperado, não prejudicando o sistema.

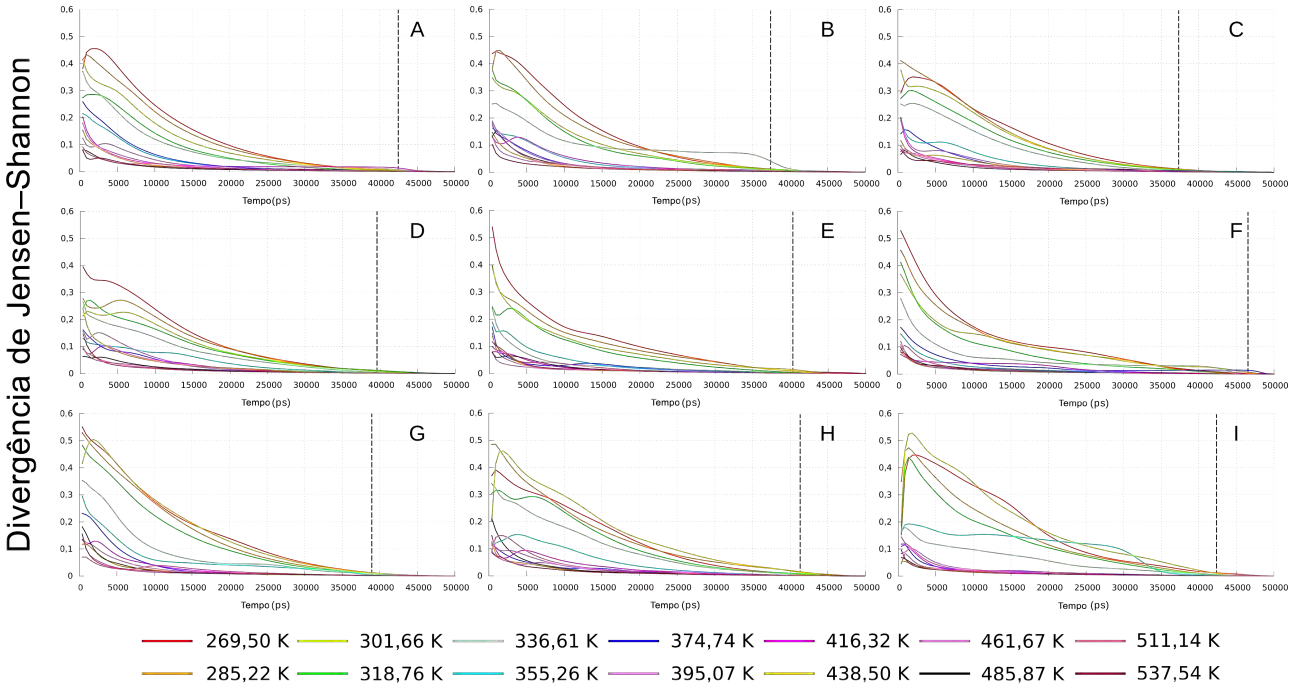

Figura 6.7 – Taxa de convergência em função do tempo de simulação para todos os protocolos, medidos pela divergência de Jensen-Shannon. A divergência de Jensen-Shannon foi calculada com o algoritmo DRES no ENCORE para todas as temperaturas (curvas coloridas). Os cálculos de convergência usaram uma janela de tempo de 20 quadros (400 ps). A linha tracejada vertical destaca o instante em que o sistema obtém 99 % de amostragem conformacional.

Em simulações REMD, as réplicas em temperaturas mais elevadas não possuem apenas a capacidade de avançar por entre a rugosa superfície de energia, mas também convergem mais rapidamente, uma vez que mais mínimos se tornam acessíveis a essas temperaturas (Figura 6.7, protocolos A a F). Tais resultados estão de acordo com os trabalhos anteriores [ZWD05, ROS07], demonstrando um limite na capacidade de amostragem à medida que a temperatura aumenta. Isso pode indicar que, especialmente em REMD para fins de PSP, altas temperaturas operam simplesmente como um motor para permitir que o sistema salte barreiras de energia, e não visam fornecer ao sistema estruturas 3D satisfatórias. No entanto, se a simulação encontra estruturas 3D adequadas em temperaturas mais altas, o sistema tem a propensão de permutá-las até temperaturas mais baixas.

6.6 Descobrindo Estruturas Próximas à Nativa

6.6.1 Análise da Formação de EES e Estruturas Terciárias Enoveladas

Para verificar a capacidade da abordagem em retornar estruturas próximas à nativa, analisou-se a adequação de EES corretos e o número de estruturas enoveladas corretamente. A DM convencional (protocolo I) foi utilizada como simulação de controle para quantificar a Taxa de Melhoria (TM) de EES e estruturas enoveladas nos demais protocolos. TM foi arbitrariamente ajustado para 1,0 para o protocolo I. TMs superiores a 1,0 indicam aumento múltiplo em relação a esse protocolo em particular (Tabela 6.3).

Tabela 6.3 – Taxa de melhoria (TM) na formação de EES e estruturas terciárias enoveladas. Todos os valores na tabela são relativos ao protocolo I de DM convencional.

ID	TM de EES corretos	TM de estruturas enoveladas
A	4,6	1,6
B	3,9	8,9
C	7,9	2,2
D	8,8	4,5
E	3,9	6,4
F	4,4	1,6
G	2,0	1,8
H	1,8	13,1
I	1,0	1,0

Analisando a quantidade de EES atribuídas corretamente, CuT-REMD mostrou comportamento semelhante para os protocolos A e B se comparado a REMD convencional E e F. Por outro lado, CuT-REMD de protocolos C e D atingiu resultados significativamente melhores que REMD convencional, bem como Cu-MD simulações G e H quando comparado com o padrão MD (protocolo I). Os melhores resultados foram obtidos pelos protocolos C e D, onde o sistema foi mantido 2 ns a cada raio de corte antes de seu incremento.

No que diz respeito às estruturas enoveladas e considerando simulações empregando o mesmo EAF, CuT-REMD (protocolos A e C) não superou REMD convencional (protocolo E). No entanto, os protocolos B e D de CuT-REMD obtiveram os melhores resultados, destacando-se o impacto positivo do uso de EAFs inferiores em detrimento a moderados. Esse desempenho também foi observado para simulações por MD convencional (sem REMD), para as quais foi obtido o melhor desempenho (Cu-MD H), um resultado inicialmente surpreendente. A partir de uma análise mais criteriosa sobre o fato, percebe-se que tal resultado foi contabilizado considerando o limiar de 3,5 Å RMSD para considerar-se estruturas como enoveladas. Agora, por exemplo, e se 100 % dessas estruturas enoveladas possuíssem RMSD de exatamente 3,5 Å? Tais estruturas seriam contabilizadas, porém existe a possibilidade de, dentre as estruturas entendidas como enoveladas, não existirem

estruturas com RMSD de 2,0 Å ou 1,5 Å, valores muito mais próximos das estruturas nativas. Dada essa introdução e após verificação, na seção abaixo (6.6.2) são exibidas evidências contundentes de que o CuT-REMD é a abordagem que retornou melhores resultados. De todo modo, os resultados da Tabela 6.3 reforçam a proposição de que um esquema de raio de corte incremental como metodologia alternativa é capaz de melhorar a capacidade de simulações em atingir estruturas nativas.

Impacto de CuT-REMD na Estabilização de Hélices

Ainda em relação à formação de EES, a fim de se verificar o impacto da utilização de raios de corte curtos na estabilização das hélices presentes na proteína *villin headpiece*, analisou-se separadamente cada uma das três hélices que a formam, sendo a primeira formada pelos resíduos Ile3 até Gln8; a segunda, pelos resíduos Pro14 até Ala 19; e a terceira, pelos resíduos Arg22 até Glu31. O protocolo D, de melhor desempenho geral e também escolhido como o protocolo a ser aplicado ao conjunto teste de proteínas (o que será abordado mais à frente), foi avaliado, em comparação à simulação por REMD convencional correspondente (em relação aos demais parâmetros de simulação) F.

A Figura 6.8 exibe, para cada hélice, o RMSD computado entre a estrutura nativa e a simulada, durante toda a simulação (50 ns), para as 4 temperaturas mais baixas. A partir da análise, mais uma vez foi perceptível ser nas temperaturas mais baixas onde as melhores estruturas se situam, para essa proteína, e assim sendo, será apresentado apenas um dos gráficos gerados (temperaturas mais baixas). No Apêndice D, estão contidos os gráficos referentes às demais temperaturas. Uma vez que os 50.000 pontos (1 a cada ps) para cada temperatura tornaram o gráfico de difícil entendimento, optou-se pela aplicação de um filtro de suavização das curvas (*smoothing*). Enfatiza-se, no entanto, que embora o filtro facilite a inspeção visual do que acontece durante as simulações, ele retira do gráfico o fator precisão. Desse modo, os resultados devem ser entendidos como padrões porém não como representantes dos valores exatos (em RMSD) atingidos pelas simulações.

Para a hélice de número 1, os protocolos aplicados retornaram comportamento semelhante, ainda que se verifique que a utilização de raios de corte mais curtos levou o sistema a atingir valores de RMSD não obtidos pela simulação REMD convencional, e em menos tempo. Os menores valores de RMSD da hélice 1, no entanto, não demonstraram se estabilizar.

Seguindo a análise, CuT-REMD demonstrou evidente maior aptidão na estabilização da segunda hélice. Embora ambas as abordagens tenham atingido limiares similares quanto ao menor valor de RMSD atingido pelas hélices, CuT-REMD foi a único capaz de manter a hélice em tal limiar até o final da simulação. Os resultados atestam que REMD manteve a estabilidade da segunda hélice até aproximadamente 30 ns de simulação, não sendo capaz de levá-la até o final da simulação. Não obstante, CuT-REMD ainda propagou

a estabilidade da hélice a mais de uma temperatura/réplica, conferindo ao sistema maior capacidade de estabilizar tal estrutura regular, o que suporta a ideia de que raios de corte mais baixos favorecerem formação de hélices.

O comportamento relativo à terceira hélice foi similar ao obtido para a hélice de número 2: uma vez estabilizada a estrutura, CuT-REMD manteve a estabilidade da hélice até o final da simulação, mantendo ainda a alta difusão entre temperaturas diferentes, o que significa ter mais de uma trajetória com hélices estáveis. Além disso, ao comparar-se os resultados de CuT-REMD aos de REMD convencional, verifica-se que, embora a hélice em REMD tenha permanecido estável, esta não foi capaz de estabilizar-se em RMSDs tão baixos quanto os obtidos por CuT-REMD, diferenciando-se cerca de 0,5 Å.

6.6.2 Habilidade de Amostrar Estados Próximos ao Nativo

Para cada protocolo de simulação, foi capturado o melhor RMSD/GDT-TS encontrado em cada temperatura (Figura 6.9 e Figura 6.10). Para facilitar a comparação visual entre as abordagens, os dados foram agrupados em quatro gráficos diferentes, cada um mostrando os resultados para os protocolos CuT-REMD e Cu-MD contra REMD e MD convencionais.

É possível observar que, em todos os casos, CuT-REMD foi capaz de encontrar melhores valores de RMSD/GDT-TS, chegando a estruturas com RMSD abaixo de 1,8 Å e GDT-TS acima de 0,8, o que não foi atingido com REMD convencional. Conforme esperado, também se observou que em temperaturas mais altas há uma propensão diminuída para obtenção de estruturas de alta qualidade [ZWD05, ROS07].

Para cada protocolo, foram agrupadas as estruturas mais semelhantes das trajetórias em *clusters*, uma prática comum na PSP58, e foram calculados *Best5Pop* e *BestStruc*58. A Tabela 6.4 mostra os resultados para todos os abordagens/protocolos testados.

Tabela 6.4 – Avaliação do desempenho dos diferentes protocolos testados de acordo com os critérios *Best5Pop* e *BestStruc*. Os cálculos de RMSD (em Å) foram realizados utilizando apenas os carbonos α das predições e da estrutura de RMN experimental (código PDB 1UNC).

ID	Best5Pop	BestStruc
A	4,9	1,5
B	4	1,7
C	4,8	1,5
D	3,8	1,4
E	4,3	2,0
F	4,8	1,8
G	6,3	1,9
H	5,4	1,3
I	5,7	2,1

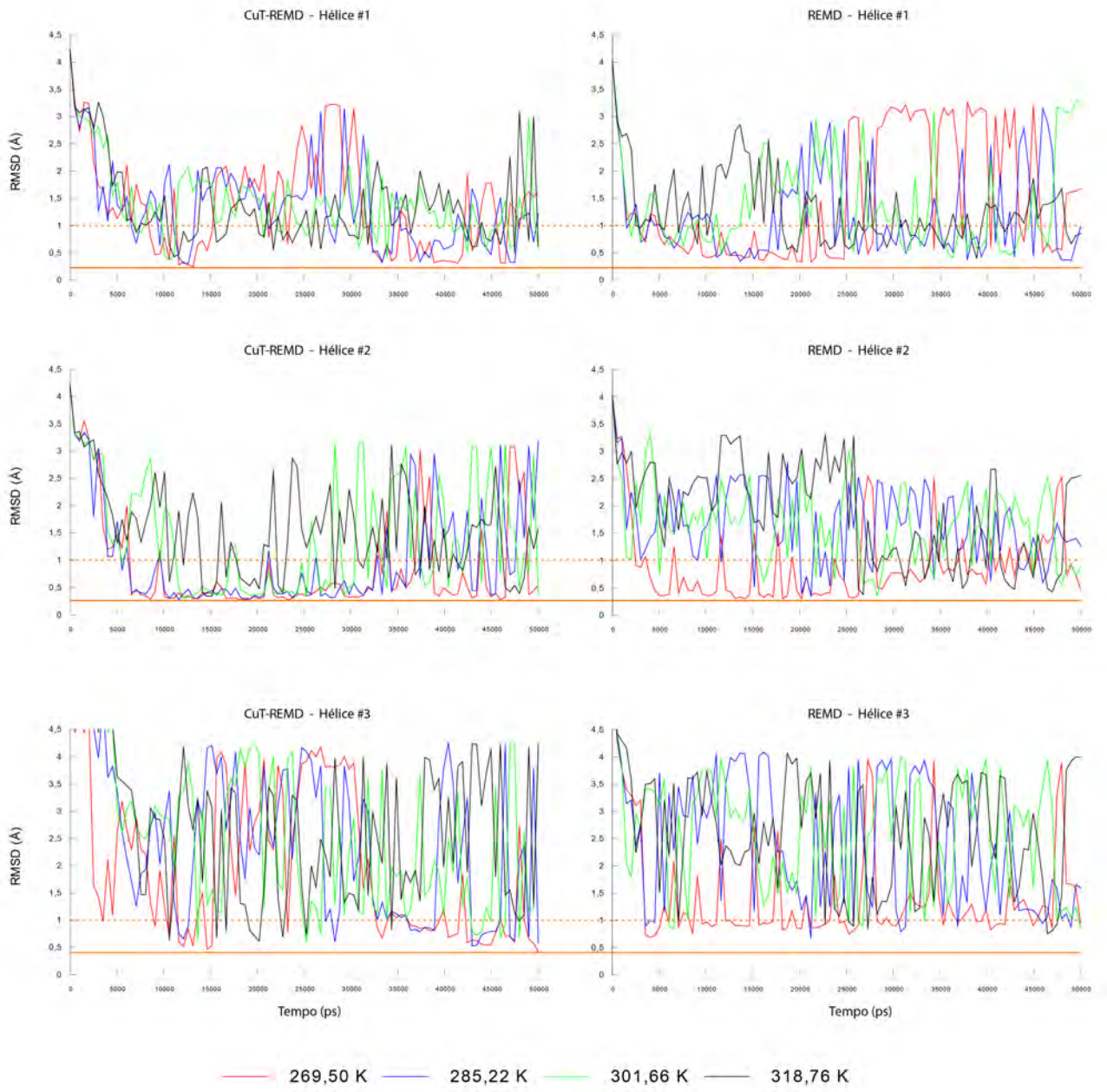

Figura 6.8 – Análise comparativa entre CuT-REMD e REMD convencional na formação individual das três hélices que compõem a proteína *villin headpiece*. Resultados para as temperaturas 269,50 K, 285,22 K, 301,66 K e 318,76 K. Em laranja, as linhas pontilhadas e contínuas representam, respectivamente, o limiar de 1 Å e o menor valor de RMSD (considerando a suavização da linha).

Pelo critério *BestStruc*, CuT-REMD e Cu-MD apresentaram estruturas preditas com RMSDs mais baixas que as abordagens convencionais (Tabela 6.4). Analisando o critério *Best5Pop*, verificou-se que as melhores estruturas preditas foram obtidas por CuT-REMD B e D, com uma melhora de 0,5 Å em comparação com a melhor estrutura de REMD convencional E. Comparando os protocolos A, B, C, D, G e H, quanto mais tempo a simulação permaneceu em raios de corte mais curtos e empregou EAF mais baixo, melhores foram os resultados. Observou-se também que quanto maior o tempo de permanência em

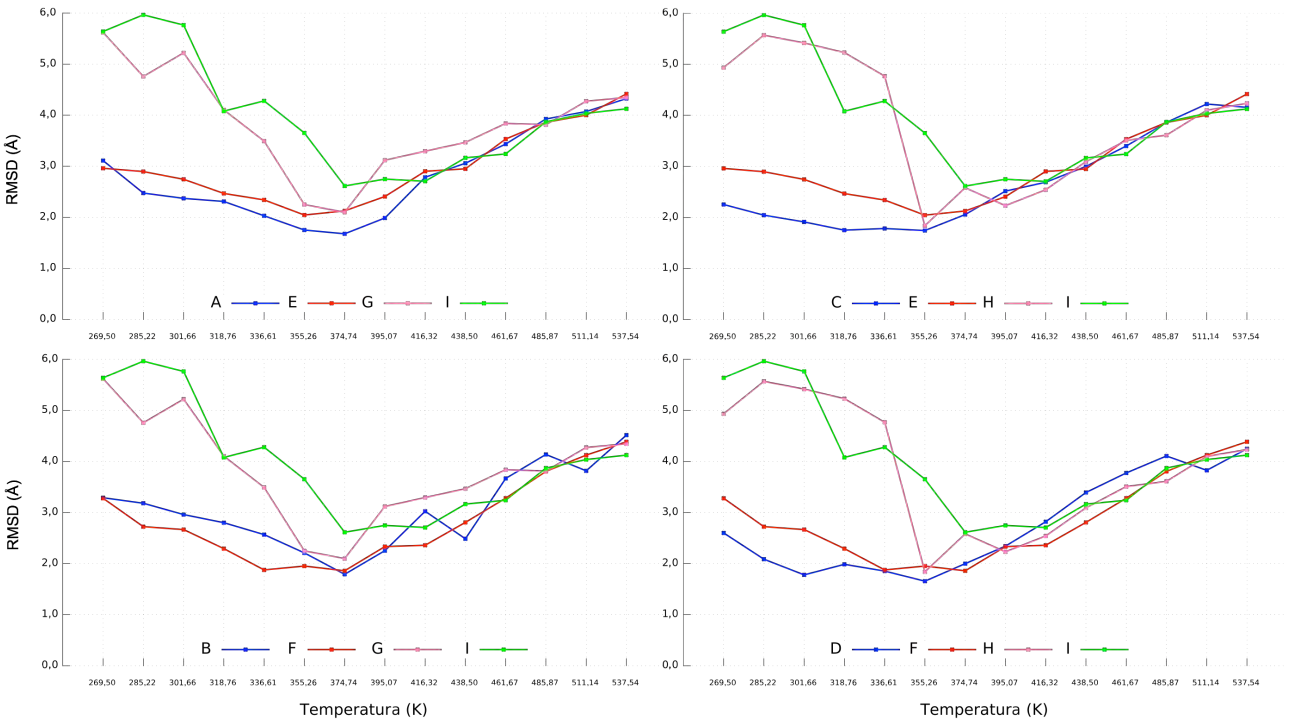

Figura 6.9 – Média do melhor RMSD para cada temperatura. Comparação de desempenho de CuT-REMD (protocolos A, B, C e D) e Cu-MD (protocolos G e H) contra simulações de REMD convencional (protocolos E e F) e DM convencional (protocolo I).

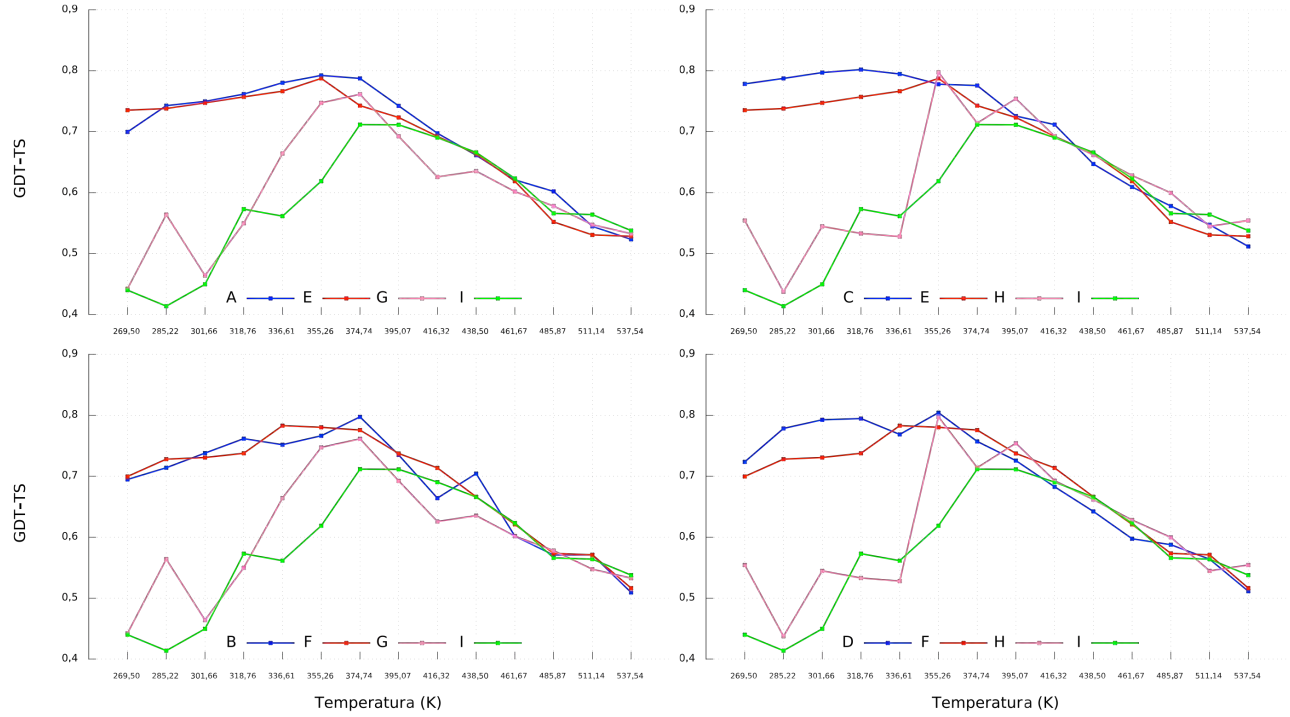

Figura 6.10 – Média do melhor GDT-TS para cada temperatura. Comparação de desempenho de CuT-REMD (protocolos A, B, C e D) e Cu-MD (protocolos G e H) contra simulações de REMD convencional (protocolos E e F) e DM convencional (protocolo I).

raios de corte mais curtos, maior o número total de *clusters* na etapa de captura, indicando

uma exploração mais ampla do espaço conformacional. A Figura 6.11 exibe as estruturas *Best5Pop* e *BestStruc* obtidas por CuT-REMD D.

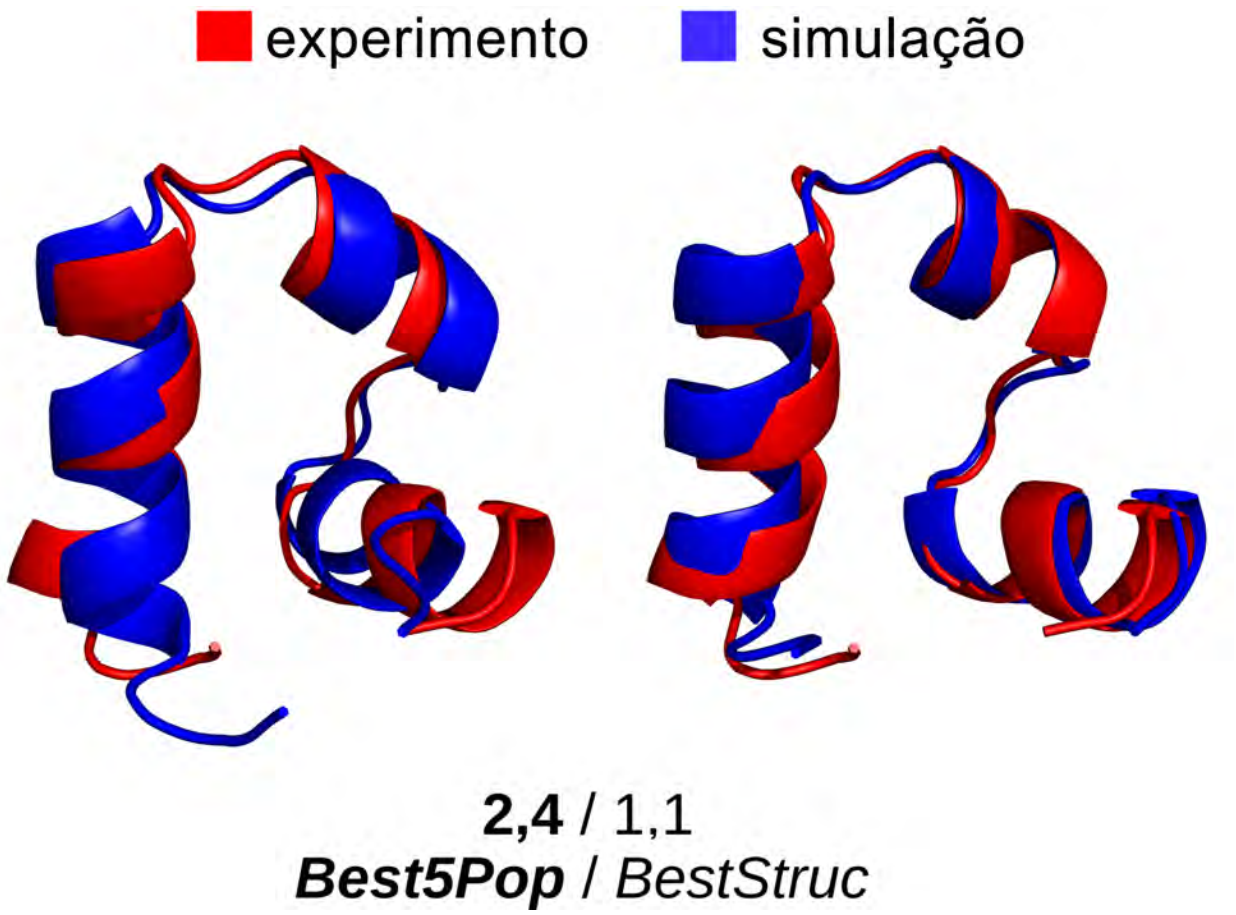

Figura 6.11 – Superposição da estrutura predita em comparação à experimental para a proteína de código PDB 1UNC. *Best5Pop* e *BestStruc* são mostrados em azul; e a estrutura nativa, em vermelho. Resultados obtidos para CuT-REMD D, com tempo de permanência de 2 ns em raios de corte mais curtos e EAF de 0,02 ps^{-1}.

Seguindo as considerações de Roitberg *et al.* [ROS07] e Zhang *et al.* [ZWD05], decidiu-se investigar ainda mais as trajetórias resultantes das quatro temperaturas mais baixas. Essas temperaturas também representam as temperaturas mais próximas daquela (294,0 K) empregada na resolução experimental de RMN de 1UNC. Neste estudo, analisou-se a distribuição de RMSDs das estruturas previstas por diferentes protocolos, para todos os métodos (Figura 6.12A). Também, calculou-se a porcentagem de estruturas preditas geradas por cada protocolo como uma função de um conjunto de faixas de RMSD (Figura 6.12B).

Para além do avanço proporcionado pelas abordagens baseadas em REMD sobre as baseadas em DM (Figura 6.12A), considerações significativas podem ser feitas a partir da Figura 6.12B, onde fica claro que os protocolos CuT-REMD, de fato, possibilitaram a obtenção de estruturas com menores RMSDs. Embora todos os protocolos tenham mostrado desempenho semelhante visitando estruturas acima de 5,0 Å (Figura 6.12B), uma vez que

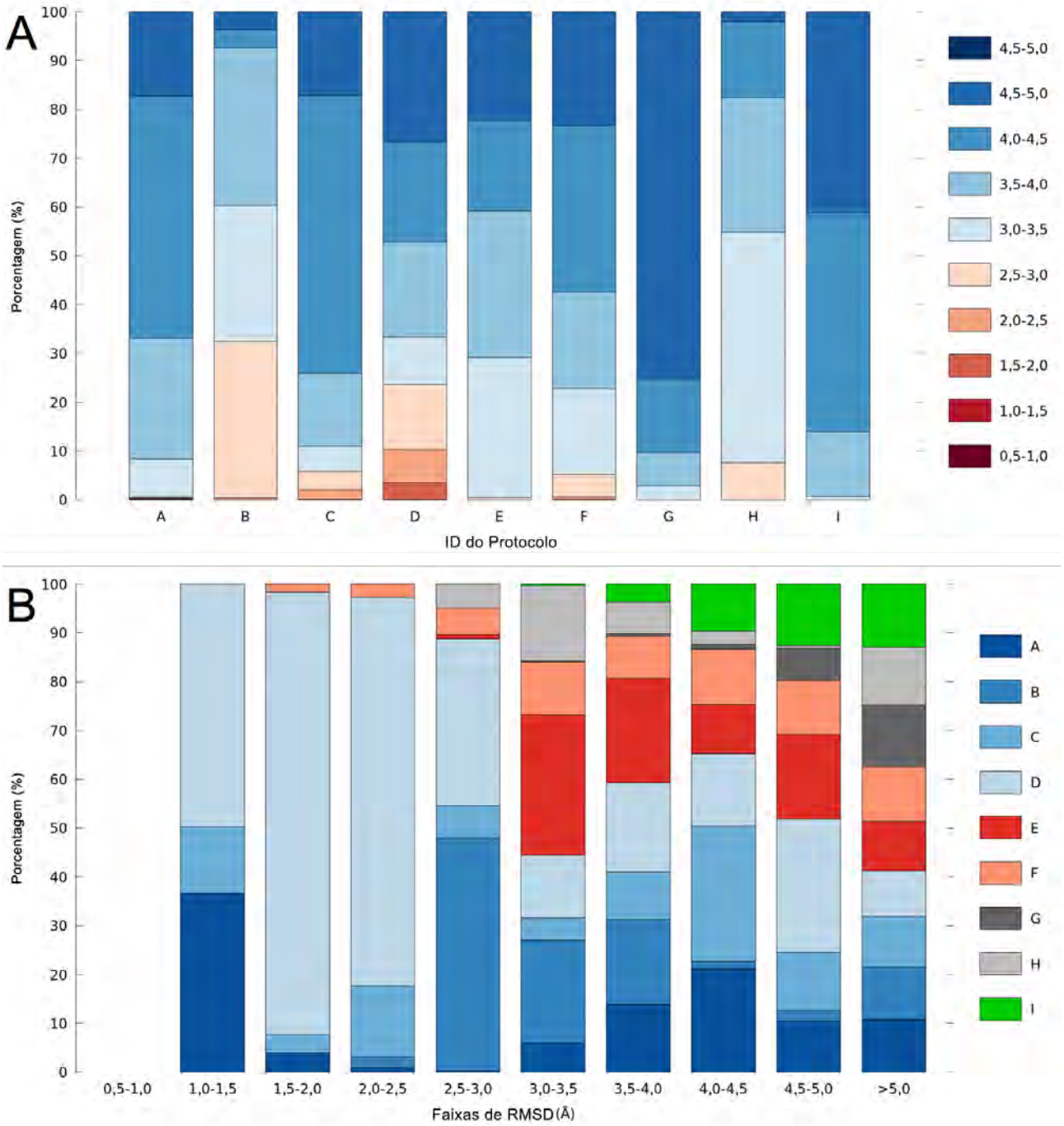

Figura 6.12 – Análise de RMSD utilizando histogramas empilhados por percentagem. (A) exibe a distribuição de RMSDs das estruturas amostradas por cada um dos diferentes protocolos, individualmente. (B) quantifica a porcentagem de estruturas amostradas por cada protocolo em faixas específicas de RMSD (intervalos de 0,5 Å).

foram investigadas estruturas de menor RMSD, as simulações CuT-REMD destacam-se dos outros protocolos, liderados pelo protocolo D, o qual revelou excelente capacidade de atingir estruturas entre 1,0 e 1,5 Å. Apesar das suas limitações, o protocolo Cu-MD H também mostrou avanços contra MD convencional.

7. RESULTADOS E DISCUSSÃO - PARTE 3: CONJUNTO TESTE DE PROTEÍNAS

Neste capítulo, serão expostos os resultados referentes à aplicação da abordagem CuT-REMD para conjunto teste de proteínas (ver Tabela 5.2). As avaliações foram feitas em duas etapas:

1. CuT-REMD *versus* REMD convencional; e
2. CuT-REMD *versus* Literatura.

Nesta tese, buscou-se um método de obtenção de estruturas que não comparasse todas as estruturas amostradas com a nativa e retornasse aquela mais similar à estrutura de referência, motivo da utilização da métrica *Best5Pop*. No entanto, pelo fato de *Best5Pop* considerar apenas estruturas oriundas de temperaturas mais baixas, existe a possibilidade de uma dada abordagem ser efetiva na amostragem de estruturas próximas à nativa porém obtendo tais conformações em temperaturas não contempladas pela metodologia de captura. Tal situação ocasiona desperdício de conformações de alta qualidade. Por esse motivo, para ambas as etapas enumeradas acima, foi avaliada a capacidade exploratória das abordagens na obtenção não apenas de *Best5Pop* como também de *BestStruc* [PMD15].

Com o intuito de facilitar o entendimento, enfatiza-se que tais métricas fazem parte da abordagem de captura de estruturas desta tese e estão descritas na metodologia do trabalho (seção 4.4).

7.1 CuT-REMD *versus* REMD Convencional

A Tabela 7.1 faz um comparativo entre o CuT-REMD e REMD convencional, para todas as proteínas do conjunto teste.

Avaliando-se quantitativamente os resultados de *Best5Pop* e *BestStruc*, CuT-REMD se comportou de maneira mais eficaz para as proteínas de classe α e $\alpha\beta$, sendo menos eficaz para proteínas de classe β. Percebe-se ainda relevante diferença entre as comparações envolvendo cálculos de GDT-TS e RMSD. Considerando-se GDT-TS, CuT-REMD atingiu melhores resultados de *Best5Pop* para cinco das nove proteínas testadas ou 56%; já considerando-se RMSD, esse número aumenta, passando para sete das nove proteínas ou 78%.

Observando os resultados de *BestStruc* (embora não sejam o foco de uma abordagem cega de predição), verificou-se que, considerando GDT-TS, para apenas 1 das proteínas (11%), CuT-REMD obteve resultados inferiores aos de REMD convencional. Considerando-se RMSD e valores mais baixos, no entanto, esse número passa para 4 ou 44%.

Tabela 7.1 – Comparação entre CuT-REMD e REMD convencional em relação aos resultados obtidos para *Best5Pop* e *BestStruc* para cada proteína presente no conjunto de testes. Células em cinza simbolizam melhores resultados.

		GDT-TS			
		CuT-REMD		REMD	
Classe	Código PDB	*Best5Pop*	*BestStruc*	*Best5Pop*	*BestStruc*
α	1L2Y	0,95	1,00	0,93	1,00
α	1RIJ	0,91	0,96	0,84	0,93
α	1VII	0,60	0,69	0,56	0,68
β	1UAO	0,78	1,00	0,80	1,00
β	1LE1	0,65	0,85	0,67	0,85
β	1E0L	0,38	0,61	0,43	0,55
$\alpha\beta$	1FME	0,43	0,66	0,54	0,65
$\alpha\beta$	1PSV	0,62	0,71	0,54	0,65
$\alpha\beta$	2WXC	0,38	0,49	0,34	0,50

		RMSD (Å)			
		CuT-REMD		REMD	
Classe	Código PDB	*Best5Pop*	*BestStruc*	*Best5Pop*	*BestStruc*
α	1L2Y	0,53	0,30	0,70	0,34
α	1RIJ	0,83	0,64	1,39	0,83
α	1VII	4,57	2,35	5,08	2,55
β	1UAO	2,70	0,39	2,22	0,36
β	1LE1	3,29	1,94	3,41	1,40
β	1E0L	6,30	5,11	6,23	4,78
$\alpha\beta$	1FME	4,34	2,84	5,33	2,85
$\alpha\beta$	1PSV	3,93	2,93	4,81	2,85
$\alpha\beta$	2WXC	5,18	3,95	7,49	7,38

Fazendo-se a análise de *Best5Pop* e *BestStruc*, foi possível constatar também que em nenhum caso foi possível capturar, por meio de *Best5Pop*, a melhor estrutura amostrada pelas simulações. Ainda assim, no entanto, percebe-se que a abordagem de captura foi efetiva, pelo fato das estruturas retornadas situarem-se, em média, 1,2 Å distantes de *BestStruc* (de RMSD) nas simulações CuT-REMD, e 1,5 Å nas simulações de REMD convencional. A fim de explorar melhor tais resultados, seguem as subseções 7.1.2 e 7.1.1.

7.1.1 Capacidade Exploratória *Best5Pop*

As análises a seguir levam em consideração apenas as trajetórias (demultiplexadas) das réplicas a temperaturas mais baixas. Tal restrição tem relação com o procedimento empregado para captura de conformações representativas da abordagem de predição de estruturas (considera apenas estruturas oriundas das 4 temperaturas mais baixas). A Figura 7.1 apresenta um panorama geral comparativo entre CuT-REMD e REMD convencional, para todas a proteínas do conjunto teste, considerando intervalos de GDT-TS e RMSD.

Por meio da análise da figura e das colunas apresentadas, ao serem comparadas colunas adjacentes de uma mesma proteína, é possível verificar a capacidade de cada abordagem amostrar estruturas mais ou menos próximas à nativa. Nota-se que, para todos os casos, as melhores estruturas obtidas configuram pequeno percentual (normalmente < 3 %) das estruturas amostradas, o que mais uma vez destaca a dificuldade do problema que está sendo abordado.

Desta vez, de maneira visual e considerando a quantidade de estruturas amostrada em cada faixa, avaliando especificamente os intervalos de GDT mais altos e de RMSD mais baixos, é possível verificar novamente a melhor capacidade de CuT-REMD obter bons resultados para as proteínas de classe α (códigos PDB 1L2Y, 1RIJ e 1VII). O mesmo não ocorre com as proteínas de classe β (códigos PDB 1UAO, 1E0L e 1LE1), as quais também confirmam visualmente os resultados obtidos anteriormente. Para a classe $\alpha\beta$, CuT-REMD mostra-se como mais efetivo para as proteínas de código PDB 1PSV e 2WXC, e menos efetivo para a proteína de código 1FME. Quanto à uniformidade dos resultados obtidos por meio de diferentes métricas, não foi possível perceber diferenças significativas entre RMSD e GDT-TS.

Embora útil, a análise acima descrita é prejudicada pelo fato da quantidade de estruturas de interesse ser muito baixa, o que prejudica a inspeção visual acerca do comportamento das abordagens em avaliação. Assim sendo, novo estudo foi feito e novos gráficos foram gerados com o intuito de esclarecer, de maneira minimalista, a capacidade de amostragem das abordagens em questão. Para tal, dividiu-se as proteínas presentes no conjunto teste em grupos de acordo com suas classes, analisando-as individualmente tanto em relação a suas faixas de GDT-TS quanto RMSD.

A Figura 7.2 confirma os resultados de seções anteriores, demonstrando a maior capacidade de CuT-REMD para amostrar estruturas próximas à nativa para as proteínas de classe α presentes no conjunto teste de proteínas (ver Tabela 5.2).

Para as proteínas de código PDB 1L2Y e 1RIJ, as melhores estruturas amostradas nas quatro primeiras temperaturas atingiram valores 0,9 a 1,0 GDT-TS, configurando estruturas praticamente idênticas à nativa. Para 1L2Y, CuT-REMD foi capaz de amostrar $\approx$ 70% das estruturas obtidas nessa faixa. Já para 1RIJ, $\approx$ 99% das estruturas nas melhores faixas

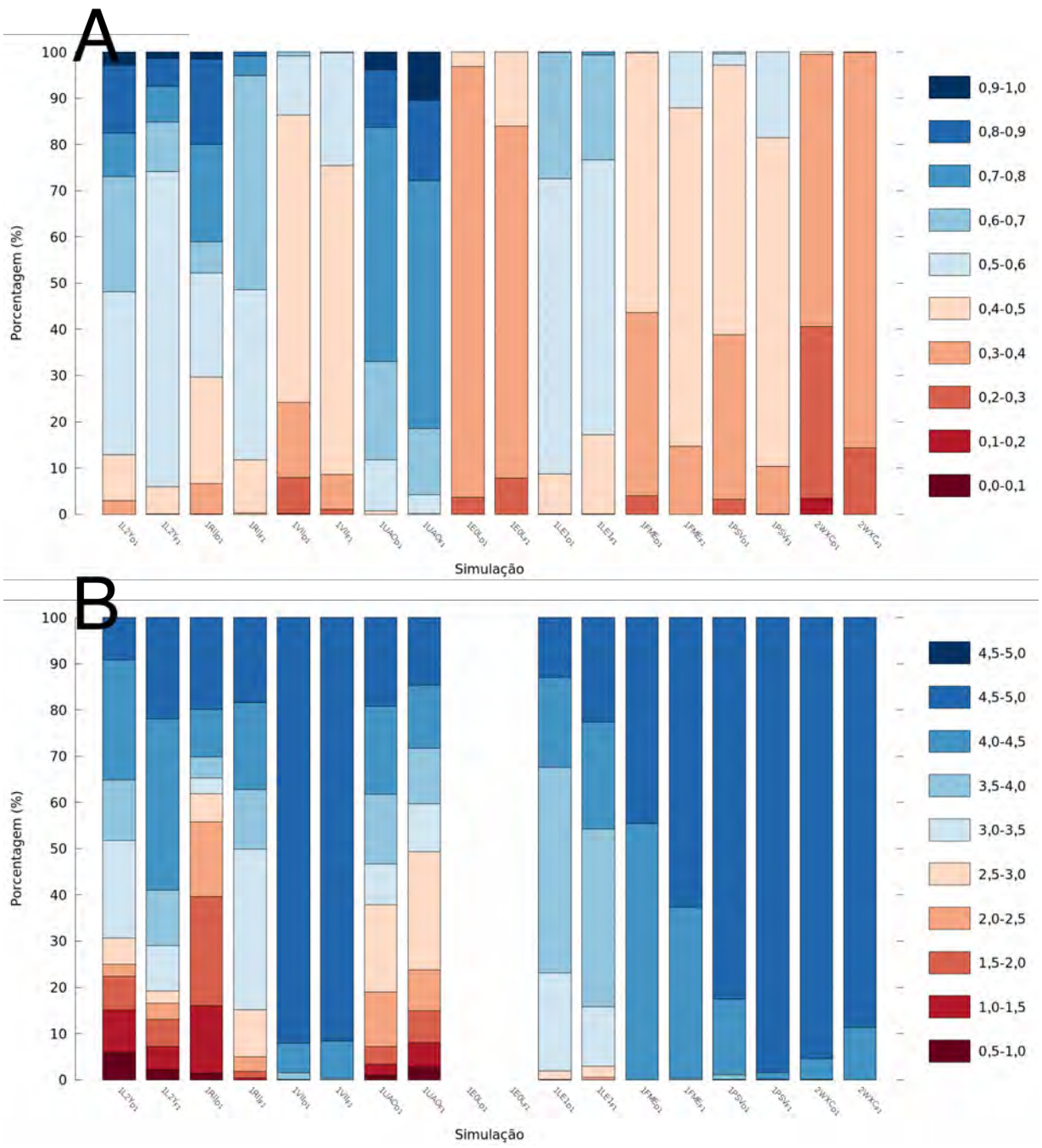

Figura 7.1 – Comparativo de CuT-REMD (D1) e REMD convencional (F1), para todas as proteínas do conjunto teste (estruturas obtidas a temperaturas baixas). A) contém resultados utilizando GDT-TS e B) utilizando RMSD. A proteína 1E0L não retornou RMSDs < 5,0 Å.

de GDT-TS/RMSD foram obtidas por CuT-REMD. Em relação a 1VII, as melhores estruturas situaram-se na faixa de 0,6 a 0,7 de GDT-TS ou 3,0 a 3,5 (Å) de RMSD e, novamente, CuT-REMD foi responsável pela exploração da maioria das estruturas nessas faixas ($\approx$ 80% para GDT-TS e $\approx$ 95% para RMSD).

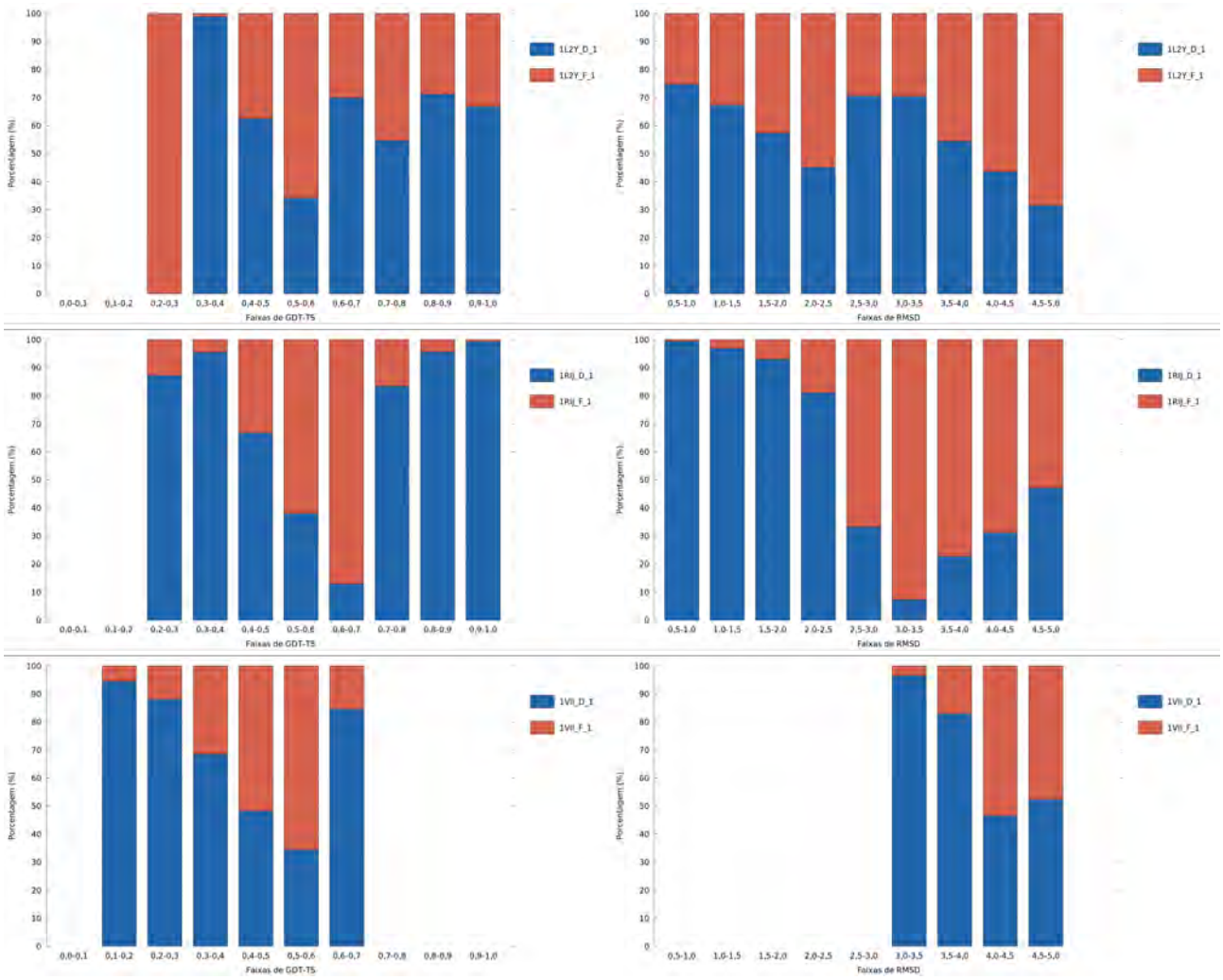

Figura 7.2 – Inspeção minimalista quanto às faixas de RMSD e GDT-TS das estruturas amostradas pelas abordagens CuT-REMD e REMD convencional. Em azul CuT-REMD; e em vermelho, REMD convencional. Resultados para as proteínas de classe α.

Para as proteínas de classe β presentes no conjunto teste de proteínas (ver Tabela 5.2), mais uma vez foi possível confirmar os resultados obtidos em seções anteriores, os quais indicavam que CuT-REMD demonstraria menor capacidade de explorar conformações próximas da nativa, para essa classe de proteínas.

Analisando a Figura 7.3 e os resultados obtidos para a proteína de código PDB 1UAO, verifica-se que ambas as abordagens foram capazes de obter estruturas na faixa de 0,9 a 1,0 GDT-TS, porém a maior parte do número total dessas estruturas foi obtida por REMD convencional ($\approx$ 70%). Para a proteína de código PDB 1LE1, as melhores estruturas obtiveram GDT-TS na faixa de 0,7 a 0,8, o que configura estruturas de alta similaridade (GDT-TS acima de 0,6). No entanto, avaliando-se apenas RMSD, o método convencional de REMD foi capaz de amostrar estruturas na faixa de 1,5 a 2,0 Å, o que não foi possível com CuT-REMD. Para a proteína de código PDB 1E0L, as melhores estruturas obtiveram GDT-TS na faixa de 0,5 a 0.6 utilizando-se REMD convencional e de 0,4 a 0.5 utilizando-se CuT-REMD, e assim sendo, nenhuma das abordagens foi suficientemente hábil para amos-

trar estruturas satisfatórias.

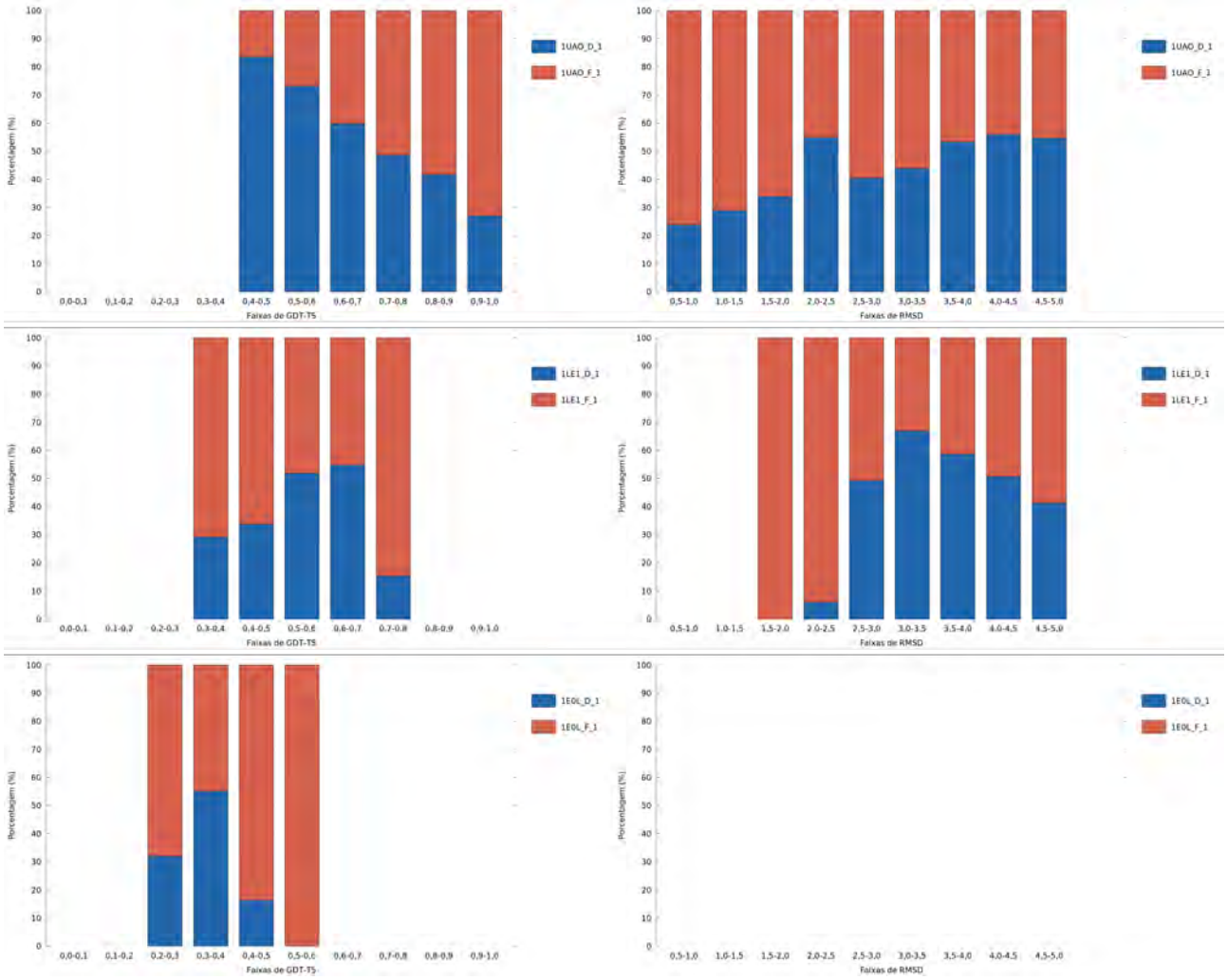

Figura 7.3 – Inspeção minimalista quanto às faixas de RMSD e GDT-TS das estruturas amostradas pelas abordagens CuT-REMD e REMD convencional. Em azul CuT-REMD; e em vermelho, REMD convencional. Resultados para as proteínas de classe β. As simulações da proteína de código PDB 1E0L não amostraram estruturas de RMSD < 5,0 Å.

A Figura 7.4 exibe os resultados referentes às proteínas de classe $\alpha\beta$ presentes no conjunto teste de proteínas (ver Tabela 5.2).

Em seções anteriores, fora constatado que, para as proteínas de código PDB 1PSV e 2WXC, CuT-REMD retornara melhores resultados e, para 1FME, piores. Tal fato foi confirmado por meio da análise minimalista aqui exposta. Atentando-se mais uma vez às trajetórias retornadas pelas simulações e à amostragem de estruturas mais próximas da nativa, CuT-REMD visitou (nas quatro temperaturas mais baixas) 100% das melhores estruturas para as proteínas de código PDB 1PSV e 2WXC, respectivamente com valores de RMSD em torno de 3,0 a 3,5 e 3,5 a 4,0 (Å). Para a proteína de código PDB 1FME, no entanto, embora CuT-REMD tenha sido capaz de amostrar estruturas de GDT-TS > 0,6

(consideradas de alta similaridade) a única simulação capaz de amostrar estruturas abaixo de 3,5 Å foi a simulação por REMD convencional.

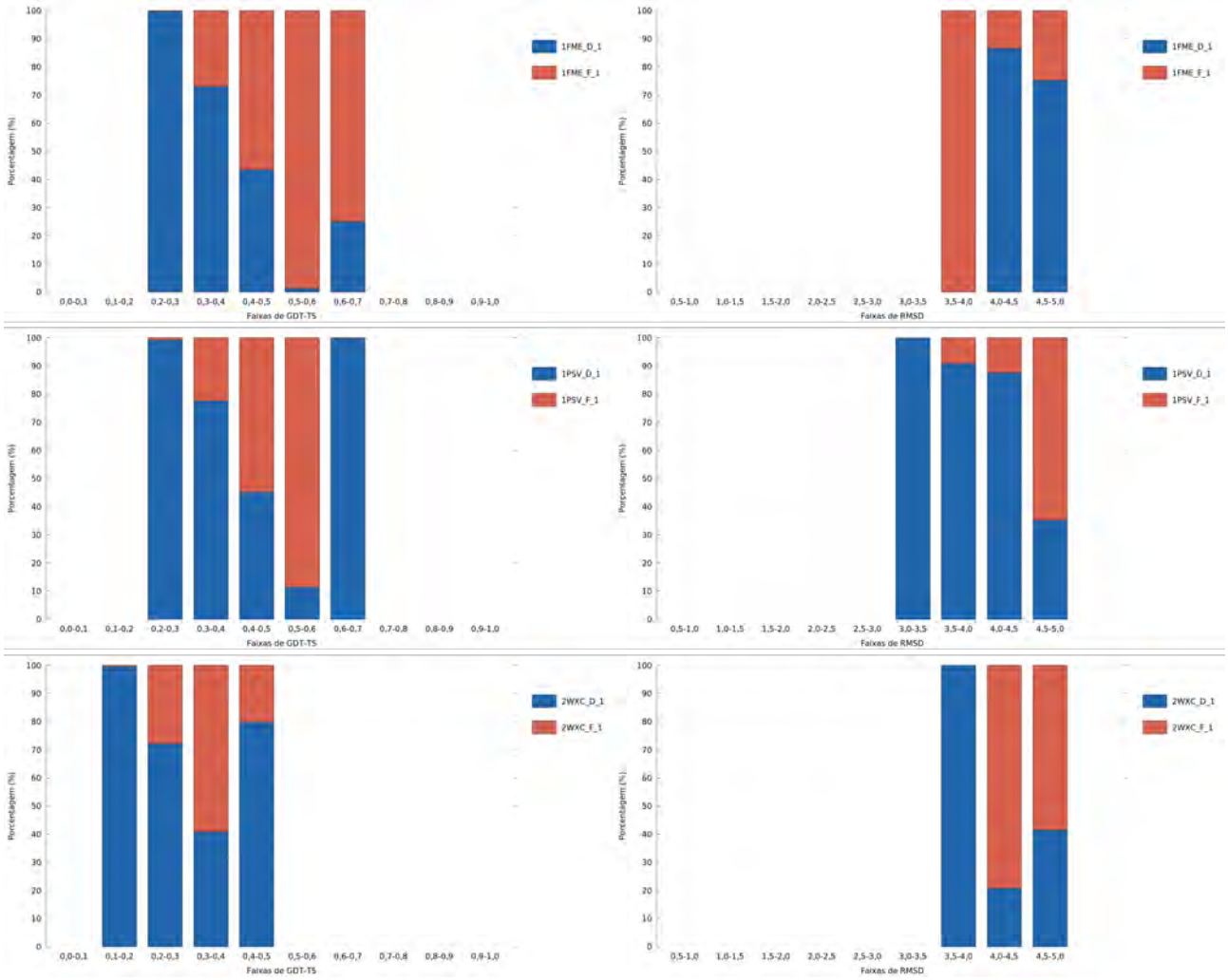

Figura 7.4 – Inspeção minimalista quanto às faixas de RMSD e GDT-TS das estruturas amostradas pelas abordagens CuT-REMD e REMD convencional. Em azul, a abordagem CuT-REMD; e em vermelho, o método REMD convencional. Resultados para as proteínas de classe $\alpha\beta$.

7.1.2 Capacidade Exploratória *BestStruc*

As análises a seguir levam em consideração as trajetórias de todas as réplicas (ou todas as temperaturas) geradas pelas simulações. As Figuras 7.5, 7.6 e 7.7 representam, respectivamente, resultados obtidos para as proteínas de classe α, β e $\alpha\beta$. O objetivo dos gráficos é demonstrar a distribuição (em faixas de GDT-TS) da totalidade de conformações obtidas pelas simulações CuT-REMD e REMD convencional. Tal análise é importante pois possibilita a inspeção visual acerca de em qual(is) temperatura(s) são obtidas as melhores estruturas. Uma vez que se busca uma abordagem que não se utilize de todas as estruturas

(nesse caso 700 mil por simulação), tal informação possui impacto direto nos resultados, já que abordagens como *Best5Pop* consideram apenas as estruturas obtidas a temperaturas mais baixas. Além disso, conforme já destacado no início do capítulo, como *Best5Pop* considera apenas estruturas oriundas de temperaturas mais baixas, existe a possibilidade de uma dada abordagem ser efetiva na amostragem de estruturas próximas à nativa, porém tais estruturas podem estar presentes apenas em temperaturas não contempladas pela metodologia de captura. Tal ocorrência também será analisada nesta subseção.

Considerando-se que as estruturas *BestStruc* - conforme esperado - retornaram sempre estruturas mais próximas da nativa se comparado às estruturas *Best5Pop*, calculou-se a diferença $|BestStruc - Best5Pop|$ a fim de obter-se uma medida capaz de informar o quão longe do ótimo a abordagem de captura de estruturas proposta por este estudo está. Dessa avaliação, foi possível perceber resultados semelhantes, tanto para CuT-REMD quanto para REMD convencional. Para as simulações CuT-REMD, as médias das diferenças entre as diferentes proteínas foram, respectivamente para GDT-TS e RMSD: 0,14 $\pm$ 0,07 e 1,2 $\pm$ 0,5. Considerando apenas REMD convencional e mais uma vez computando a média para todas as proteínas, os valores retornados para GDT-TS e RMSD foram, respectivamente: 0,13 $\pm$ 0,03 e 1,5 $\pm$ 0,7.

Embora os resultados acima demonstrem que, na média, a abordagem de captura foi efetiva, a análise a seguir detalha com mais propriedade o comportamento das abordagens para com as diferentes proteínas testadas.

A Figura 7.5 exibe os resultados para as proteínas de classe α. Para as proteínas de código PDB 1L2Y, 1RIJ e 1VII, as diferenças entre as estruturas *BestStruc* e *Best5Pop*, computadas em GDT-TS foram (em média) 0,06 e 0,09 respectivamente para CuT-REMD e REMD convencional, valores que podem ser considerados baixos. Ao analisar as temperaturas em que as estruturas *BestStruc* foram obtidas, embora existam diferenças entre as abordagens, a estipulação das 4 temperaturas mais baixas como restrição para as trajetórias serem analisadas mostrou-se efetiva, ou seja, não resultou em desperdício relevante de estruturas.

Em adição, cabe ressaltar que, para todas as proteínas dessa classe, os melhores resultados de *Best5Pop* foram obtidos pelas simulações por CuT-REMD. Nota-se ainda que, considerando apenas as temperaturas mais baixas, CuT-REMD foi capaz de concentrar uma maior proporção de estruturas de qualidade satisfatória.

A Figura 7.6 exibe os resultados para as proteínas de classe β. Para as proteínas de código PDB 1UAO, 1LE1 e 1E0L, as diferenças entre as estruturas *BestStruc* e *Best5Pop*, computadas em GDT-TS, foram (em média) 0,21 e 0,17 respectivamente para CuT-REMD e REMD convencional. Tais valores foram os mais altos entre as classes de proteínas testadas. Ao analisar as temperaturas em que as estruturas *BestStruc* foram obtidas, foi necessário analisar individualmente cada proteína.

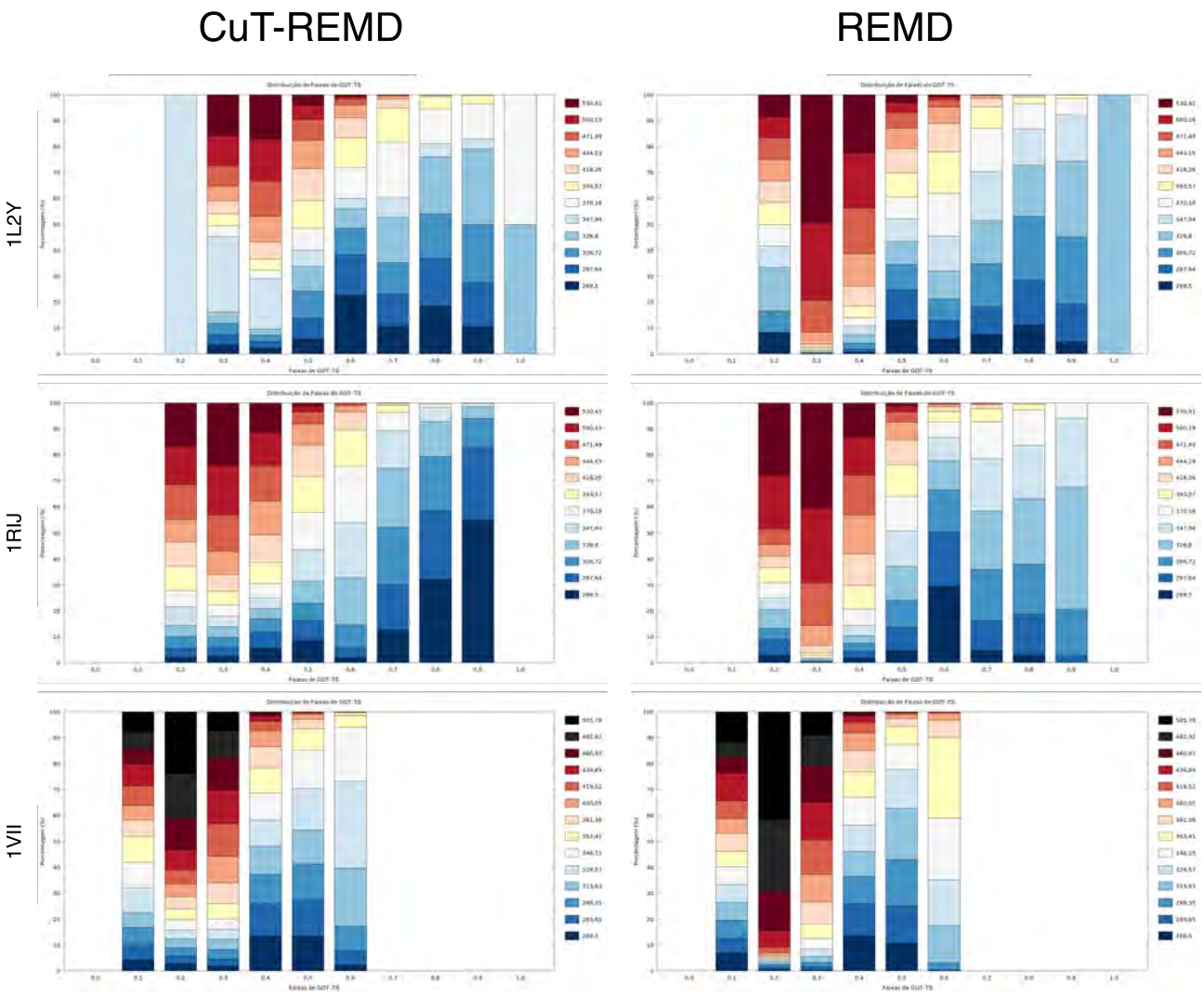

Figura 7.5 – Distribuição em faixas de GDT-TS da totalidade das estruturas obtidas por CuT-REMD e REMD, classificadas por temperatura (em K). Proteínas de classe α.

Para a proteína de código PDB 1UAO, o comportamento de CuT-REMD e REMD foi praticamente idêntico no que se refere aos valores de GDT-TS retornados e, embora os resultados de *Best5Pop* tenham sido cerca de 20% inferiores aos de *BestStruc*, ambas as abordagens consideraram para clusterização as temperaturas em que as melhores estruturas estavam presentes. Assim sendo, a razão das estruturas *Best5Pop* terem menor similaridade com a nativa pode ser entendida como inerente ao processo de clusterização utilizado, mas não relacionado à quantidade de temperaturas utilizada. De todo modo, as estruturas obtidas são relevantes, uma vez que estruturas de GDT-TS $\approx$ 0,8 são consideradas ótimas.

Para a proteína de código PDB 1LE1, o comportamento de CuT-REMD e REMD foi novamente similar, contando no entanto com apenas 3 temperaturas por volta de 0,8 de GDT-TS em CuT-REMD e 4 temperaturas em simulações REMD convencional. No entanto, verificando-se tais temperaturas, nota-se que, em ambas as abordagens, estas são as temperaturas mais altas das simulações, ou seja, as trajetórias e consequentemente as

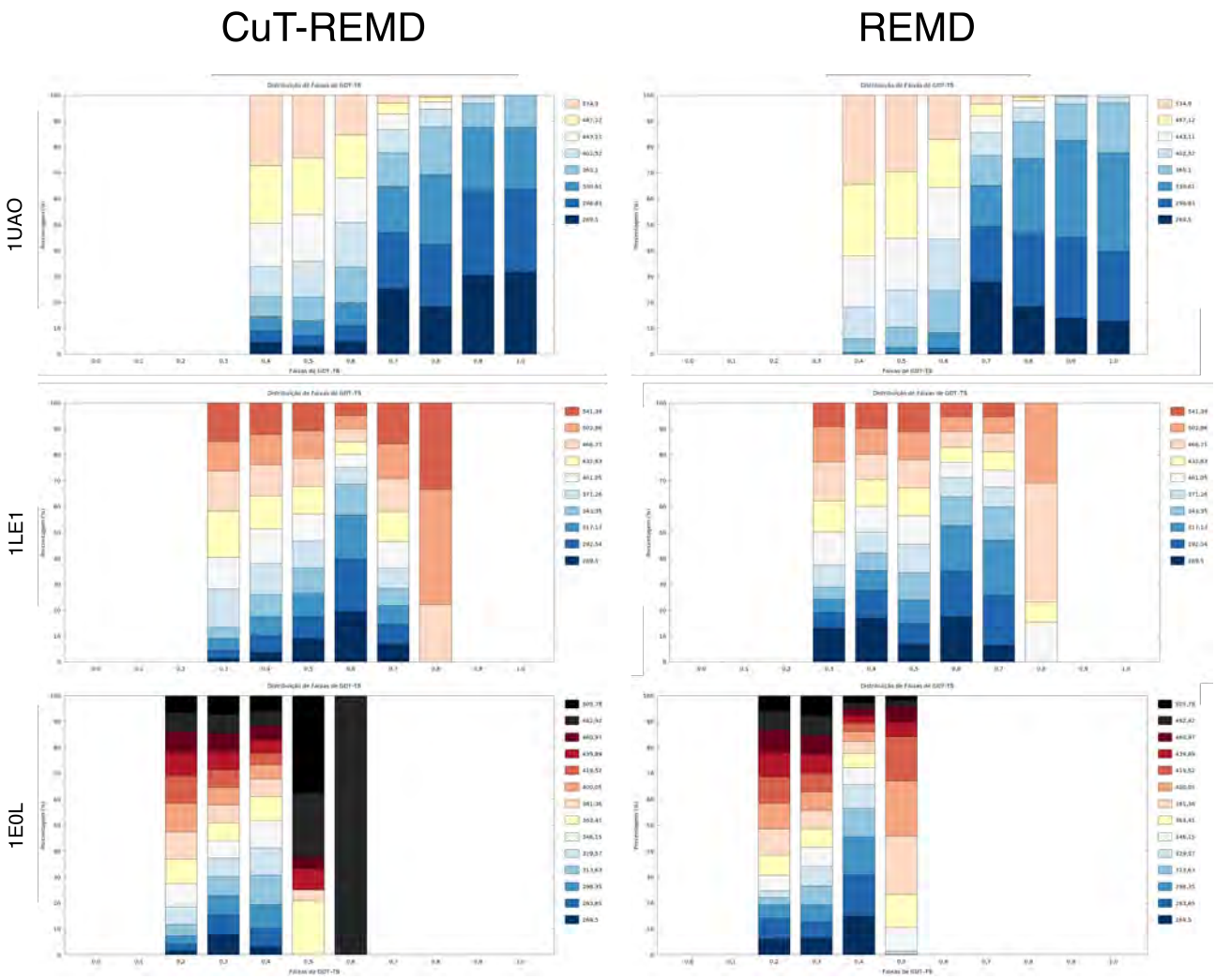

Figura 7.6 – Distribuição em faixas de GDT-TS da totalidade das estruturas obtidas por CuT-REMD e REMD, classificadas por temperatura (em K). Proteínas de classe β.

estruturas geradas nessas temperaturas não são consideradas no processo de clusterização.

Ao avaliar os resultados de *Best5Pop*, estes são percebidos como cerca de 20% piores do que os de *BestStruc*. A razão das estruturas *Best5Pop* terem menor similaridade com a nativa pode então, nesse caso, ser atribuída ao filtro inicial de temperaturas, o que de imediato impede que estruturas por volta dos 0,8 de GDT-TS sejam obtidas por *Best5Pop*.

Para a proteína de código PDB 1E0L, o comportamento do método REMD convencional foi superior, porém retornando valor de *Best5Pop* cerca de apenas 5% melhor. Os resultados de *BestStruc*, no entanto, foram superiores para CuT-REMD. Tal resultado poderia ser entendido como contraditório. Pela Figura 7.6, percebe-se, no entanto, que CuT-REMD foi capaz de amostrar, de fato, estruturas melhores (GDT-TS > 0,6) que o método convencional, porém estas estruturas se situaram em temperaturas elevadas e, por isso, foram negligenciadas no momento da clusterização e captura da estrutura predita. Além disso, as estruturas de qualidade imediatamente inferior a 0,6 de GDT-TS também

se posicionaram em temperaturas mais altas, o que explica a melhor adequação de REMD convencional quando se utiliza a métrica *Best5Pop*.

A Figura 7.7 exibe os resultados para as proteínas de classe $\alpha\beta$. Para as proteínas de código PDB 1FME, 1PSV e 2WXC, as diferenças entre as estruturas *BestStruc* e *Best5Pop*, computadas em GDT-TS foram (em média) 0,14 e 0,13 respectivamente para CuT-REMD e REMD convencional.

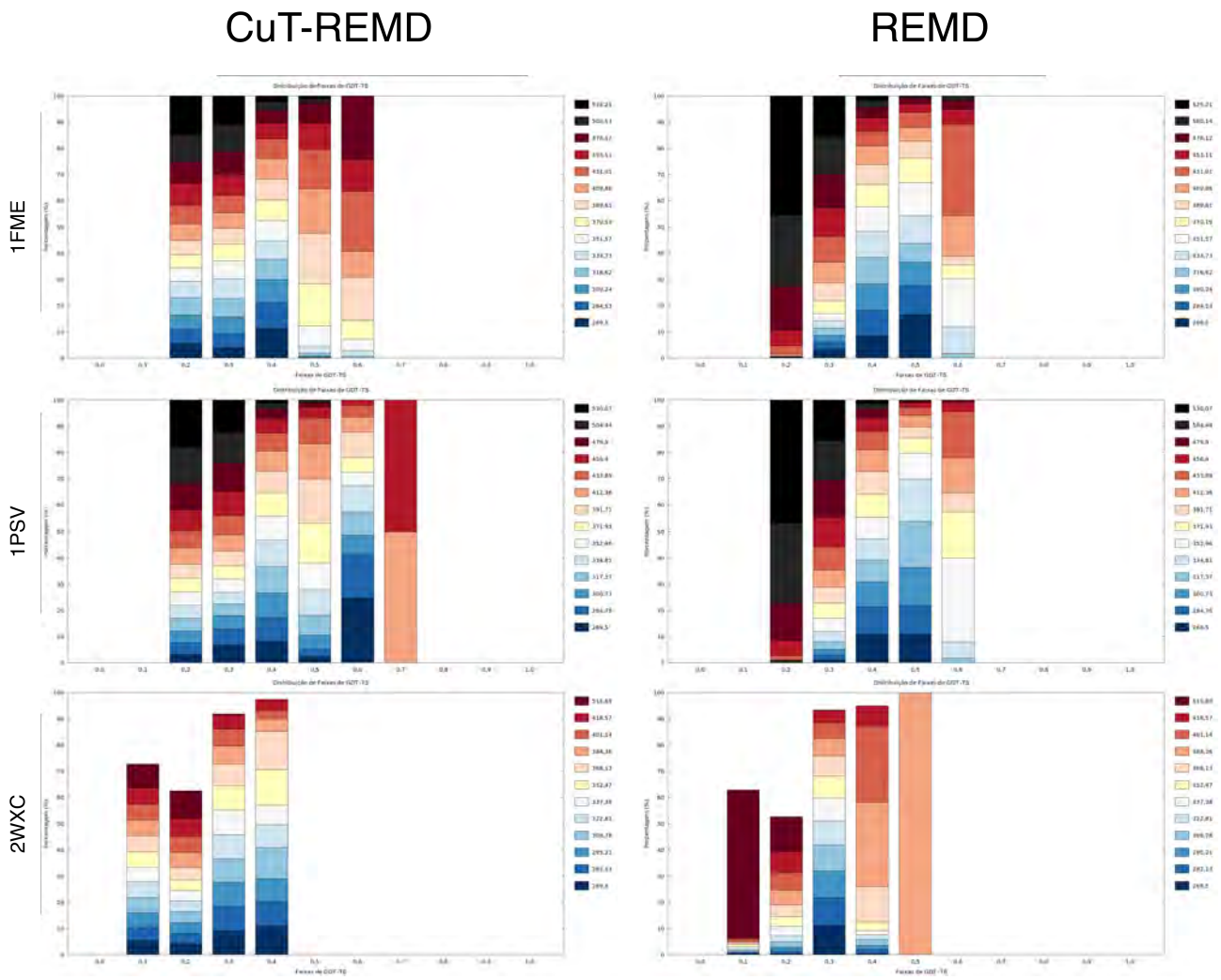

Figura 7.7 – Distribuição em faixas de GDT-TS da totalidade das estruturas obtidas por CuT-REMD e REMD, classificadas por temperatura (em K). Proteínas de classe $\alpha\beta$.

Para a proteína de código PDB 1FME, o comportamento do método REMD convencional foi consideravelmente superior, retornando valores de *Best5Pop* cerca de 10% melhores. Os resultados de *BestStruc*, no entanto, foram similares. Ao se analisar com atenção os resultados, percebe-se que, considerando apenas as temperaturas mais baixas, foi amostrada uma maior proporção de estruturas de GDT-TS mais elevado nas simulações por REMD convencional, o que explica a melhor capacidade da abordagem em retornar valores de *Best5Pop* satisfatórios.

Quanto à proteína de código PDB 1PSV, verifica-se que CuT-REMD foi superior tanto na obtenção de *BestStruc* quanto *Best5Pop*. Além disso, a proporção de estruturas de GDT-TS > 0,6 foi maior para a simulação baseada na abordagem CuT-REMD, sem mencionar estruturas de GDT-TS acima de 0,7, as quais, embora não tenham sido obtidas em temperaturas baixas (e por isso não consideradas quando da clusterização) foram as melhores estruturas amostradas, nível não atingido por REMD convencional.

Avaliando a proteína de código PDB 2WXC, os resultados demonstram que CuT-REMD obteve melhores resultados quanto a *Best5Pop*, enquanto *BestStruct* retornou resultados similares por meio das duas abordagens. Curiosamente, nesse caso, REMD convencional foi capaz de amostrar uma única estrutura de GDT-TS > 0,5, fato não atingido por CuT-REMD. Considerando-se estruturas de GDT-TS > 0,4, no entanto, verifica-se que a proporção destas em temperaturas mais baixas foi consideravelmente mais alta na simulação por CuT-REMD, o que consequentemente conferiu a esse tipo de simulação maior capacidade em encontrar uma estrutura *Best5Pop* de mais alta qualidade.

7.2 CuT-REMD *versus* Literatura

Com o intuito de verificar a capacidade da abordagem aqui proposta em comparação à literatura, CuT-REMD foi comparado com os trabalhos resultantes do mapeamento sistemático (Apêndice A) apresentado nesta tese. Além destes, foram incluídos também na comparação os principais programas de predição de estruturas 3D do CASP: SCRATCH [VW09], BHAGEERATH [JBS+06], QUARK [XZ12] e PEP-FOLD3 [LTR+16].

Vale a ressalva, no entanto, de que os métodos BHAGEERATH, QUARK, SCRATCH e PEP-FOLD3 não são métodos *ab initio*, ou seja, valem-se de informações adicionais acerca da proteína (como bibliotecas de fragmentos de estruturas depositadas em bancos de dados) que não apenas sua estrutura primária (métodos *de novo*).

O servidor BHAGEERATH faz uso de informações de estrutura secundária com o objetivo de gerar modelos que, na sequência, passam por filtros biofísicos e otimização, apresentando ao final um total de 10 modelos para o usuário. Já o método QUARK, um dos métodos de mais destaque na área de PSP, fundamenta-se em dividir a estrutura primária da proteína em pequenos fragmentos que podem variar de 1 a 20 resíduos de aminoácidos. Tais fragmentos são então comparados com uma biblioteca de fragmentos, e mapas de restrições são obtidos. Segue-se então a etapa de criação de modelos iniciais, baseada na união dos fragmentos, e aplica-se REMC, utilizando um campo de força baseado em conhecimento. Quanto ao método SCRATCH, este é baseado no algoritmo 3Dpro, o qual utiliza a predição de estrutura secundária em concomitância a uma biblioteca de fragmentos do PDB, aplicando *Simulated Annealing* para atingir energias mais baixas e estruturas de qualidade. Por fim, tem-se o método PEP-FOLD3, o qual é baseado em um conceito

intitulado alfabeto estrutural e utiliza um Modelo Escondido de Markov ou *Hidden Markov Model* derivado de um alfabeto estrutural de 27 letras para descrever proteínas como uma série de fragmentos sobrepostos (cada um composto por 4 aminoácidos). PEP-FOLD3 é baseado na predição desses fragmentos seguida pela aplicação de um algoritmo guloso, orientado por um campo de força genérico e de *coarse-grained*. Um ponto importante em relação ao método é o fato de ser bem adaptado exatamente para a predição da estrutura de peptídeos e pequenas proteínas, foco deste trabalho. Tal característica coloca-o como passível de maior atenção quanto aos resultados.

Antes de iniciar as comparações, cabe a ressalva de que nem todos os métodos foram aplicados a todas as proteínas do conjunto teste. Enquanto certos grupos de pesquisa estudaram apenas um grupo limitado de proteínas, certos métodos disponibilizados *online* possuem limitações individuais, como por exemplo a quantidade de aminoácidos mínima aceita. BHAGEERATH e SCRATCH não possuem restrições quanto ao tamanho de proteínas, porém o método QUARK, por exemplo, aceita apenas sequências de 20 ou mais resíduos de aminoácidos. PEP-FOLD3, em sua nova versão (janeiro de 2016), trabalha atualmente com proteínas de 5 a 50 resíduos de aminoácidos.

Salienta-se também que algumas proteínas não foram avaliadas em relação a certos métodos pois, embora tais métodos possuíssem servidores divulgados como ativos, alguns deles não estiveram disponíveis para as consultas desta pesquisa. Alguns dos métodos simplesmente não retornaram respostas às submissões realizadas, sendo o contato via correio eletrônico insuficiente para que tal tarefa fosse cumprida, caso do servidor BHAGEERATH e, em parte, do servidor SCRATCH.

As Tabelas 7.2, 7.3, 7.4, 7.2, 7.6, 7.7, 7.8, 7.9 e 7.10 a seguir representam valores de RMSD calculados a partir do mesmo intervalo de resíduos, agrupando os diferentes métodos em tabelas individuais para cada proteína do conjunto de testes. De acordo com a disponibilidade dos dados e quando aplicável, foram adicionados também, nas tabelas, detalhes quanto ao tipo de simulação executado por cada trabalho, como a quantidade de réplicas, o tempo de simulação (ns) e o tipo de solvente empregado.

A proteína de código PDB 1L2Y, conhecida como gaiola de triptofanos ou *tryptophan cage* é uma das mais estudadas em simulações baseadas em DM para a predição de estrutura de proteínas, o que se confirma na quantidade de métodos *ab initio* selecionados como relacionados a esta tese. Dando atenção aos 14 trabalhos listados envolvendo métodos *ab initio*, percebe-se que CuT-REMD, levando em consideração seja a estrutura predita pelos métodos ou as estruturas mais próximas da nativa amostradas (*BestStruc*), foi o método que retornou melhores resultados. Conforme pode ser verificado pela Tabela 7.2, grande parte dos métodos *ab initio* obteve estruturas abaixo de 1,5 Å, porém apenas 3 métodos (o trabalho de Kannan e Zacharias [KZ09a], o trabalho do grupo do professor Carlos Simmerling [NMH+14] e CuT-REMD) foram capazes de atingir estruturas de RMSD < 1,0 Å, o que reforça o desempenho do método aqui apresentado. Quanto aos resulta-

dos obtidos em comparação aos trabalhos envolvendo métodos *de novo*, mais uma vez CuT-REMD mostrou melhor desempenho, desta vez destacando-se significativamente dos demais métodos, já que o melhor método comparado (PEP-FOLD3) atingiu apenas 3,0 Å de RMSD, enquanto CuT-REMD atingiu 0,5 Å.

Tabela 7.2 – Tabela comparativa entre CuT-REMD e os métodos disponíveis na literatura, sejam eles *ab initio* ou *de novo*. Proteína de código PDB 1L2Y. Resultados em Å. Tempo representa o tempo total de simulação (em ns).

CuT-REMD *versus* métodos *ab initio*

Referência	Estrutura Predita	*BestStruc*	# Rép	Tempo	Solv
[Zho04]	-	1,3	50	250	Exp
[DTND08]	-	∼1,5	8	800	Exp
[KZ09a]	0,9	0,4	5	200	Exp
[KZ09b]	1,1	0,4	16	640	Exp
[LO10]	-	0,9	16	1600	Imp
[BPJV11]	-	1,5	8(x2)	320	Imp
[MBFP12]	2,6	2,4	-	-	-
[Fer14]	1,1	1,0	-	1000	Exp
[JW14a]	1,2	0,4	12	13200	Exp
[MJG+14]	1,1	<1,0	12	1920	Imp
[NMH+14]	0,7	0,3	9	540	Exp
[OZ14]	-	<2.0	9	540	Exp
[SKS+15]	∼3,0	<2,0	34(x2)	15648	Imp
CuT-REMD	0,5	0,3	12	600	Imp

CuT-REMD *versus* métodos *de novo*

Referência	Estrutura Predita	*BestStruc*
BHAGEERATH [JBS+06]	-	-
SCRATCH [VW09]	-	-
QUARK [XZ12]	3,5	-
PEP-FOLD3 [LTR+16]	3,0	-
CuT-REMD	0,5	0,3

Para a proteína de código PDB 1RIJ, na comparação entre CuT-REMD e trabalhos com métodos *ab initio*, CuT-REMD foi o de melhor desempenho e atingiu 0,8 Å como resultado de sua estrutura predita, e 0,6 Å como a melhor estrutura amostrada (*BestStruc*). O método que mais se aproximou de CuT-REMD foi o método de Fernandes [Fer14], o qual obteve 1,9 Å e 3,6 Å, respectivamente para a estrutura predita e *BestStruc*. Assim sendo, CuT-REMD demonstrou, para essa proteína, capacidade de diminuir o RMSD para menos da metade do obtido até então. Considerando também os métodos *de novo*, CuT-REMD mostra-se ainda mais promissor, uma vez que os 0.8 Å atingidos configuram uma melhoria de 2,8 Å em relação ao método SCRATCH, o de segundo melhor desempenho.

Tabela 7.3 – Tabela comparativa entre CuT-REMD e os métodos disponíveis na literatura, sejam eles *ab initio* ou *de novo*. Proteína de código PDB 1RIJ. Resultados em Å. Tempo representa o tempo total de simulação (em ns).

CuT-REMD *versus* métodos *de novo*

Referência	Estrutura Predita	*BestStruc*	# Rép	Tempo	Solv
[MBFP12]	4,3	2,4	-	-	-
[Fer14]	1,9	3,6	-	1000	Exp
[SKS+15]	∼4,0	∼3,5	34(x2)	15648	Imp
CuT-REMD	0,8	0,6	12	600	Imp

CuT-REMD *versus* métodos *de novo*

Referência	Estrutura Predita	*BestStruc*
BHAGEERATH [JBS+06]	-	-
SCRATCH [VW09]	3,6	-
QUARK [XZ12]	4,7	-
PEP-FOLD3 [LTR+16]	4,2	-
CuT-REMD	0,8	0,6

A proteína de código PDB 1VII é, assim como 1UNC (estudo de caso), uma *villin headpiece*, porém desta vez presente em galinha, enquanto 1UNC em humanos. Para essa proteína, CuT-REMD obteve resultados de 4,6 Å para a estrutura predita e 2,4 Å para a melhor estrutura amostrada. Verificando os resultados obtidos pelos métodos *ab initio*, verificou-se que o método de melhor resultados foi o método de Nguyen, o qual obteve respectivamente 2,3 Å e 1,1 Å para a estrutura predita e a melhor estrutura amostrada. Assim sendo, CuT-REMD distanciou-se 2,3 Å do melhor resultado obtido na literatura. Ambos os métodos utilizam solvente implícito, porém ao analisar os detalhes da simulação de ambos os métodos, verifica-se que o trabalho de Nguyen (grupo Simmerling), embora se utilize de apenas 8 réplicas para sua simulação por REMD, necessitou de um total de 33.600 ns para atingir seu resultado, um tempo de simulação 48 vezes mais longo que o utilizado por CuT-REMD. Avaliando-se CuT-REMD contra os métodos que utilizam informações de bancos de dados, verificou-se que, embora tenham sido amostradas por CuT-REMD estruturas abaixo de 2.5 Å, o método não foi capaz de retorná-las como a estrutura predita e, assim sendo, foi capaz apenas de atingir melhor desempenho que dois dos métodos *de novo* (BHAGEERATH e SCRATCH). Quanto aos outros dois métodos (QUARK e PEP-FOLD3), estes obtiveram o mesmo resultado: 3,2 Å de RMSD em relação à estrutura nativa, representando uma melhoria de 1,4 Å no RMSD se comparado a CuT-REMD.

A proteína de código PDB 1UAO foi a menor proteína testada e, embora fosse esperado bom desempenho do método, uma vez que estima-se um espaço de busca menor, isso não se confirmou. CuT-REMD não foi eficaz na formação das fitas de folha e, consequentemente, embora tenha amostrado a melhor estrutura (0,4 Å de RMSD) dentre todos

Tabela 7.4 – Tabela comparativa entre CuT-REMD e os métodos disponíveis na literatura, sejam eles *ab initio* ou *de novo*. Proteína de código PDB 1VII. Resultados em Å. Tempo representa o tempo total de simulação (em ns).

CuT-REMD *versus* métodos *ab initio*

Referência	Estrutura Predita	*BestStruc*	# Rép	Tempo	Solv
[JSJ11]	-	2,5	20	30000	Exp
[MBFP12]	2,5	2,2	-	-	-
[Fer14]	6,6	4,2	-	1000	Exp
[NMH$^+$14]	2,3	1,1	8	33600	Imp
CuT-REMD	4,6	2,4	14	700	Imp

CuT-REMD *versus* métodos *de novo*

Referência	Estrutura Predita	*BestStruc*
BHAGEERATH [JBS$^+$06]	5,5	-
SCRATCH [VW09]	5,2	-
QUARK [XZ12]	3,2	-
PEP-FOLD3 [LTR$^+$16]	3,2	-
CuT-REMD	4,6	2,4

os métodos (sejam *ab initio* ou *de novo*), não foi capaz de estabilizar tal estrutura a ponto de explorá-la por mais tempo, o que consequentemente ocasionaria a formação de nichos de estruturas mais acessadas e posteriormente retornaria um *cluster* de boa qualidade no momento da captura da estrutura a ser apresentada como nativa. Avaliando-se apenas os métodos *de novo*, no entanto, percebe-se que a estrutura predita por CuT-REMD (de 2,7 Å de RMSD em relação à nativa) foi apenas 0,4 Å menos similar que a estrutura de 2,3 Å predita pelo servidor PEP-FOLD3, único método *denovo* comparável.

Para a proteína de código PDB 1LE1, quando comparando os resultados de CuT-REMD com os métodos *ab initio* presentes na literatura, verifica-se que este obteve resultado similar aos métodos de melhor desempenho, com predições $\sim$3,0 Å de RMSD em relação à estrutura nativa. Cabe ainda a ressalva de que os resultados de CuT-REMD foram obtidos em 500 ns de simulação, o menor tempo de simulação entre os métodos, com diferença significativa para os demais. Ao comparar-se o desempenho de CuT-REMD com métodos *de novo*, no entanto, CuT-REMD mostra desempenho inferior e, mesmo que o RMSD de 3,3 Å não possa ser considerado ruim, não foi possível superar o método PEP-FOLD3, uma vez que este atingiu estrutura apenas 1,6 Å distante da nativa.

A proteína de código PDB 1E0L mostrou-se uma das mais difíceis para a tarefa da obtenção de sua estrutura terciária. Com exceção do método *ab initio* de Ozkan e colaboradores (RMSD de 2,2 Å) e do método *de novo* PEP-FOLD3 (RMSD de 1,6 Å), os outros seis métodos avaliados (incluindo CuT-REMD) não foram capazes de retornar resultados satisfatórios, tendo suas estruturas preditas em média 6,0 Å de RMSD para a nativa.

Tabela 7.5 – Tabela comparativa entre CuT-REMD e os métodos disponíveis na literatura, sejam eles *ab initio* ou *de novo*. Proteína de código PDB 1UAO. Os símbolos * e e † representam, respectivamente, simulações com solvente explícito e com solvente implícito. Resultados em Å. Tempo representa o tempo total de simulação (em ns).

CuT-REMD *versus* métodos *ab initio*

Referência	Estrutura Predita	*BestStruc*	# Rép	Tempo	Solv
[SPHvdS05]*	-	2,9	16	7360	Exp
[SPHvdS05]†	-	1,0	16	16000	Imp
[KZ07]	-	<1	7	140	Exp
[MBFP12]	0,7	0,6	-	-	-
[Fer14]	0,6	0,6	-	1000	Exp
[OZ14]	$\sim$1.0	$\sim$1.0	9	135	Exp
CuT-REMD	2,7	0,4	8	400	Imp

CuT-REMD *versus* métodos *de novo*

Referência	Estrutura Predita	*BestStruc*
BHAGEERATH [JBS^{+}06]	-	-
SCRATCH [VW09]	-	-
QUARK [XZ12]	-	-
PEP-FOLD3 [LTR^{+}16]	2,3	-
CuT-REMD	2,7	0,4

Tabela 7.6 – Tabela comparativa entre CuT-REMD e os métodos disponíveis na literatura, sejam eles *ab initio* ou *de novo*. Proteína de código PDB 1LE1. Resultados em Å. Tempo representa o tempo total de simulação (em ns).

CuT-REMD *versus* métodos *ab initio*

Referência	Estrutura Predita	*BestStruc*	# Rép	Tempo	Solv
[MBFP12]	4,5	2,2	-	-	-
[Fer14]	2,9	2,9	-	1000	Exp
[SKS^{+}15]	$\sim$3,0	$\sim$1,0	34(x2)	15648	Imp
CuT-REMD	3,3	1,9	10	500	Imp

CuT-REMD *versus* métodos *de novo*

Referência	Estrutura Predita	*BestStruc*
BHAGEERATH [JBS^{+}06]	-	-
SCRATCH [VW09]	-	-
QUARK [XZ12]	-	-
PEP-FOLD3 [LTR^{+}16]	1,6	
CuT-REMD	3,3	1,9

Em relação à proteína de código PDB 1FME, ao comparar-se CuT-REMD e os resultados obtidos por métodos *ab initio* disponíveis na literatura, verifica-se que a estrutura

Tabela 7.7 – Tabela comparativa entre CuT-REMD e os métodos disponíveis na literatura, sejam eles *ab initio* ou *de novo*. Proteína de código PDB 1E0L. Resultados em Å. Tempo representa o tempo total de simulação (em ns).

CuT-REMD *versus* métodos *ab initio*

Referência	Estrutura Predita	*BestStruc*	# Rép	Tempo	Solv
[OWCD07]	2,2	-	-	-	Imp
[MBFP12]	7,6	5,9	-	-	-
[Fer14]	6,5	4,9	-	1000	Exp
CuT-REMD	6,3	5,1	14	700	Imp

CuT-REMD *versus* métodos *de novo*

Referência	Estrutura Predita	*BestStruc*
BHAGEERATH [JBS+06]	-	-
SCRATCH [VW09]	7,5	-
QUARK [XZ12]	4,7	-
PEP-FOLD3 [LTR+16]	1,6	-
CuT-REMD	6,3	5,1

predita por CuT-REMD (de 4,3 Å de RMSD em relação à nativa) foi melhor que 3 dos 5 métodos com os quais foi comparado e pior que 2 métodos. Jiang e Wu obtiveram uma estrutura de 2,7 Å enquanto Perez *et al.* obtiveram uma estrutura de 3,4 Å de distância da estrutura nativa. Fica aqui novamente a ressalva de que os resultados por CuT-REMD foram obtidos em um tempo total de simulação de 700 ns, tempo este extremamente menor (87 vezes) que o tempo requerido por Jiang e Wu para a obtenção de seus resultados. Os métodos *de novo* apresentaram comportamento similar no que se refere aos resultados; enquanto os métodos BHAGEERATH e SCRATCH atingiram resultados considerados inferiores aos de CuT-REMD, os métodos QUARK e PEP-FOLD3 atingiram resultados melhores. Nenhum método *de novo*, no entanto, foi capaz de superar os resultados de Jiang e Wu destacados anteriormente.

Os resultados comparativos para a proteína de código PDB 1PSV compõem a Tabela 7.9. Por meio desta, é possível constatar que, para os métodos *ab initio*, CuT-REMD mostrou-se o de melhor desempenho, embora a estrutura predita de CuT-REMD não tenha RMSD considerado baixo (3,9 Å). Avaliando-se *BestStruc*, a estrutura de CuT-REMD atinge 2,9 Å, o que a configura como resultado satisfatório. Na comparação com os métodos que em adição utilizam informações de bancos de dados, o método QUARK foi o único capaz de superar CuT-REMD, enquanto SCRATCH e PEP-FOLD3 atingiram resultados, no mínimo, 1,6 Å piores.

A proteína de código PDB 2WXC, assim como destacado quando da análise de 1FME, também se apresentou como de difícil predição estrutural porém, mesmo assim, o método CuT-REMD foi, dentre os métodos *ab initio*, aquele que atingiu os melhores resul-

Tabela 7.8 – Tabela comparativa entre CuT-REMD e os métodos disponíveis na literatura, sejam eles *ab initio* ou *de novo*. Proteína de código PDB 1FME. Resultados em Å. Tempo representa o tempo total de simulação (em ns).

CuT-REMD *versus* métodos *ab initio*

Referência	Estrutura Predita	*BestStruc*	# Rép	Tempo	Solv
[MBFP12]	6,8	3,5	-	-	-
[Fer14]	5,3	3,7	-	1000	Exp
[JW14a]	2.7	1,2	36	61200	Exp
[NMH$^+$14]	4,6	0,9	6	54600	Imp
[PMD15]	3,4	2,0	30	15000	Exp
CuT-REMD	4,3	2,8	14	700	Imp

CuT-REMD *versus* métodos *de novo*

Referência	Estrutura Predita	*BestStruc*
BHAGEERATH [JBS$^+$06]	4,9	-
SCRATCH [VW09]	4,9	-
QUARK [XZ12]	3,3	-
PEP-FOLD3 [LTR$^+$16]	3,2	-
CuT-REMD	4,3	2,8

Tabela 7.9 – Tabela comparativa entre CuT-REMD e os métodos disponíveis na literatura, sejam eles *ab initio* ou *de novo*. Proteína de código PDB 1PSV. Resultados em Å. Tempo representa o tempo total de simulação (em ns).

CuT-REMD *versus* métodos *ab initio*

Referência	Estrutura Predita	*BestStruc*	# Rép	Tempo	Solv
[Sue03]	-	3.3 (4.2)	29	87	Imp
[MBFP12]	5,8	3,5	-	-	-
[Fer14]	4,5	4,5	-	1000	Exp
CuT-REMD	3,9	2,9	14	700	Imp

CuT-REMD *versus* métodos *de novo*

Referência	Estrutura Predita	*BestStruc*
BHAGEERATH [JBS$^+$06]	-	-
SCRATCH [VW09]	5,5	-
QUARK [XZ12]	2,4	-
PEP-FOLD3 [LTR$^+$16]	5,8	-
CuT-REMD	3,9	2,9

tados, mesmo que distantes do satisfatório (RMSD de 5,2 Å). Expandindo a análise aos métodos *de novo*, os métodos QUARK e PEP-FOLD3 foram os que demonstraram melhor capacidade de predição para 2WXC, já o método SCRATCH teve desempenho inferior a

CuT-REMD e individualmente ruim, retornando estruturas praticamente sem relação com a nativa (8,3 Å). O único método capaz de predizer uma estrutura abaixo de 3,0 Å, neste caso, foi QUARK.

Tabela 7.10 – Tabela comparativa entre CuT-REMD e os métodos disponíveis na literatura, sejam eles *ab initio* ou *de novo*. Proteína de código PDB 2WXC. Resultados em Å. Tempo representa o tempo total de simulação (em ns).

CuT-REMD *versus* métodos *ab initio*

Referência	Estrutura Predita	*BestStruc*	# Rép	Tempo	Solv
[JW14a]	6.2	3,0	36	54000	Exp
[NMH+14]	8,3	2,1	16	35200	Imp
CuT-REMD	5,2	3,9	14	700	Imp

CuT-REMD *versus* métodos *de novo*

Referência	Estrutura Predita	*BestStruc*
BHAGEERATH [JBS+06]	-	-
SCRATCH [VW09]	8,4	-
QUARK [XZ12]	2,8	-
PEP-FOLD3 [LTR+16]	4,3	-
CuT-REMD	5,2	3,9

8. CONCLUSÕES

Neste capítulo, serão apresentadas as principais conclusões desta pesquisa. Em um primeiro momento, serão expostas as conclusões oriundas dos testes realizados com a proteína de código PDB 1UNC, alvo de estudo de caso exaustivo, e na sequência, os resultados obtidos da aplicação da abordagem CuT-REMD para um conjunto teste de proteínas. Por fim são expostas as limitações da abordagem.

8.1 Estudo de Caso com a Proteína *villin headpiece* de Código PDB 1UNC

Foi proposta a abordagem CuT-REMD para abordar o problema PSP. Este é baseado em simulações por REMD padrão, exceto pelo fato de que, em vez de utilizar um único valor de raio de corte para contabilizar interações não ligadas, introduziu-se o esquema de raio de corte incremental por tempo de execução no qual o raio de corte varia de 4,0 Å a 8,0 Å. A ideia é que os raios de corte mais curtos permitem o início da formação de estruturas locais e o início da estabilização de EES dentro de diferentes segmentos ao longo da cadeia polipeptídica, e à medida que o raio de corte aumenta gradualmente de 4,0 Å para 8,0 Å, os EES de tamanho apropriado são formados, e por sua vez se reúnem para favorecer a formação de estruturas próximas à nativa.

Para o problema do PSP, o uso de uma abordagem com raio de corte incremental resultou em ganhos substanciais em relação às simulações por REMD convencional e DM convencional. Uma vez que a abordagem CuT-REMD não altera o formalismo REMD, a superfície de energia amostrada por seus *ensembles* segue uma distribuição de Boltzmann, o que foi demonstrado sem complicações. Mesmo contando com pouco tempo de simulação (50 ns por réplica), os resultados indicaram que os valores de raio de corte entre 4,0 Å e 8,0 Å geram distribuições de energia que se aproximam bem da curva teórica esperada por abordagens que seguem uma distribuição de Boltzmann.

CuT-REMD é uma abordagem reprodutível em relação ao espaço de energia potencial acessado e amostra energias consideravelmente distintas daquelas amostradas por REMD convencional, destacando o impacto de empregar raio de corte incremental (mesmo que somente no início da simulação). Constatou-se também que, passando mais tempo em raios de corte mais curtos, tem como consequência uma maior sobreposição de energia entre réplicas adjacentes, facilitando a propagação das estruturas em intercâmbios envolvendo diferentes temperaturas. Propõe-se que isso permita uma redução no número de réplicas simuladas, aumentando assim o ganho computacional.

O exame acerca do fenômeno de aprisionamento entre intercâmbios (ETR) demonstrou que menores taxas de aprisionamento são favorecidas sob o esquema de raio de

corte incremental, endossando CuT-REMD como uma forma de proporcionar mais liberdade ao sistema para se mover por entre a rugosa superfície de energia. Embora não tenham sido testados EAFs mais altos do que o valor moderado de 1 ps^{-1}, os resultados obtidos confirmam as constatações de Sindhikara *et al.* [SMR08, SER10], segundo as quais de EAFs mais elevados são esperadas maiores taxas de aprisionamento.

Uma vez que se pretende utilizar CuT-REMD para predizer estruturas 3D de proteínas, sua eficiência na amostragem conformacional é de extrema relevância. Em relação a isso, constatou-se que quanto mais tempo a simulação permaneceu em raios de corte mais curtos, menor foi o tempo médio necessário para completar um evento de tunelamento, caracterizando assim melhor amostragem do espaço conformacional em relação à abordagem convencional. Embora se tenha obtido uma melhor amostragem, o que muitas vezes implica menor taxa de convergência, não foram encontradas evidências de que os raios de corte incrementais fossem um fator prejudicial nesse quesito.

Dando seguimento ao estudo, avaliou-se o número de estruturas que adotaram corretamente EES, bem como o número total de estruturas entendidas como enoveladas. Para todos os protocolos de simulação (o que inclui métodos convencionais), mais uma vez os melhores resultados foram obtidos ao passar mais tempo em raios de corte mais curtos, seja nas simulações baseadas em REMD ou DM. A quantificação do número total de estruturas enoveladas mostrou, no entanto, um padrão diferente. Embora os protocolos CuT-REMD A e C não tenham sido capazes de atingir o mesmo desempenho que REMD convencional, os protocolos CuT-REMD B e D, além de Cu-MD, revelaram capacidade de obtenção de um maior número de estruturas enoveladas, destacando o impacto positivo dos EAEs inferiores sobre o moderado. A avaliação da qualidade relativa das estruturas enoveladas preditas revelou valores de GDT-TS e RMSD iguais a 0,8 e 1,0 Å, respectivamente, demonstrando a capacidade de CuT-REMD para amostrar estruturas quase nativas, especialmente a temperaturas mais baixas.

Ainda em relação a estruturas secundárias, avaliou-se a capacidade de CuT-REMD em comparação a REMD convencional (protocolos D contra F) no que diz respeito à estabilização individual de cada uma das três hélices que formam a proteína alvo do estudo de caso. Desse estudo, verificou-se que, embora para a primeira hélice os resultados não tenham diferido consideravelmente, tanto para a segunda quanto para a terceira hélice CuT-REMD demonstrou maior capacidade em estabilizar as estruturas regulares, conferindo ao sistema inclusive maior difusão das hélices estáveis entre diferentes temperaturas, suportando mais uma vez a ideia de raios de corte mais baixos favorecerem formação de hélices.

Adicionalmente, foram computados os cinco *clusters* mais populosos em termos estruturais e examinados seus centroides (*Best5Pop* e *BestStruc*).

Os protocolos de simulação CuT-REMD B e D exibiram uma melhoria em termos de RMSD de até 1,0 Å em comparação com as estruturas obtidas com REMD convencional. A inspeção das distribuições globais de RMSD revelou excelente capacidade dos protocolos

D, A e C de CuT-REMD, respectivamente, na obtenção de estruturas entre 1,0 e 1,5 Â de RMSD em relação à estrutura de RMN experimental esperada (código PDB 1UNC), uma qualidade estrutural não obtida com os protocolos convencionais.

Ressalta-se o fato de que os raios de corte incrementais foram aplicados somente nos primeiros 5-10 ns de um total de 50 ns de tempo de simulação, e ainda assim, os efeitos mostraram-se dignos de atenção. De acordo com os indicativos, quanto mais a simulação se mantiver em raios de corte mais curtos, mais abrangente é sua capacidade de amostragem do espaço de energia. Em relação à escala de tempo de simulação, trabalhos anteriores [RPES16] utilizaram simulações por REMD em solvente implícito de curta duração apenas como forma de gerar as coordenadas iniciais para simulações com solvente explícito mais longas. Os resultados desta tese demonstram que simulações de curta duração podem, de fato, ser merecedoras de investigações mais detalhadas.

Por fim, a aplicação da metodologia CuT-REMD para melhorar a qualidade e a velocidade (simulações em escala de tempo mais curta) da predição da estrutura 3D de proteínas mostrou-se eficaz. Apresentaram-se resultados de nove protocolos diferentes, incluindo REMD e DM convencionais, e esses resultados foram comparados. Embora CuT-REMD tenha sido testado de forma exaustiva em apenas uma proteína-alvo (código PDB 1UNC), os resultados obtidos instigaram a expansão da aplicabilidade e verificação de comportamento de CuT-REMD para com uma gama maior de proteínas.

8.2 Conjunto Teste de Proteínas

Uma vez tendo CuT-REMD demonstrado ser uma abordagem passível de investigações futuras, aplicou-se o protocolo verificado como de melhor desempenho no estudo de caso (protocolo D) para um conjunto teste de proteínas composto ao todo por 9 proteínas, sendo 3 de cada classe (α, β e $\alpha\beta$).

8.2.1 CuT-REMD *versus* REMD Convencional

Comparando CuT-REMD com REMD convencional e analisando os resultados de *Best5Pop* e *BestStruc* obtidos por ambos os métodos, CuT-REMD se comportou de maneira mais eficaz para as proteínas de classe α (códigos PDB 1L2Y, 1RIJ e 1VII) e $\alpha\beta$ (códigos PDB 1FME, 1PSV e 2WXC), sendo menos eficaz para proteínas de classe β (códigos PDB 1UAO, 1LE1 e 1E0L).

Isto posto, verfica-se, como hipótese a ser tratada futuramente, que o CuT-REMD promove, por meio do diminuto raio de corte inicial a ser incrementado, a formação e estabilização de hélices. Todavia, uma vez que a formação de folhas depende de duas fitas que distantes umas das outras interagem até que se aproximem e se estabilizem, tal interação

é de maior alcanse, o que fica claro se notarmos a incapacidade de CuT-REMD (com raios de corte curtos) estabilizar tais interações.

Considerando-se a métrica RMSD e comparando os valores absolutos atingidos pelas estruturas preditas (*Best5Pop*) por cada método após etapa de clusterização, verificou-se que para sete das nove proteínas ou 78% dos casos CuT-REMD retornou valores melhores, ou seja, atingiu estruturas de resposta mais atrativas. Já observando as melhores estruturas amostradas dentro das simulações ou *BestStruc*, verificou-se que, considerando GDT-TS, para apenas 1 das proteínas (11%) CuT-REMD obteve resultados inferiores aos de REMD convencional.

Verificou-se ainda que apenas um diminuto percentual das estruturas amostradas, sejam as simulações por CuT-REMD ou REMD convencional, são estruturas de qualidade satisfatória (cerca de 3%), o que destaca a dificuldade do problema abordado neste trabalho. Uma vez que o protocolo de captura e apresentação da estrutura predita pelas abordagens utiliza um filtro inicial de temperaturas, verificou-se a possibilidade de, mesmo as abordagens amostrando estruturas de boa qualidade, estas não se fazerem presentes como resultados das predições. Com esse intuito, foram analisadas as distribuições de RMSD e GDT-TS de cada proteína, avaliando mais uma vez o comportamento de cada abordagem e comparando CuT-REMD com REMD convencional.

Para as proteínas de código PDB 1L2Y e 1RIJ, embora ambos as abordagens tenham retornado estruturas de boa qualidade, CuT-REMD destacou-se na obtenção das melhores estruturas, sendo capaz de amostrar $\approx$ 70% das estruturas obtidas na faixa de 0,9 a 1,0 de GDT-TS para 1L2Y e $\approx$99 99% para 1RIJ. Em relação à proteína de código PDB 1VII, as melhores estruturas situaram-se na faixa 3,0 a 3,5 (Å) de RMSD, e novamente, CuT-REMD foi responsável pela exploração da maioria das estruturas nessas faixas ($\approx \approx$ 95%). Tais resultados evidenciam a boa adaptação de CuT-REMD para as proteínas de classse α testadas.

Para as proteínas de classe β presentes no conjunto teste de proteínas, no entanto, mais uma vez foi possível notar os resultados de CuT-REMD como inferiores (pelo menos para 1UAO e 1LE1) aos de REMD convencional. Para a proteína de código PDB 1UAO, verificou-se que ambas as abordagens foram capazes de obter estruturas na faixa de 0,9 a 1,0 GDT-TS, porém a maior parte do número total dessas estruturas foi obtida por REMD convencional ($\approx$ 70%). Para a proteína de código PDB 1LE1, as melhores estruturas obtiveram RMSD na faixa de 1,5 a 2,0 Å, atingidas apenas pelo método convencional de REMD, e não por CuT-REMD. Por fim, para a proteína de código PDB 1E0L, nenhuma abordagem foi capaz de amostrar (e apresentar pós-processo de captura) estruturas de boa qualidade, limitando-se a estruturas com GDT-TS na faixa de 0,5 a 0,6 para REMD convencional e de 0,4 a 0,5 para CuT-REMD.

Em relação à classe $\alpha\beta$, 2 das 3 proteínas, as de código PDB 1PSV e 2WXC, obtiveram em CuT-REMD seus melhores resultados. Já a proteína de código 1FME ob-

teve melhor desempenho sendo simulada por REMD convencional. Da análise minimalista, se constatou que CuT-REMD visitou (nas quatro temperaturas mais baixas) 100% das melhores estruturas para as proteínas de código PDB 1PSV e 2WXC, respectivamente com valores de RMSD em torno de 3,0 a 3,5 e 3,5 a 4,0 (Å). Para a proteína de código PDB 1FME, no entanto, embora CuT-REMD tenha sido capaz de amostrar estruturas de GDT-TS > 0,6 (consideradas de alta similaridade), a única simulação capaz de amostrar estruturas abaixo de 3,5 Å foi a simulação por REMD convencional.

Uma vez elucidada a relação entre a capacidade preditiva de CuT-REMD *versus* REMD convencional, os pontos fortes de CuT-REMD ficaram por conta das estruturas de classes α (melhores resultados em 100% dos casos) e $\alpha\beta$ (melhores resultados em 66% dos casos), já o ponto fraco de CuT-REMD evidenciou-se como as proteínas de classe β, onde embora para uma das proteínas (33%) nenhum método tenha sido hábil o suficiente na descoberta de sua estrutura 3D, o método REMD convencional melhor se adaptou às demais proteínas testadas (66%). Isso posto, passou-se à etapa seguinte de avaliação do método: a comparação com os métodos disponíveis na literatura.

8.2.2 CuT-REMD *versus* Literatura

Em âmbito geral, CuT-REMD foi capaz de, para 4 das 9 proteínas (1L2Y, 1RIJ, 1UAO e 1LE1), chegar a RMSDs abaixo de 3,5 Å, sendo que para 1L2Y e 1RIJ o RMSD foi menor que 1,0 Å. Em relação às 5 proteínas restantes, apenas 1 obteve como retorno uma estrutura predita acima de 6,0 Å (1E0L), ficando as demais em um intervalo de 3,9 Å (1PSV) a 5,2 Å (2WXC).

Na comparação com a literatura, CuT-REMD foi avaliado primeiramente em relação aos métodos que compartilham com ele o fato de serem métodos *ab initio*. Em um segundo momento, uma vez que os resultados foram atrativos, estendeu-se a comparação a métodos *de novo*.

Métodos *ab initio*

Verificando-se não a estrutura predita mas a melhor estrutura visitada pela simulação (o que muitas vezes foi a única informação encontrada na literatura), CuT-REMD foi capaz de, para 7 das 9 proteínas (12Y, 1RIJ, 1VII, 1UAO, 1LE1, 1FME e 1PSV), chegar a estruturas de RMSD abaixo de 3,0 Å, restando apenas as proteínas 1E0L e 2WXC com resultados de RMSD mais altos (5,1 Å e 3,9 Å, respectivamente).

Quanto ao tamanho das proteínas e o desempenho da abordagem, não foi possível verificar padrão que se repetisse, uma vez que para proteínas de até 15 aminoácidos os resultados foram piores que para proteínas de 15 a 25 aminoácidos. Além disso, as proteínas no intervalo de 25 a 30 aminoácidos obtiveram resultados variados se compara-

dos aos obtidos pelas proteínas de até 40 aminoácidos. De todo modo, maior abrangência em termos de proteínas teste é entendida como necessária para que uma análise nesse âmbito seja bem sucedida. De todo modo, vale o destaque de que, mesmo com variações no tamanho das proteínas, as proteínas da classe $\alpha\beta$ foram as que demonstraram maior variação (em média) entre *BestStruc* e *Best5Pop*.

Na comparação com os métodos *ab initio* e considerando tanto a estrutura predita pelos trabalhos quanto a melhor estrutura observada durante as simulações (*BestStruc*), CuT-REMD obteve o melhor resultado dentre todas abordagens para as proteínas de código PDB 1L2Y (sendo o melhor entre 14 métodos avaliados), 1RIJ (sendo o melhor entre 4 métodos avaliados), 1PSV (sendo o melhor entre 4 métodos avaliados) e 2WXC (sendo o melhor entre 3 métodos avaliados). Tais proteínas pertencem, respectivamente, às classes α, α, $\alpha\beta$ e $\alpha\beta$, confirmando resultados anteriores.

Para a proteína de código PDB 1VII (de classe α), CuT-REMD foi o terceiro melhor método dentre 5, tanto na comparação entre estruturas *Best5Pop* quanto *BestStruc*.

Ao avançar para os resultados relativos às proteínas de classe β, diferente do comportamento significativamente deficitário observado em CuT-REMD em relação a REMD convencional, CuT-REMD sendo individualmente comparado à literatura não pode ser considerado uma abordagem ruim pois, embora não tenha se destacado dos demais métodos, seu desempenho foi regular. Para a proteína de código PDB 1UAO, CuT-REMD foi a melhor abordagem dentre 7 na avaliação acerca da estrutura *BestStruc*, entretanto não foi capaz de superar as estruturas preditas de 3 dos métodos, configurando-se nesse quesito apenas como a quarto melhor abordagem. Para a proteína de código PDB 1LE1, 3 dos 4 métodos situaram-se em predições por volta de 3.0 Å, e CuT-REMD faz parte desse rol. Além disso, avaliando-se exaustivamente todas as estruturas amostradas (*BestStruc*), CuT-REMD configurou-se como a segunda melhor abordagem. Para a proteína de código PDB 1E0L, um comportamento incomum foi observado: apenas um único método se mostrou bem adaptado para predizer sua estrutura, o de Ozkan *et al.*, que mesmo sendo o trabalho mais antigo, foi o único a obter estruturas de qualidade satisfatória. Os métodos de Melo *et al.*, Fernandes e o próprio CuT-REMD (ainda que este tenha sido o melhor entre os 3) não foram hábeis o suficiente para acompanhar os resultados de Ozkan e colaboradores.

Por fim, para a proteína de código PDB 1FME e classe $\alpha\beta$, Cut-REMD posicionou-se, dentre os 6 diferentes métodos avaliados, como o terceiro melhor método na comparação entre as estruturas preditas e o quarto na comparação entre as estruturas *BestStruc*.

Em suma, a aplicação do protocolo D para o conjunto teste de proteínas, em comparação aos métodos *ab initio*, foi capaz de demonstrar a boa aptidão de CuT-REMD para predizer as estruturas de proteínas que contenham hélices, sejam estas proteínas da classe α ou $\alpha\beta$. Em contrapartida, verificou-se também CuT-REMD como sendo menos apto a predizer estruturas da classe β. Na comparação direta com outros métodos, CuT-REMD teve bom desempenho, colocando-se na grande maioria das vezes ou como o melhor método

de predição ou com resultados próximos aos melhores métodos, dependendo da proteína estudada.

Tempo de Simulação *versus* Tempo Computacional

Ainda em relação aos métodos *ab initio*, é importante destacar que, dentre todos os métodos envolvendo simulações moleculares expostos nas Tabelas 7.2, 7.3, 7.4, 7.2, 7.6, 7.7, 7.8, 7.9 e 7.10, CuT-REMD é, prioritariamente, aquele que demanda menor tempo de simulação. Comparando o tempo computacional dos métodos, percebeu-se que CuT-REMD chega a ser simulado por até 48 vezes menos tempo que, por exemplo, o trabalho do grupo do professor Carlos Simmerling [NMH+14] ou até 87 vezes menos tempo, na simulação de 1FME, em comparação ao trabalho de Jiang e Wu [JW14a].

Consequentemente, CuT-REMD pode ser entendido não como um método de baixo custo computacional (pelo fato das dinâmicas demandarem bastante esforço computacional), mas como um método que diminui o tempo de simulação necessário quando se tem em vista a obtenção de boas estruturas. A complexidade computacional de CuT-REMD é a mesma que a de REMD convencional.

Além do mais, dada a escassez de recursos disponíveis nas Universidades e Faculdades situadas no Brasil, a diminuição do tempo de simulação coloca-se como de suma importância, uma vez que viabiliza pesquisas mesmo contando com recursos computacionais limitados.

Métodos *de novo*

Uma vez que os resultados de CuT-REMD foram satisfatórios quando comparados com os métodos *ab initio* disponíveis na literatura, optou-se por estender a comparação aos métodos *de novo* disponíveis na literatura, tendo ciência de que estes utilizam informações adicionais, provenientes de bases de dados.

Repetindo o ocorrido com os métodos *ab initio*, para as proteínas de código PDB 1L2Y e 1RIJ (classe α), CuT-REMD obteve o melhor resultado dentre todos métodos, e para a proteína de código PDB 1VII (também de classe α), foi novamente o terceiro melhor método dentre 5, distanciando-se menos de 1,5 Å do melhor resultado.

Para as proteínas de classe β, mais uma vez a maioria dos métodos demonstrou dificuldade na obtenção das estruturas 3D próximas da nativa. Para a proteína de código PDB 1UAO, CuT-REMD posicionou-se como o segundo melhor método (entre apenas dois métodos, porém com uma diferença de apenas 0,4 Å de RMSD), tendo o mesmo desempenho para a proteína de código PDB 1LE1, dessa vez com maior diferença de desempenho (RMSD 1,7 Å mais alto). Quanto à proteína de código PDB 1E0L, o único método hábil o suficiente na obtenção de estruturas 3D satisfatoriamente similares (< 3.5 Å) foi PEP-FOLD,

enquanto QUARK, CuT-REMD e SCRATCH alcançaram apenas estruturas de RMSD > 4.5 Å. Para todas as proteínas dessa classe, o método mais bem adaptado foi PEP-FOLD.

Para as proteínas de classe $\alpha\beta$ de código PDB 1PSV e 2WXC, o método QUARK mostrou capacidade de superar os resultados obtidos pelos métodos *ab initio*, chegando a RMSDs de 2,4 Å e 2,8 Å para 1PSV e 2WXC, respectivamente. Assim sendo, QUARK destacou-se dos demais como melhor método, enquanto CuT-REMD posicionou-se como o segundo melhor para 1PSV e o terceiro melhor para 2WXC. Por fim, para a proteína de código PDB 1FME, CuT-REMD posicionou-se, dentre os 5 diferentes métodos avaliados, como o terceiro melhor na comparação entre as estruturas preditas, a uma diferença de 1,1 Å do melhor método (PEP-FOLD).

Assim sendo, com base nos resultados da comparação de CuT-REMD com a literatura *de novo*, embora tenha encontrado maior dificuldade, CuT-REMD manteve seu bom desempenho, inclusive superando certos servidores (SCRATCH) em todas as ocasiões. Em suma, os resultados obtidos pelo estudo mostram-se encorajadores, e embora muitas descobertas tenham sido feitas, a quantidade de novas perguntas surgidas ao longo da pesquisa foi ainda maior, abrindo espaço para novos desafios e novos trabalhos relacionados.

8.3 Limitações

As limitações da abordagem CuT-REMD são:

1. Limitação quanto ao número de resíduos na cadeia de aminoácidos: Embora não exista uma restrição para o tamanho das proteínas a serem alvo de CuT-REMD, sabe-se que, uma vez que o método utiliza-se de REMD, as simulações necessitarão de uma quantidade muito grande de réplicas afim de se obter o resultado desejado. Com o aumento do número de réplicas, o custo computacional inerente à abordagem cresce também, o que dificulta a aplicação de CuT-REMD a proteínas maiores que 50 resíduos de aminoácidos.

2. Proteínas de Classe β: Segundo os testes proferidos para este trabalho, percebe-se que CuT-REMD promove, por meio do diminuto raio de corte inicial a ser incrementado, a formação e estabilização de hélices. Todavia, uma vez que a formação de folhas β depende de interações de maior alcance, CuT-REMD mostrou-se menos apto a estabilizar tais estruturas.

3. Uma vez que o código do AMBER (até a presente versão) não permite alterações nos códigos que utilizam GPU para realizar as simulações, não é executar simulações com raio de corte reduzido utilizando GPUs, o que limita o desempenho da abordagem.

9. PERSPECTIVAS

O desenvolvimento desta tese resultou no aparecimento de diversas questões de pesquisa a serem exploradas em trabalhos futuros:

- Os resultados obtidos em relação à proteína *villin headpiece* indicam que a aplicação de raio de corte incremental pode não apenas melhorar a capacidade exploratória do sistema, como também possibilitar a utilização de um número menor de réplicas, para a mesma proteína, visto que o grau de difusão de simulações CuT-REMD foi mais alto que o de simulações por REMD convencional. Isso posto, cabe a investigação mais profunda acerca do tema, avaliando-se o impacto da diminuição/aumento do número de réplicas (e consequentemente de temperaturas) no desempenho do sistema, levando em consideração todas as outras variáveis passíveis de parametrização (tempo de permanência em baixos raios de corte, EAF, tempo de simulação, valor de incremento de raio, etc) e também as taxas retornáveis pelas simulações (EAR, ETR, reprodutibilidade, diversidade na amostragem, convergência, etc).

- Uma vez que os resultados para as proteínas contendo hélices foram satisfatórios, pretende-se, de imediato, iniciar estudos com CuT-REMD para proteínas de tamanho superior a 50 resíduos de aminoácidos.

- Outro ponto de destaque é o tempo de permanência em raios de corte mais baixos, uma vez que, nos testes executados durante o período da pesquisa, o tempo de 2 ns demonstrou resultados consideravelmente melhores que o tempo de permanência de 1 ns. Abre-se a hipótese, então, de que tempos de permanência maiores melhorem ainda mais os resultados de CuT-REMD.

- Uma vez que CuT-REMD foi aplicado e testado exaustivamente apenas para a proteína de código PDB 1UNC, da classe α, a replicação de tal etapa utilizando como alvo proteínas de classes diferentes deverá conferir a CuT-REMD maior adaptabilidade quando executado com proteínas da classe β, para as quais CuT-REMD obteve seus piores resultados.

- A modificação do incremento (em Å) no raio de corte das simulações é outro teste a ser executado, o que pode ser feito de maneira facilitada por meio da interface gráfica disponibilizada por esta tese. Acredita-se que tal alteração gere grande impacto nos resultados obtidos, principalmente pelo fato do valor de raio de corte de 4,0 Å ter sido entendido, após análises, como não benéfico para as simulações. Abre-se a possibilidade de simulações iniciando de 4,5 Å levarem a melhores resultados.

- Tendo em vista o impacto positivo do uso de EAFs inferiores em detrimento do uso de EAFs moderados, e percebendo que os EAFs poderiam ser ainda menores, destaca-se também essa possível alteração na busca de uma abordagem mais eficiente.

- A quantidade de temperaturas levada em consideração no momento da clusterização, parte da metodologia de captura e apresentação da estrutura predita e estipulada nesta tese, também demonstrou pontos a serem melhor estudados. Devido à complexidade envolvida, os estudos iniciais desenvolvidos pelo autor da tese resultaram no trabalho de mestrado (já em desenvolvimento) realizado pelo aluno Rafael C. O. Macedo, vinculado ao Programa de Pós-Graduação em Ciência da Computação da Faculdade de Informática (FACIN) da PUCRS.

- A interface gráfica GKT-REMD, embora acate o que consta em um dos objetivos específicos desta tese e forneça ao usuário as funcionalidades necessárias para a configuração dos arquivos de entrada de simulações CuT-REMD (ou REMD) no AMBER, assim como análises, tem potencial para ser mais robusta, considerar diferentes tipos de pacotes de simulação e fornecer maior gama de possíveis análises ao usuário, o que a tornaria significativamente mais atrativa à comunidade que utiliza simulações por REMD. O autor entende que, levando em consideração a complexidade no entendimento das métricas e análises, além da quantidade de dados a serem tratados, o desenvolvimento de uma nova versão de GTK-REMD que enderece tais questões pode ser considerado como trabalho de grande valia aos pesquisadores de PSP e simulação molecular em geral.

- Por fim, estima-se que os avanços em relação à CuT-REMD possibilitem ainda a adaptação da abordagem para ser utilizado com proteínas maiores e, consequentemente, tornar possível a inscrição da abordagem como participante na modalidade *Free Modelling* do próximo CASP.

REFERÊNCIAS BIBLIOGRÁFICAS

[ABG06] Alonso, H.; Bliznyuk, A. A.; Gready, J. E. “Combining docking and molecular dynamic simulations in drug design”, *Medicinal Research Reviews*, vol. 26–5, 2006, pp. 531–568.

[AG08] Abraham, M. J.; Gready, J. E. “Ensuring mixing efficiency of replica-exchange molecular dynamics simulations”, *Journal of Chemical Theory and Computation*, vol. 4–7, 2008, pp. 1119–1128.

[Anf73] Anfinsen, C. B. “Principles that govern the folding of protein chains”, *Science*, vol. 181–96, 1973, pp. 223–230.

[ANZ95] Alexandrov, N. N.; Nussinov, R.; Zimmer, R. M. “Fast protein fold recognition via sequence to structure alignment and contact capacity potentials”, *Pacific Symposium on Biocomputing*, 1995, pp. 53–72.

[AT89] Allen, M. P.; Tildesley, D. J. “Computer simulation of liquids”. New York: Clarendon Press, 1989, 385p.

[AT94] Abagyan, R.; Totrov, M. “Biased probability monte carlo conformational searches and electrostatic calculations for peptides and proteins”, *Journal of Molecular Biology*, vol. 235–3, 1994, pp. 983–1002.

[BAD05] Beck, D. A. C.; Armen, R. S.; Daggett, V. “Cutoff size need not strongly influence molecular dynamics results for solvated polypeptides”, *Biochemistry*, vol. 44–2, 2005, pp. 609–616.

[BAS09] Bahamish, H. A. A.; Abdullah, R.; Salam, R. A. “Protein tertiary structure prediction using artificial bee colony algorithm”. In: 3^{rd} Asia International Conference on Modelling Simulation, 2009, pp. 258–263.

[BBBP09] Bowman, G. R.; Beauchamp, K. A.; Boxer, G.; Pande, V. S. “Progress and challenges in the automated construction of markov state models for full protein systems”, *Journal of Chemical Physics*, vol. 131–12, 2009.

[BBO+83] Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. “Charmm: A program for macromolecular energy, minimization, and dynamics calculations”, *Journal of Computational Chemistry*, vol. 4, 1983, pp. 187–217.

[BBW+14] Biasini, M.; Bienert, S.; Waterhouse, A.; Arnold, K.; Studer, G.; Schmidt, T.; Kiefer, F.; Cassarino, T. G.; Bertoni, M.; Bordoli, L.; Schwede, T. “Swiss-

model: modelling protein tertiary and quaternary structure using evolutionary information", *Nucleic Acids Research*, vol. 42–W1, 2014, pp. W252–W258.

[BDdS13] Brasil, C. R. S.; Delbem, A. C. B.; da Silva, F. L. B. "Multiobjective evolutionary algorithm with many tables for purely ab initio protein structure prediction", *Journal of Computational Chemistry*, vol. 34–20, 2013, pp. 1719–1734.

[Beu11] Beu, T. A. "Molecular dynamics simulations of ion transport through carbon nanotubes", *The Journal of Chemical Physics*, vol. 135–4, 2011, pp. 445–453.

[Bha43] Bhattacharyya, A. "On a measure of divergence between two statistical populations defined by their probability distributions", *Bulletin of the Calcutta Mathematical Society*, vol. 35, 1943, pp. 99–109.

[BHR+14] Bergonzo, C.; Henriksen, N. M.; Roe, D. R.; Swails, J. M.; Roitberg, A. E.; Cheatham, T. E. "Multidimensional replica exchange molecular dynamics yields a converged ensemble of an rna tetranucleotide", *Journal of Chemical Theory and Computation*, vol. 10–1, 2014, pp. 492–499.

[BJKK13] Blaszczyk, M.; Jamroz, M.; Kmiecik, S.; Kolinski, A. "Cabs-fold: server for the de novo and consensus-based prediction of protein structure", *Nucleic Acids Research*, vol. 41–W1, 2013, pp. W406–W411.

[BLE91] Bowie, J. U.; Luthy, R.; Eisenberg, D. "A method to identify protein sequences that fold into a known three-dimensional structure", *Science*, vol. 253–5016, 1991, pp. 164–170.

[BN92] Berg, B. A.; Neuhaus, T. "Multicanonical ensemble: A new approach to simulate first-order phase transitions", *Physical Review Letters*, vol. 68, 1992, pp. 9–12.

[BPBP12] Bramucci, E.; Paiardini, A.; Bossa, F.; Pascarella, S. "Pymod: sequence similarity searches, multiple sequence-structure alignments, and homology modeling within pymol", *BMC Bioinformatics*, vol. 13–4, 2012, pp. 1–6.

[BPJV11] Balaraman, G. S.; Park, I. H.; Jain, A.; Vaidehi, N. "Folding of small proteins using constrained molecular dynamics", *Journal of Physical Chemistry B*, vol. 115–23, 2011, pp. 7588–7596.

[Bro02] Brooks, C. L. "Protein and peptide folding explored with molecular simulations", *Accounts of Chemical Research*, vol. 35–6, 2002, pp. 447–454.

[BSBNDS07] Breda, A.; Santos, D. S.; Basso, L. A.; Norberto De Souza, O. "Ab initio 3-d structure prediction of an artificially designed three-α-helix bundle via all-atom

molecular dynamics simulations", *Genetics and Molecular Research*, vol. 6–4, 2007, pp. 901–910.

[BSVI07] Brenner, P.; Sweet, C. R.; VonHandorf, D.; Izaguirre, J. A. "Accelerating the replica exchange method through an efficient all-pairs exchange", *The Journal of Chemical Physics*, vol. 126–7, 2007, pp. 074103.

[BWD07] Beck, D. A. C.; White, G. W. N.; Daggett, V. "Exploring the energy landscape of protein folding using replica-exchange and conventional molecular dynamics simulations", *Journal of Structural Biology*, vol. 157–3, 2007, pp. 514–523.

[BWF+00] Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. "The protein data bank", *Nucleic Acids Research*, vol. 28–1, 2000, pp. 235–242.

[CAH+10] Chen, V. B.; Arendall, III, W. B.; Headd, J. J.; Keedy, D. A.; Immormino, R. M.; Kapral, G. J.; Murray, L. W.; Richardson, J. S.; Richardson, D. C. "*MolProbity*: all-atom structure validation for macromolecular crystallography", *Acta Crystallographica Section D*, vol. 66–1, 2010, pp. 12–21.

[Cal05] Calvo, F. "All-exchanges parallel tempering", *The Journal of Chemical Physics*, vol. 123–12, 2005, pp. 124106.

[CBB+14] Case, D. A.; Babin, V.; Berryman, J. T.; Betz, R. M.; Cai, Q.; Cerutti, D. S.; Cheatham, T. E.; Darden, T. A.; Duke, R. E.; Gohlke, H.; Goetz, A. W.; Gusarov, S.; Homeyer, N.; Janowski, P.; Kaus, J.; Kolossváry, I.; Kovalenko, A.; Lee, T. S.; LeGrand, S.; Luchko, T.; Luo, R.; Madej, B.; Merz, K. M.; Paesani, F.; Roe, D. R.; Roitberg, A.; Sagui, C.; Salomon-Ferrer, R.; Seabra, G.; Simmerling, C. L.; Smith, W.; Swails, J.; Walker; Wang, J.; Wolf, R. M.; Wu, X.; Kollman, P. A. "Amber 14". San Francisco: University of California Press, 2014.

[CBD10] Custódio, F. L.; Barbosa, H. J. C.; Dardenne, L. E. "Full-atom ab initio protein structure prediction with a genetic algorithm using a similarity-based surrogate model". In: IEEE Congress on Evolutionary Computation, CEC 2010, Barcelona, Spain, 18-23 July 2010, 2010, pp. 1–8.

[CCB+95] Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz Jr, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. "A second generation force field for the simulation of proteins, nucleic acids, and organic molecules", *Journal of the American Chemical Society*, vol. 117–19, 1995, pp. 5179–5197.

[CCID+05] Case, D. A.; Cheatham Iii, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz Jr, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. “The amber biomolecular simulation programs”, *Journal of Computational Chemistry*, vol. 26–16, 2005, pp. 1668–1688.

[CCOS06] Chinchio, M.; Czaplewski, C.; Ołdziej, S.; Scheraga, H. A. “A hierarchical multiscale approach to protein structure prediction: Production of low-resolution packing arrangements of helices and refinement of the best models with a united-residue force field”, *Multiscale Modeling and Simulation*, vol. 5–4, 2006, pp. 1175–1195.

[CFT03] Chikenji, G.; Fujitsuka, Y.; Takada, S. “A reversible fragment assembly method for de novo protein structure prediction”, *The Journal of Chemical Physics*, vol. 119–13, 2003, pp. 6895.

[CGP+98] Crescenzi, P.; Goldman, D.; Papadimitriou, C.; Piccolboni, A.; Yannakakis, M. “On the complexity of protein folding”, *Journal of Computational Biology*, vol. 5, 1998, pp. 597–603.

[CHB+05] Christen, M.; Hünenberger, P. H.; Bakowies, D.; Baron, R.; Bürgi, R.; Geerke, D. P.; Heinz, T. N.; Kastenholz, M. A.; Kräutler, V.; Oostenbrink, C.; Peter, C.; Trzesniak, D.; Van Gunsteren, W. F. “The gromos software for biomolecular simulation: Gromos05”, *Journal of Computational Chemistry*, vol. 26–16, 2005, pp. 1719–1751.

[CHLL03] Chou, C. I.; Han, R. S.; Lee, T. K.; Li, S. P. “A Guided Monte Carlo Approach to Optimization Problems”. Berlin: Springer, 2003, pp. 447–451.

[CKML+16] Clark, K.; Karsch-Mizrachi, I.; Lipman, D. J.; Ostell, J.; Sayers, E. W. “Genbank”, *Nucleic Acids Research*, vol. 44–Database issue, 2016, pp. D67–D72.

[CKP+11] Cong, Q.; Kinch, L. N.; Pei, J.; Shi, S.; Grishin, V. N.; Li, W.; Grishin, N. V. “An automatic method for casp9 free modeling structure prediction assessment”, *Bioinformatics*, vol. 27–24, 2011, pp. 3371.

[Cle08] Clementi, C. “Coarse-grained models of protein folding: toy models or predictive tools?”, *Current Opinion in Structural Biology*, vol. 18–1, 2008, pp. 10–15.

[CLXD03] Chowdhury, S.; Lee, M. C.; Xiong, G.; Duan, Y. “Ab initio folding simulation of the trp-cage mini-protein approaches nmr resolution”, *Journal of Molecular Biology*, vol. 327–3, 2003, pp. 711–717.

[CM06] Colombo, G.; Micheletti, C. “Protein folding simulations: combining coarse-grained models and all-atom molecular dynamics”, *Theoretical Chemistry Accounts*, vol. 116–1-3, 2006, pp. 75–86.

[CRBB03] Chivian, D.; Robertson, T.; Bonneau, R.; Baker, D. “Ab initio methods”, *Methods of Biochemical Analysis*, vol. 44, 2003, pp. 547–557.

[CRSB05] Cheng, J.; Randall, A. Z.; Sweredoski, M. J.; Baldi, P. “Scratch: a protein structure and structural feature prediction server”, *Nucleic Acids Research*, vol. 33–Web Server issue, 2005, pp. W72–6.

[CS11] Chodera, J. D.; Shirts, M. R. “Replica exchange and expanded ensemble simulations as gibbs sampling: Simple improvements for enhanced mixing”, *The Journal of Chemical Physics*, vol. 135–19, 2011, pp. 194110.

[CSS01] Cochran, A. G.; Skelton, N. J.; Starovasnik, M. A. “Tryptophan zippers: Stable, monomeric β-hairpins”, *Proceedings of the National Academy of Sciences*, vol. 98–10, 2001, pp. 5578–5583.

[CTTM03] Carnevali, P.; Tóth, G.; Toubassi, G.; Meshkat, S. N. “Fast protein structure prediction using monte carlo simulations with modal moves”, *Journal of the American Chemical Society*, vol. 125–47, 2003, pp. 14244–14245.

[DA92] Dandekar, T.; Argos, P. “Potential of genetic algorithms in protein folding and protein engineering simulations”, *Protein Engineering*, vol. 5–7, 1992, pp. 637–645.

[Dal12] Dall’Agno, K. C. d. M. “Um estudo sobre a predição da estrutura 3d aproximada de proteínas utilizando o método cref com refinamento”, Dissertação de Mestrado, Programa de Pós-Graduação em Ciência da Computação, PUCRS, 2012, 132p.

[DBL13] Dorn, M.; Buriol, L. S.; Lamb, L. C. “A molecular dynamics and knowledge-based computational strategy to predict native-like structures of polypeptides”, *Expert Systems with Applications*, vol. 40–2, 2013, pp. 698–706.

[Der99] Derreumaux, P. “From polypeptide sequences to structures using monte carlo simulations and an optimized potential”, *Journal of Chemical Physics*, vol. 111–5, 1999, pp. 2301–2310.

[DeSBL14] Dorn, M.; e Silva, M. B.; Buriol, L. S.; Lamb, L. C. “Three-dimensional protein structure prediction: Methods and computational strategies”, *Computational Biology and Chemistry*, vol. 53, Part B–0, 2014, pp. 251–276.

[DGJ+99] Daura, X.; Gademann, K.; Jaun, B.; Seebach, D.; VanGunsteren, W. F.; Mark, A. E. “Peptide folding: when simulation meets experiment”, *Angewandte Chemie, International Edition*, vol. 38–1/2, 1999, pp. 236–240.

[DK98] Duan, Y.; Kollman, P. A. “Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution”, *Science*, vol. 282–5389, 1998, pp. 740–744.

[DK01] Duan, Y.; Kollman, P. A. “Computational protein folding: from lattice to all-atom”, *IBM Systems Journal*, vol. 40, 2001, pp. 297–309.

[DKAA+11] Darbandi, M.; Khaledi-Alidusti, R.; Abbaspour, M.; Abbasi, H. R.; Schneider, G. “Study of cut-off radius and temperature effects on water molecular behavior using molecular dynamics method c3”. In: 9^{th} International Conference on Nanochannels, Microchannels, and Minichannels, ICNMM 2011, 2011, pp. 277–282.

[DM12] Dill, K. A.; MacCallum, J. L. “The protein-folding problem, 50 years on”, *Science*, vol. 338–6110, 2012, pp. 1042–1046.

[DNdS10a] Dorn, M.; Norberto de Souza, O. “A3n: An artificial neural network n-gram-based method to approximate 3-d polypeptides structure prediction”, *Expert Systems with Applications*, vol. 37–12, 2010, pp. 7497–7508.

[DNdS10b] Dorn, M.; Norberto de Souza, O. “Mining the protein data bank with cref to predict approximate 3-d structures of polypeptides”, *International Journal of Data Mining and Bioinformatics*, vol. 4–3, 2010, pp. 281–299.

[DSM97] Dahiyat, B. I.; Sarisky, C. A.; Mayo, S. L. “De novo protein design: towards fully automated sequence selection”, *Journal of Molecular Biology*, vol. 273–4, 1997, pp. 789 – 796.

[DTND08] Ding, F.; Tsao, D.; Nie, H.; Dokholyan, N. V. “Ab initio folding of proteins with all-atom discrete molecular dynamics”, *Structure*, vol. 16–7, 2008, pp. 1010–1018.

[DYP98] Darden, T.; York, D.; Pedersen, L. “Particle mesh ewald: An nlog(n) method for ewald sums in large systems”, *The Journal of Chemical Physics*, vol. 98–12, 1998, pp. 10089–10092.

[EG14] English, C. A.; García, A. E. “Folding and unfolding thermodynamics of the tc10b trp-cage miniprotein”, *Physical Chemistry Chemical Physics*, vol. 16–7, 2014, pp. 2748–2757.

[EHLSW02] Eastwood, M. P.; Hardin, C.; Luthey-Schulten, Z.; Wolynes, P. G. “Statistical mechanical refinement of protein structure prediction schemes: Cumulant expansion approach”, *Journal of Chemical Physics*, vol. 117–9, 2002, pp. 4602–4615.

[Fei85] Feistel, R. “Application of the monte carlo method in statistical physics”, *Journal of Applied Mathematics and Mechanics*, vol. 65–10, 1985, pp. 521–521.

[Fer14] Fernandes, T. V. A. “Desenvolvimento e aplicação de métodos computacionais para predição de estrutura de proteínas”, Tese de Doutorado, Instituto de Biofísica Carlos Chagas Filho, UFRJ, 2014, 229p.

[FFM+06] Floudas, C. A.; Fung, H. K.; McAllister, S. R.; Mönnigmann, M.; Rajgaria, R. “Advances in protein structure prediction and de novo protein design: A review”, *Chemical Engineering Science*, vol. 61–3, 2006, pp. 966–988.

[Fie07] Field, M. J. “A Practical Introduction to the Simulation of Molecular Systems”. Cambridge: Cambridge University Press, 2007, 2 ed., 344p.

[Flo07] Floudas, C. A. “Computational methods in protein structure prediction”, *Biotechnology and Bioengineering*, vol. 97–2, 2007, pp. 207–213.

[FPW10] Fonseca, R.; Paluszewski, M.; Winter, P. “Protein structure prediction using bee colony optimization metaheuristic”, *Journal of Mathematical Modelling and Algorithms*, vol. 9–2, 2010, pp. 181–194.

[Fra93] Fraenkel, A. S. “Complexity of protein folding”, *Bulletin of Mathematical Biology*, vol. 55–6, 1993, pp. 1199–1210.

[FSW91] Frauenfelder, H.; Sligar, S.; Wolynes, P. “The energy landscapes and motions of proteins”, *Science*, vol. 254–5038, 1991, pp. 1598–1603.

[FWT02] Fukunishi, H.; Watanabe, O.; Takada, S. “On the hamiltonian replica exchange method for efficient sampling of biomolecular systems: Application to protein structure prediction”, *Journal of Chemical Physics*, vol. 116–20, 2002, pp. 9058–9067.

[GCS01] Gibbs, N.; Clarke, A. R.; Sessions, R. B. “Ab initio protein structure prediction using physicochemical potentials and a simplified off-lattice model”, *Proteins: Structure, Function and Genetics*, vol. 43–2, 2001, pp. 186–202.

[GHK00] Gohlke, H.; Hendlich, M.; Klebe, G. “Knowledge-based scoring function to predict protein-ligand interactions”, *Journal of Molecular Biology*, vol. 295–2, 2000, pp. 337–356.

[GHP06] Garcia, A. E.; Herce, H.; Paschek, D. “Simulations of temperature and pressure unfolding of peptides and proteins with replica exchange molecular dynamics”, *Annual Reports in Computational Chemistry*, vol. 2, 2006, pp. 83–95.

[GKH05] Gront, D.; Kolinski, A.; Hansmann, U. H. E. “Protein structure prediction by tempering spatial constraints”, *Journal of Computer-Aided Molecular Design*, vol. 19–8, 2005, pp. 603–608.

[GKKG14] Gniewek, P.; Kolinski, A.; Kloczkowski, A.; Gront, D. “Bioshell-threading: versatile monte carlo package for protein 3d threading”, *BMC Bioinformatics*, vol. 15, 2014, pp. 22–22.

[GPW+03] Ginalski, K.; Pas, J.; Wyrwicz, L. S.; von Grotthuss, M.; Bujnicki, J. M.; Rychlewski, L. “Orfeus: Detection of distant homology using sequence profiles and predicted secondary structure”, *Nucleic Acids Research*, vol. 31–13, 2003, pp. 3804–3807.

[GWX+12] Götz, A. W.; Williamson, M. J.; Xu, D.; Poole, D.; Le Grand, S.; Walker, R. C. “Routine microsecond molecular dynamics simulations with amber on gpus. 1. generalized born”, *Journal of Chemical Theory and Computation*, vol. 8–5, 2012, pp. 1542–1555.

[Gü04] Güntert, P. “Automated nmr structure calculation with cyana”, *Methods in Molecular Biology*, vol. 278, 2004, pp. 353–378.

[HACD05] Hoque, M. T.; Andl. Chetty, M.; Dooley, S. “A new guided genetic algorithm for 2d hydrophobic-hydrophilic model to predict protein folding”, *Evolutionary Computation*, vol. 1, 2005, pp. 259–266.

[Han97] Hansmann, U. H. E. “Parallel tempering algorithm for conformational studies of biological molecules”, *Chemical Physics Letters*, vol. 281–1-3, 1997, pp. 140–150.

[HAO+06] Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. “Comparison of multiple amber force fields and development of improved protein backbone parameters”, *Proteins: Structure, Function and Genetics*, vol. 65–3, 2006, pp. 712–725.

[HBE13] Henry, E. R.; Best, R. B.; Eaton, W. A. “Comparing a simple theoretical model for protein folding with all-atom molecular dynamics simulations”, *Proceedings of the National Academy of Sciences*, vol. 110–44, 2013, pp. 17880–17885.

[HCT95] Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. “Pairwise solute descreening of solute charges from a dielectric medium”, *Chemical Physics Letters*, vol. 246–1, 1995, pp. 122–129.

[HCT96] Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. “Parametrized models of aqueous free energies of solvation based on pairwise descreening of solute atomic charges from a dielectric medium”, *The Journal of Physical Chemistry*, vol. 100–51, 1996, pp. 19824–19839.

[HD06] Ho, B. K.; Dill, K. A. “Folding very short peptides using molecular dynamics”, *PLoS Computational Biology*, vol. 2–4, 2006.

[HDS96] Humphrey, W.; Dalke, A.; Schulten, K. “Vmd: Visual molecular dynamics”, *Journal of Molecular Graphics*, vol. 14–1, 1996, pp. 33–38.

[Hee86] Heermann, D. W. “Computer simulation methods: in theoretical physics”. Heidelberg: Springer-Verlag, 1986, 2 ed., 148p.

[HI97] Hart, W. E.; Istrail, S. “Robust proofs of np-hardness for protein folding: General lattices and energy potentials”, *Journal of Computational Biology*, vol. 4–1, 1997, pp. 1–22.

[HLS+09] Hegler, J. A.; Lätzer, J.; Shehu, A.; Clementi, C.; Wolynes, P. G. “Restriction versus guidance in protein structure prediction”, *Proceedings of the National Academy of Sciences*, vol. 106–36, 2009, pp. 15302–15307.

[HMW02] Herges, T.; Merlitz, H.; Wenzel, W. “Stochastic optimisation methods for biomolecular structure prediction”, *Journal of the Association for Laboratory Automation*, vol. 7–3, 2002, pp. 98–104.

[HN96] Hukushima, K.; Nemoto, K. “Exchange monte carlo method and application to spin glass simulations”, *Journal of the Physical Society of Japan*, vol. 65–6, 1996, pp. 1604.

[Hop16] Hopkins, W. G. “A new view of statistics”. Capturado em: http://www.sportsci.org/resource/stats/, Jan 2016.

[HPLS02] Hardin, C.; Pogorelov, T. V.; Luthey-Schulten, Z. “Ab initio protein structure prediction”, *Current Opinion in Structural Biology*, vol. 12–2, 2002, pp. 176–181.

[HS99] Hao, M. H.; Scheraga, H. A. “Designing potential energy functions for protein folding”, *Current Opinion in Structural Biology*, vol. 9–2, 1999, pp. 184–188.

[HSD14] Hatch, H. W.; Stillinger, F. H.; Debenedetti, P. G. “Computational study of the stability of the miniprotein trp-cage, the gb1 β-hairpin, and the ak16 peptide, under negative pressure”, *Journal of Physical Chemistry B*, vol. 118–28, 2014, pp. 7761–7769.

[HVKS14] Hoffmann, F.; Vancea, I.; Kamat, S. G.; Strodel, B. “Protein structure prediction: Assembly of secondary structure elements by basin-hopping”, *ChemPhysChem*, vol. 15–15, 2014, pp. 3378–3390.

[HYSM04] Honda, S.; Yamasaki, K.; Sawada, Y.; Morii, H. “10 residue folded peptide designed by segment statistics”, *Structure*, vol. 12–8, 2004, pp. 1507–1518.

[IOP96] IOPScience. “New dimensions in simulation”, *Physics World*, vol. 9–7, 1996, pp. 29.

[JBS+06] Jayaram, B.; Bhushan, K.; Shenoy, S. R.; Narang, P.; Bose, S.; Agrawal, P.; Sahu, D.; Pandey, V. “Bhageerath: An energy based web enabled computer software suite for limiting the search space of tertiary structures of small globular proteins”, *Nucleic Acids Research*, vol. 34–21, 2006, pp. 6195–6204.

[Jef97] Jeffrey, G. A. “An Introduction to Hydrogen Bonding”. New York: Oxford University Press, 1997, 303p.

[JK13] Jamroz, M.; Kolinski, A. “Clusco: clustering and comparison of protein models”, *BMC Bioinformatics*, vol. 14–1, 2013, pp. 62.

[JMTR96] Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. “Development and testing of the opls all-atom force field on conformational energetics and properties of organic liquids”, *Journal of the American Chemical Society*, vol. 118–45, 1996, pp. 11225–11236.

[Jon99] Jones, D. T. “Protein secondary structure prediction based on position-specific scoring matrices”, *Journal of Molecular Biology*, vol. 292–2, 1999, pp. 195–202.

[Jon01] Jones, D. T. “Predicting novel protein folds by using fragfold”, *Proteins: Structure, Function and Genetics*, vol. 45–SUPPL. 5, 2001, pp. 127–132.

[JRL+05] Jaroszewski, L.; Rychlewski, L.; Li, Z.; Li, W.; Godzik, A. “Ffas03: A server for profile-profile sequence alignments”, *Nucleic Acids Research*, vol. 33–SUPPL. 2, 2005, pp. W284–W288.

[JSJ11] Jani, V.; Sonavane, U. B.; Joshi, R. “Microsecond scale replica exchange molecular dynamic simulation of villin headpiece: an insight into the folding landscape”, *Journal of Biomolecular Structure and Dynamics*, vol. 28–6, 2011, pp. 845–60.

[JSJ14] Jani, V.; Sonavane, U. B.; Joshi, R. “Remd and umbrella sampling simulations to probe the energy barrier of the folding pathways of engrailed homeodomain”, *Journal of Molecular Modeling*, vol. 20–6, 2014, pp. 2283.

[JTT92] Jones, D. T.; Taylor, W. R.; Thornton, J. M. “A new approach to protein fold recognition”, *Nature*, vol. 358–6381, 1992, pp. 86–89.

[JW14a] Jiang, F.; Wu, Y.-D. “Folding of fourteen small proteins with a residue-specific force field and replica-exchange molecular dynamics”, *Journal of the American Chemical Society*, vol. 136–27, 2014, pp. 9536–9539.

[JW14b] Jiang, F.; Wu, Y. D. “Folding of fourteen small proteins with a residue-specific force field and replica-exchange molecular dynamics”, *Journal of the American Chemical Society*, vol. 136–27, 2014, pp. 9536–9.

[JWS08] Jagielska, A.; Wroblewska, L.; Skolnick, J. “Protein model refinement using an optimized physics-based all-atom force field”, *Proceedings of the National Academy of Sciences*, vol. 105–24, 2008, pp. 8268–8273.

[KDN+04] Krieger, E.; Darden, T.; Nabuurs, S. B.; Finkelstein, A.; Vriend, G. “Making optimal use of empirical energy functions:force-field parameterization in crystal space”, *Proteins: Structure, Function and Genetics*, vol. 57–4, 2004, pp. 678–683.

[KF03] Klepeis, J. L.; Floudas, C. A. “Astro-fold: A combinatorial and global optimization framework for ab initio prediction of three-dimensional structures of proteins from the amino acid sequence”, *Biophysical Journal*, vol. 85–4, 2003, pp. 2119–2146.

[KFNH08] Katagiri, D.; Fuji, H.; Neya, S.; Hoshino, T. “Ab initio protein structure prediction with force field parameters derived from water-phase quantum chemical calculation”, *Journal of Computational Chemistry*, vol. 29–12, 2008, pp. 1930–1944.

[KK99] Keseru, G.; Kolossvary, I. “Molecular Mechanics and Conformational Analysis inDrug Design”. Oxford: Wiley, 1999, 176p.

[KK05] Kone, A.; Kofke, D. A. “Selection of temperature intervals for parallel-tempering simulations”, *The Journal of Chemical Physics*, vol. 122–20, 2005, pp. 206101.

[Kof02] Kofke, D. A. “On the acceptance probability of replica-exchange monte carlo trials”, *The Journal of Chemical Physics*, vol. 117–15, 2002, pp. 6911–6914.

[Kol04] Kolinski, A. “Reduced models of proteins and their applications”, *Polymer*, vol. 45–2, 2004, pp. 511–524.

[KS95] Koppensteiner, W. A.; Sippl, M. J. “Knowledge-based potentials-back to the roots”, *Biochemistry*, vol. 63, 1995, pp. 247.

[KS09] Kelley, L. A.; Sternberg, M. J. “Protein structure prediction on the web: a case study using the phyre server”, *Nature Protocols*, vol. 4–3, 2009, pp. 363–371.

[KSB+99] Kalé, L.; Skeel, R.; Bhandarkar, M.; Brunner, R.; Gursoy, A.; Krawetz, N.; Phillips, J.; Shinozaki, A.; Varadarajan, K.; Schulten, K. “Namd2: Greater scalability for parallel molecular dynamics”, *Journal of Computational Physics*, vol. 151–1, 1999, pp. 283–312.

[KSJ10] Koulgi, S.; Sonavane, U.; Joshi, R. “Insights into the folding pathway of the engrailed homeodomain protein using replica exchange molecular dynamics simulations”, *Journal of Molecular Graphics* & *Modelling*, vol. 29–3, 2010, pp. 481–491.

[KWW+12] Källberg, M.; Wang, H.; Wang, S.; Peng, J.; Wang, Z.; Lu, H.; Xu, J. “Template-based protein structure modeling using the raptorx web server”, *Nature Protocols*, vol. 7–8, 2012, pp. 1511–1522.

[KZ07] Kannan, S.; Zacharias, M. “Enhanced sampling of peptide and protein conformations using replica exchange simulations with a peptide backbone biasing-potential”, *Proteins: Structure, Function, and Bioinformatics*, vol. 66–3, 2007, pp. 697–706.

[KZ09a] Kannan, S.; Zacharias, M. “Folding of trp-cage mini protein using temperature and biasing potential replica-exchange molecular dynamics simulations”, *International Journal of Molecular Sciences*, vol. 10–3, 2009, pp. 1121–1137.

[KZ09b] Kannan, S.; Zacharias, M. “Folding simulations of trp-cage mini protein in explicit solvent using biasing potential replica-exchange molecular dynamics simulations”, *Proteins: Structure, Function and Bioinformatics*, vol. 76–2, 2009, pp. 448–460.

[KZ10] Kannan, S.; Zacharias, M. “Application of biasing-potential replicaexchange simulations for loop modeling and refinement of proteins in explicit solvent”, *Proteins: Structure, Function and Bioinformatics*, vol. 78–13, 2010, pp. 2809–2819.

[LAW+12] Lindert, S.; Alexander, N.; Wötzel, N.; Karakaş, M.; Stewart, P. L.; Meiler, J. “Ab initio protein modeling into cryoem density maps using em-fold”, *Biopolymers*, vol. 97–9, 2012, pp. 669–677.

[LB02] Liu, Y.; Beveridge, D. L. “Exploratory studies of ab initio protein structure prediction: Multiple copy simulated annealing, amber energy functions, and a generalized born/solvent accessibility solvation model”, *Proteins: Structure, Function and Genetics*, vol. 46–1, 2002, pp. 128–146.

[LC76] Levitt, M.; Chothia, C. “Structural patterns in globular proteins”, *Nature*, vol. 261–5561, 1976, pp. 552–558.

[LDK01] Lee, M. R.; Duan, Y.; Kollman, P. A. “State of the art in studying protein folding and protein structure prediction using molecular dynamics methods”, *Journal of Molecular Graphics and Modelling*, vol. 19–1, 2001, pp. 146–149.

[LDM08] Lu, M.; Dousis, A. D.; Ma, J. “Opus-psp: An orientation-dependent statistical all-atom potential derived from side-chain packing”, *Journal of Molecular Biology*, vol. 376–1, 2008, pp. 288 – 301.

[Les00] Lesk, A. M. “Introduction to Protein Architecture: The Structural Biology of Proteins”. New York: Oxford University Press, 2000, 1 ed., 147p.

[Les08] Lesk, A. M. “Introduction to bioinformatics”. New York: Oxford University Press, 2008, 3 ed., 474p.

[Lev68] Levinthal, C. “Are there pathways for protein folding?”, *Journal of Medical Physics*, vol. 65–1, 1968, pp. 44–45.

[LGMJ93] Le Grand, S. M.; Merz Jr, K. M. “The application of the genetic algorithm to the minimization of potential energy functions”, *Journal of Global Optimization*, vol. 3–1, 1993, pp. 49–66.

[LHZB06] Liu, P.; Huang, X.; Zhou, R.; Berne, B. J. “Hydrophobic aided replica exchange: an efficient algorithm for protein folding in explicit solvent”, *The Journal of Physical Chemistry B*, vol. 110–38, 2006, pp. 19018–19022.

[LK00] Lazaridis, T.; Karplus, M. “Effective energy functions for protein structure prediction”, *Current Opinion in Structural Biology*, vol. 10–2, 2000, pp. 139–145.

[LKJK04] Lee, J.; Kim, S. Y.; Joo, K.; Kim, I. “Prediction of protein tertiary structure using profesy, a novel method based on fragment assembly and conformational space annealing”, *Proteins: Structure, Function and Genetics*, vol. 56–4, 2004, pp. 704–714.

[LKS05] Liwo, A.; Khalili, M.; Scheraga, H. A. “Ab initio simulations of protein-folding pathways by molecular dynamics with the united-residue model of polypeptide chains”, *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102–7, 2005, pp. 2362–2367.

[LLA+04] Liu, Y.; Liu, Z.; Androphy, E.; Chen, J.; Baleja, J. D. “Design and characterization of helical peptides that inhibit the e6 protein of papillomavirus”, *Biochemistry*, vol. 43–23, 2004, pp. 7421–7431.

[LLFB09] Lindorff-Larsen, K.; Ferkinghoff-Borg, J. “Similarity measures for protein ensembles”, *PLoS ONE*, vol. 4–1, 2009, pp. e4203.

[LLPDS11] Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E. “How fast-folding proteins fold”, *Science*, vol. 334–6055, 2011, pp. 517–520.

[LM14] Lyras, D. P.; Metzler, D. “Reformalign: improved multiple sequence alignments using a profile-based meta-alignment approach”, *BMC Bioinformatics*, vol. 15–1, 2014, pp. 265.

[LMMT93] Laskowski, R. A.; Macarthur, M. W.; Moss, D. S.; Thornton, J. M. “Procheck: a program to check the stereochemical quality of protein structures”, *Journal of Applied Crystallography*, vol. 26, 1993, pp. 283–291.

[LMSVV92] Lyubartsev, A. P.; Martsinovski, A. A.; Shevkunov, S. V.; Vorontsov-Velyaminov, P. N. “New approach to monte carlo calculation of the free energy: Method of expanded ensembles”, *The Journal of Chemical Physics*, vol. 96–3, 1992, pp. 1776.

[LNC08] Lehninger, A.; Nelson, D. L.; Cox, M. M. “Lehninger Principles of Biochemistry”. New York: W. H. Freeman, 2008, 5 ed., 1328p.

[LO10] Lee, M. S.; Olson, M. A. “Protein folding simulations combining self-guided langevin dynamics and temperature-based replica exchange”, *Journal of Chemical Theory and Computation*, 2010.

[LPNdS12] Lipinski-Paes, T.; Norberto de Souza, O. “Cooperative multi-agent system for protein structure prediction”. In: 8^{th} International Conference of the Brazilian Association for Bioinformatics and Computational Biology, 2012, pp. 117.

[LPNdS14] Lipinski-Paes, T.; Norberto de Souza, O. “Masters: A general sequence-based multiagent system for protein tertiary structure prediction”, *Electronic Notes in Theoretical Computer Science*, vol. 306, 2014, pp. 45–59.

[LRO07] Lee, D.; Redfern, O.; Orengo, C. “Predicting protein function from sequence and structure”, *Nature Reviews Molecular Cell Biology*, vol. 8–12, 2007, pp. 995–1005.

[LSW+09] Lindert, S.; Staritzbichler, R.; Wötzel, N.; Karakaş, M.; Stewart, P. L.; Meiler, J. “Em-fold: De novo folding of α/-helical proteins guided by intermediate-resolution electron microscopy density maps”, *Structure*, vol. 17–7, 2009, pp. 990–1003.

[LTBK01] Lee, M. R.; Tsai, J.; Baker, D.; Kollman, P. A. “Molecular dynamics in the endgame of protein structure prediction”, *Journal of Molecular Biology*, vol. 313–2, 2001, pp. 417–430.

[LTR+16] Lamiable, A.; Thevenet, P.; Rey, J.; Vavrusa, M.; Derreumaux, P.; Tuffery, P. “Pep-fold3: faster de novo structure prediction for linear peptides in solution and in complex”, *Nucleic Acids Research*, vol. 44–W1, 2016, pp. W449–54.

[LWLD07] Lei, H.; Wu, C.; Liu, H.; Duan, Y. “Folding free-energy landscape of villin headpiece subdomain from molecular dynamics simulations”, *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104–12, 2007, pp. 4925–4930.

[LWW+08] Lei, H.; Wu, C.; Wang, Z.-X.; Zhou, Y.; Duan, Y. “Folding processes of the b domain of protein a to the native state observed in all-atom ab initio folding simulations”, *The Journal of Chemical Physics*, vol. 128–23, 2008.

[LWWD09] Lei, H.; Wang, Z. X.; Wu, C.; Duan, Y. “Dual folding pathways of an α/β protein from all-atom ab initio folding simulations”, *Journal of Chemical Physics*, vol. 131–16, 2009.

[MBFP12] Melo, M. C. R.; Bernardi, R. C.; Fernandes, T. V. A.; Pascutti, P. G. “Gsafold: A new application of gsa to protein structure prediction”, *Proteins: Structure, Function and Bioinformatics*, vol. 80–9, 2012, pp. 2305–2310.

[MBN+98] MacKerell, A. D.; Brooks, C. L.; Nilsson, L.; Roux, B.; Won, Y.; Karplus, M. “CHARMM: The Energy Function and Its Parameterization with an Overview of the Program”. New York: Wiley, 1998, vol. 1, pp. 271–277.

[MD99] Manousiouthakis, V. I.; Deem, M. W. “Strict detailed balance is unnecessary in monte carlo simulation”, *The Journal of Chemical Physics*, vol. 110–6, 1999, pp. 2753–2756.

[MDK+99] Mohanty, D.; Dominy, B. N.; Kolinski, A.; Brooks Iii, C. L.; Skolnick, J. “Correlation between knowledge-based and detailed atomic potentials: Application to the unfolding of the gcn4 leucine zipper”, *Proteins: Structure, Function and Genetics*, vol. 35–4, 1999, pp. 447–452.

[MGCO00] Macias, M. J.; Gervais, V.; Civera, C.; Oschkinat, H. “Structural analysis of ww domains and design of a ww prototype”, *Nature Structural & Molecular Biology*, vol. 7–5, 2000, pp. 375–379.

[MGK77] McCammon, J. A.; Gelin, B. R.; Karplus, M. “Dynamics of folded proteins”, *Nature*, vol. 267–5612, 1977, pp. 585–590.

[MHS12] Marks, D. S.; Hopf, T. A.; Sander, C. “Protein structure prediction from sequence variation”, *Nature Biotechnology*, vol. 30–11, 2012, pp. 1072–1080.

[MJG+14] Mou, L.; Jia, X.; Gao, Y.; Li, Y.; Zhang, J. Z. H.; Mei, Y. “Folding simulation of trp-cage utilizing a new amber compatible force field with coupled main chain torsions”, *Journal of Theoretical and Computational Chemistry*, vol. 13–4, 2014, pp. 1450026.

[MMBS75] Momany, F. A.; McGuire, R. F.; Burgess, A. W.; Scheraga, H. A. “Energy parameters in polypepltides. vii. geometric parameters, partial atomic charges, nonbonded interactions, hydrogen bond interactions, and intrinsic torsional potentials for the naturally occurring amino acids”, *Journal of Physical Chemistry*, vol. 79–22, 1975, pp. 2361–2381.

[MMK97] McKnight, C. J.; Matsudaira, P. T.; Kim, P. S. “Nmr structure of the 35-residue villin headpiece subdomain”, *Nature Structural & Molecular Biology*, vol. 4–3, 1997, pp. 180–184.

[MNF14] Mirjalili, V.; Noyes, K.; Feig, M. “Physics-based protein structure refinement through multiple molecular dynamics trajectories and structure averaging”, *Proteins: Structure, Function and Bioinformatics*, vol. 82–SUPPL.2, 2014, pp. 196–207.

[Mob12] Mobley, D. L. “Let’s get honest about sampling”, *Journal of Computer-Aided Molecular Design*, vol. 26, 2012, pp. 93–95.

[MP92] Marinari, E.; Parisi, G. “Simulated tempering: A new monte carlo scheme”, *Europhysics Letters*, vol. 19–6, 1992, pp. 451.

[MPD15] MacCallum, J. L.; Perez, A.; Dill, K. A. “Determining protein structures by combining semireliable data with atomistic physical models by bayesian inference”, *Proceedings of the National Academy of Sciences*, vol. 112–22, 2015, pp. 6985–6990.

[MRR+53] Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. “Equation of state calculations by fast computing machines”, *The Journal of Chemical Physics*, vol. 21–6, 1953, pp. 1087.

[MRSF+00] Martí-Renom, M. A.; Stuart, A. C.; Fiser, A.; Sánchez, R.; Melo, F.; Sali, A. “Comparative protein structure modeling of genes and genomes”, *Annual Review of Biophysics and Biomolecular Structure*, vol. 29–1, 2000, pp. 291–325.

[MS15] Michino, M.; Shi, L. “Computational Approaches in the Structure–Function Studies of Dopamine Receptors”. New York: Springer, 2015, *Neuromethods*, vol. 96, pp. 31–42.

[MSC+10] Maisuradze, G. G.; Senet, P.; Czaplewski, C.; Liwo, A.; Scheraga, H. A. “Investigation of protein folding by coarse-grained molecular dynamics with the unres force field”, *The Journal of Physical Chemistry A*, vol. 114–13, 2010, pp. 4471–4485.

[MSLS14] Maciejczyk, M.; Spasic, A.; Liwo, A.; Scheraga, H. A. “Dna duplex formation with a coarse-grained model”, *Journal of Chemical Theory and Computation*, vol. 10–11, 2014, pp. 5020–5035.

[MSO03] Mitsutake, A.; Sugita, Y.; Okamoto, Y. “Replica-exchange multicanonical and multicanonical replica-exchange monte carlo simulations of peptides. ii. application to a more complex system”, *The Journal of Chemical Physics*, vol. 118, 2003, pp. 6676–6688.

[NBBJ06] Narang, P.; Bhushan, K.; Bose, S.; Jayaram, B. “Protein structure evaluation using an all-atom energy based empirical scoring function”, *Journal of Biomolecular Structure and Dynamics*, vol. 23–4, 2006, pp. 385–406.

[NdSO99] Norberto de Souza, O. N.; Ornstein, R. L. “Molecular dynamics simulations of a protein-protein dimer: Particle- mesh ewald electrostatic model yields far superior results to standard cutoff model”, *Journal of Biomolecular Structure and Dynamics*, vol. 16–6, 1999, pp. 1205–1218.

[NFA02] Neidigh, J. W.; Fesinmeyer, R. M.; Andersen, N. H. “Designing a 20-residue protein”, *Nature Structural & Molecular Biology*, vol. 9–6, 2002, pp. 425–430.

[NH07] Nadler, W.; Hansmann, U. H. E. “Dynamics and optimal number of replicas in parallel tempering simulations”, *Physical Review E*, vol. 76–6, 2007, pp. 065701.

[NMH+14] Nguyen, H.; Maier, J.; Huang, H.; Perrone, V.; Simmerling, C. “Folding simulations for proteins with diverse topologies are accessible in days with a physics-based force field and implicit solvent”, *Journal of the American Chemical Society*, vol. 136–40, 2014, pp. 13959–13962.

[NMK94] Ngo, J. T.; Marks, J.; Karplus, M. “Computational Complexity, Protein Structure Prediction, and the Levinthal Paradox”. Boston: Birkhäuser Boston, 1994, pp. 433–506.

[NRB12] Nagata, K.; Randall, A.; Baldi, P. “Sidepro: A novel machine learning approach for the fast and accurate prediction of side-chain conformations”, *Proteins: Structure, Function and Bioinformatics*, vol. 80–1, 2012, pp. 142–153.

[NSR+09] Neuweiler, H.; Sharpe, T. D.; Rutherford, T. J.; Johnson, C. M.; Allen, M. D.; Ferguson, N.; Fersht, A. R. “The folding mechanism of bbl: Plasticity of transition-state structure observed within an ultrafast folding protein family”, *Journal of Molecular Biology*, vol. 390–5, 2009, pp. 1060 – 1073.

[Nym08] Nymeyer, H. “How efficient is replica exchange molecular dynamics? an analytic approach”, *Journal of Chemical Theory and Computation*, vol. 4–4, 2008, pp. 626–636.

[OCB02] Onufriev, A.; Case, D. A.; Bashford, D. “Effective born radii in the generalized born approximation: The importance of being perfect”, *Journal of Computational Chemistry*, vol. 23–14, 2002, pp. 1297–1304.

[OS14] Olson, B.; Shehu, A. “Multi-objective optimization techniques for conformational sampling in template-free protein structure prediction”. In: 6^{th} International Conference on Bioinformatics and Computational Biology, 2014, pp. 143–148.

[Osg00] Osguthorpe, D. J. “Ab initio protein folding”, *Current Opinion in Structural Biology*, vol. 10–2, 2000, pp. 146–152.

[OWCD07] Ozkan, S. B.; Wu, G. A.; Chodera, J. D.; Dill, K. A. “Protein folding by zipping and assembly”, *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104–29, 2007, pp. 11987–11992.

[OZ14] Ostermeir, K.; Zacharias, M. “Hamiltonian replica-exchange simulations with adaptive biasing of peptide backbone and side chain dihedral angles”, *Journal of Computational Chemistry*, vol. 35–2, 2014, pp. 150–8.

[PCC+95] Pearlman, D. A.; Case, D. A.; Caldwell, J. W.; Ross, W. S.; Cheatham Iii, T. E.; DeBolt, S.; Ferguson, D.; Seibel, G.; Kollman, P. “Amber, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules”, *Computer Physics Communications*, vol. 91–1-3, 1995, pp. 1–41.

[PGW+12] Park, I. H.; Gangupomu, V.; Wagner, J.; Jain, A.; Vaidehi, N. “Structure refinement of protein low resolution models using the gneimo constrained dynamics method”, *Journal of Physical Chemistry B*, vol. 116–8, 2012, pp. 2365–2375.

[PJW03] Ponder J. W., C. D. A. “Force fields for protein simulations”, *Advances in Protein Biochemistry*, vol. 66–5, 2003, pp. 27–85.

[PKS03] Pokarowski, P.; Kolinski, A.; Skolnick, J. “A minimal physically realistic protein-like lattice model: Designing an energy landscape that ensures all-or-none folding to a unique native state”, *Biophysical Journal*, vol. 84–3, 2003, pp. 1518–1526.

[PL96] Park, B.; Levitt, M. “Energy functions that discriminate x-ray and near native folds from well-constructed decoys”, *Journal Molecular Biology*, vol. 258–2, 1996, pp. 367–392.

[PLLD+12] Piana, S.; Lindorff-Larsen, K.; Dirks, R. M.; Salmon, J. K.; Dror, R. O.; Shaw, D. E. “Evaluating the effects of cutoffs and treatment of long-range electrostatics in protein folding simulations”, *PLoS ONE*, vol. 7–6, 2012.

[PM97] Pedersen, J. T.; Moult, J. “Protein folding simulations with genetic algorithms and a detailed molecular description”, *Journal of Molecular Biology*, vol. 269–2, 1997, pp. 240–259.

[PM07] Periole, X.; Mark, A. E. “Convergence and sampling efficiency in replica exchange simulations of peptide folding in explicit solvent”, *The Journal of Chemical Physics*, vol. 126–1, 2007.

[PMD15] Perez, A.; MacCallum, J. L.; Dill, K. A. “Accelerating molecular simulations of proteins using bayesian inference on weak information”, *Proceedings of the National Academy of Sciences of the United States of America*, vol. 112–38, 2015.

[PMSD16] Perez, A.; Morrone, J. A.; Simmerling, C.; Dill, K. A. “Advances in free-energy-based simulations of protein folding and ligand binding”, *Current Opinion in Structural Biology*, vol. 36, 2016, pp. 25–31.

[PNG07] Paschek, D.; Nymeyer, H.; García, A. E. “Replica exchange simulation of reversible folding/unfolding of the trp-cage miniprotein in explicit solvent: On the structure and possible role of internal water”, *Journal of Structural Biology*, vol. 157–3, 2007, pp. 524–533.

[PPC05] Predescu, C.; Predescu, M.; Ciobanu, C. V. “On the efficiency of exchange in parallel tempering monte carlo simulations”, *The Journal of Physical Chemistry B*, vol. 109–9, 2005, pp. 4189–96.

[PPLB07] Pedreira, O.; Piattini, M.; Luaces, M. R.; Brisaboa, N. R. “A systematic review of software process tailoring”, *SIGSOFT Software Engineering Notes*, vol. 32, 2007, pp. 1–6.

[PS03] Pitera, J. W.; Swope, W. “Understanding folding and design: Replica-exchange simulations of trp-cage miniproteins”, *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100–13, 2003, pp. 7587–7592.

[PvdS08] Patriksson, A.; van der Spoel, D. “A temperature predictor for parallel tempering simulations”, *Physical Chemistry Chemical Physics*, vol. 10–15, 2008, pp. 2073–2077.

[RBC14] Roe, D. R.; Bergonzo, C.; Cheatham, T. E. “Evaluation of enhanced sampling provided by accelerated molecular dynamics with hamiltonian replica exchange methods”, *Journal of Physical Chemistry B*, vol. 118–13, 2014, pp. 3543–3552.

[RC03] Rao, F.; Caflisch, A. “Replica exchange molecular dynamics simulations of reversible folding”, *Journal of Chemical Physics*, vol. 119–7, 2003, pp. 4035–4042.

[RCB77] Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. “Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes”, *Journal of Computational Physics*, vol. 23–3, 1977, pp. 327–341.

[RCdP05] Rathore, N.; Chopra, M.; de Pablo, J. J. “Optimal allocation of replicas in parallel tempering simulations”, *The Journal of Chemical Physics*, vol. 122–2, 2005, pp. 024111.

[RGFP09] Roy, S.; Goedecker, S.; Field, M. J.; Penev, E. “A minima hopping study of all-atom protein folding and structure prediction”, *Journal of Physical Chemistry B*, vol. 113–20, 2009, pp. 7315–7321.

[RKZ10] Roy, A.; Kucukural, A.; Zhang, Y. “I-tasser: a unified platform for automated protein structure and function prediction”, *Nature Protocols*, vol. 5–4, 2010, pp. 725–738.

[RO09] Rentzsch, R.; Orengo, C. A. “Protein function prediction – the power of multiplicity”, *Trends in Biotechnology*, vol. 27–4, 2009, pp. 210–219.

[ROS07] Roitberg, A. E.; Okur, A.; Simmerling, C. “Coupling of replica exchange simulations to a non-boltzmann structure reservoir”, *Journal of Physical Chemistry B*, vol. 111–10, 2007, pp. 2415–2418.

[RP03] Rhee, Y. M.; Pande, V. S. “Multiplexed-replica exchange molecular dynamics method for protein folding simulation”, *Biophysical Journal*, vol. 84–2, 2003, pp. 775–786.

[RPE+12] Raval, A.; Piana, S.; Eastwood, M. P.; Dror, R. O.; Shaw, D. E. “Refinement of protein structure homology models via long, all-atom molecular dynamics simulations”, *Proteins-Structure Function and Bioinformatics*, vol. 80–8, 2012, pp. 2071–2079.

[RPES16] Raval, A.; Piana, S.; Eastwood, M. P.; Shaw, D. E. “Assessment of the utility of contact-based restraints in accelerating the prediction of protein structure using molecular dynamics simulations”, *Protein Science*, vol. 25–1, 2016, pp. 19–29.

[RSMB04] Rohl, C. A.; Strauss, C. E. M.; Misura, K. M. S.; Baker, D. “Protein structure prediction using rosetta”, *Methods in Enzymology*, vol. 383, 2004, pp. 66 – 93, numerical Computer Methods, Part D.

[Sai94] Saito, M. “Molecular dynamics simulations of proteins in solution: Artifacts caused by the cutoff approximation”, *The Journal of Chemical Physics*, vol. 101–5, 1994, pp. 4055–4061.

[SB93] Sali, A.; Blundell, T. L. “Comparative protein modelling by satisfaction of spatial restraints”, *Journal of Molecular Biology*, vol. 234–3, 1993, pp. 779–815.

[SBRB99] Simons, K. T.; Bonneau, R.; Ruczinski, I.; Baker, D. “Ab initio protein structure prediction of casp iii targets using rosetta”, *Proteins: Structure, Function, and Genetics*, vol. 37–S3, 1999, pp. 171–176.

[SDD+08] Shaw, D. E.; Deneroff, M. M.; Dror, R. O.; Kuskin, J. S.; Larson, R. H.; Salmon, J. K.; Young, C.; Batson, B.; Bowers, K. J.; Chao, J. C.; Eastwood, M. P.; Gagliardo, J.; Grossman, J. P.; Ho, C. R.; Lerardi, D. J.; Kolossváry, I.; Klepeis, J. L.; Layman, T.; McLeavey, C.; Moraes, M. A.; Mueller, R.; Priest, E. C.; Shan, Y.; Spengler, J.; Theobald, M.; Towles, B.; Wang, S. C. “Anton, a special-purpose machine for molecular dynamics simulation”, *Communications of the ACM*, vol. 51–7, 2008, pp. 91–97.

[SER10] Sindhikara, D. J.; Emerson, D. J.; Roitberg, A. E. “Exchange often and properly in replica exchange molecular dynamics”, *Journal of Chemical Theory and Computation*, vol. 6–9, 2010, pp. 2804–2808.

[SFGP+13] Salomon-Ferrer, R.; Götz, A. W.; Poole, D.; Le Grand, S.; Walker, R. C. “Routine microsecond molecular dynamics simulations with amber on gpus. 2. explicit solvent particle mesh ewald”, *Journal of Chemical Theory and Computation*, vol. 9–9, 2013, pp. 3878–3888.

[SG01] Sanbonmatsu, K. Y.; Garcia, A. E. “Structure of metenkephalin in explicit solvent using replica molecular dynamics”, *Biophysical Journal*, vol. 80, 2001, pp. 399A–399A.

[Shi13] Shirts, M. R. “Simple quantitative tests to validate sampling from thermodynamic ensembles”, *Journal of Chemical Theory and Computation*, vol. 9–2, 2013, pp. 909–926.

[SHVW05] Schug, A.; Herges, T.; Verma, A.; Wenzel, W. “Investigation of the parallel tempering method for protein folding”, *Journal of Physics Condensed Matter*, vol. 17–18, 2005, pp. S1641–S1650.

[Sip95] Sippl, M. J. “Knowledge-based potentials for proteins”, *Current Opinion in Structural Biology*, vol. 5–2, 1995, pp. 229–235.

[SK93] Smarr, L. L.; Kaufmann, W. J. “Supercomputing and the transformation of science”. New York: W.H. Freeman, 1993, 256p.

[SKS+15] Sieradzan, A. K.; Krupa, P.; Scheraga, H. A.; Liwo, A.; Czaplewski, C. “Physics-based potentials for the coupling between backbone- and side-chain-local conformational states in the united residue (unres) force field for protein simulations”, *Journal of Chemical Theory and Computation*, vol. 11–2, 2015, pp. 817–831.

[SLD98] Schneider, J. P.; Lombardi, A.; DeGrado, W. F. “Analysis and design of three-stranded coiled coils and three-helix bundles”, *Folding and Design*, vol. 3–2, 1998, pp. R29–R40.

[SM01] Sarisky, C. A.; Mayo, S. L. “The $\beta\beta\alpha$ fold: explorations in sequence space”, *Journal of Molecular Biology*, vol. 307–5, 2001, pp. 1411 – 1418.

[Smi05] Smith, J. E. “The Co-Evolution of Memetic Algorithms for Protein Structure Prediction”. Berlin: Springer, 2005, pp. 105–128.

[SMLL+10] Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y.; Wriggers, W. “Atomic-level characterization of the structural dynamics of proteins”, *Science*, vol. 330–6002, 2010, pp. 341–346.

[SMR08] Sindhikara, D.; Meng, Y.; Roitberg, A. E. “Exchange frequency in replica exchange molecular dynamics”, *Journal of Chemical Physics*, vol. 128–2, 2008.

[SO99] Sugita, Y.; Okamoto, Y. “Replica-exchange molecular dynamics method for protein folding”, *Chemical Physics Letters*, vol. 314–1–2, 1999, pp. 141–151.

[SPHvdS05] Seibert, M. M.; Patriksson, A.; Hess, B.; van der Spoel, D. “Reproducible polypeptide folding and structure prediction using molecular dynamics simulations”, *Journal of Molecular Biology*, vol. 354–1, 2005, pp. 173–183.

[SR95] Srinivasan, R.; Rose, G. D. “Linus: A hierarchic procedure to predict the fold of a protein”, *Proteins: Structure, Function and Genetics*, vol. 22–2, 1995, pp. 81–99.

[SR02] Srinivasan, R.; Rose, G. D. “Ab initio prediction of protein structure using linus”, *Proteins: Structure, Function and Genetics*, vol. 47–4, 2002, pp. 489–495.

[SS92a] Schreiber, H.; Steinhauser, O. “Cutoff size does strongly influence molecular dynamics results on solvated polypeptides”, *Biochemistry*, vol. 31–25, 1992, pp. 5856–5860.

[SS92b] Schreiber, H.; Steinhauser, O. “Molecular dynamics studies of solvated polypeptides: Why the cut-off scheme does not work”, *Chemical Physics*, vol. 168–1, 1992, pp. 75–89.

[SS92c] Schreiber, H.; Steinhauser, O. “Taming cut-off induced artifacts in molecular dynamics studies of solvated polypeptides: The reaction field method”, *Journal of Molecular Biology*, vol. 228–3, 1992, pp. 909–923.

[SSBOV+09] Scott Shell, M.; Banu Ozkan, S.; Voelz, V.; Wu, G. A.; Dill, K. A. “Blind test of physics-based prediction of protein structures”, *Biophysical Journal*, vol. 96–3, 2009, pp. 917–924.

[SSR02] Simmerling, C.; Strockbine, B.; Roitberg, A. E. “All-atom structure prediction and folding simulations of a stable protein”, *Journal of the American Chemical Society*, vol. 124–38, 2002, pp. 11258–11259.

[SSRP05] Snow, C. D.; Sorin, E. J.; Rhee, Y. M.; Pande, V. S. “How well can simulation predict protein folding kinetics and thermodynamics?” Palo Alto: Annual Reviews, 2005, vol. 34, pp. 43–69.

[STHH90] Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. “Semianalytical treatment of solvation for molecular mechanics and dynamics”, *Journal of the American Chemical Society*, vol. 112–16, 1990, pp. 6127–6129.

[STTC07] Shao, J.; Tanner, S. W.; Thompson, N.; Cheatham, T. E. “Clustering molecular dynamics trajectories: 1. characterizing the performance of different clustering algorithms”, *Journal of Chemical Theory and Computation*, vol. 3–6, 2007, pp. 2312–2334.

[Sue03] Suenaga, A. “Replica-exchange molecular dynamics simulations for a small-sized protein folding with implicit solvent”, *Journal of Molecular Structure: THEOCHEM*, vol. 634–1–3, 2003, pp. 235–241.

[Sun95] Sun, S. “A genetic algorithm that seeks native states of peptides and proteins”, *Biophysical Journal*, vol. 69–2, 1995, pp. 340–355.

[SW86] Swendsen, R. H.; Wang, J.-S. “Replica monte carlo simulation of spin-glasses”, *Physical Review Letters*, vol. 57–21, 1986, pp. 2607.

[Sö05] Söding, J. “Protein homology detection by hmm-hmm comparison”, *Bioinformatics*, vol. 21–7, 2005, pp. 951–960.

[TC00] Tsui, V.; Case, D. A. “Theory and applications of the generalized born solvation model in macromolecular simulations”, *Biopolymers*, vol. 56–4, 2000, pp. 275–291.

[TD11] Toxvaerd, S.; Dyre, J. C. “Communication: Shifted forces in molecular dynamics”, *The Journal of Chemical Physics*, vol. 134–8, 2011, pp. 081102.

[TGPE04] Teodorescu, O.; Galor, T.; Pillardy, J.; Elber, R. “Enriching the sequence substitution matrix by structural information”, *Proteins: Structure, Function and Genetics*, vol. 54–1, 2004, pp. 41–48.

[TM99] Tuckerman, M. E.; Martyna, G. J. “Understanding modern molecular dynamics: Techniques and applications”, *The Journal of Physical Chemistry B*, vol. 104–2, 1999, pp. 159–178.

[Toz05] Tozzini, V. “Coarse-grained models for proteins”, *Current Opinion in Structural Biology*, vol. 15–2, 2005, pp. 144–150.

[TPB+15] Tiberti, M.; Papaleo, E.; Bengtsen, T.; Boomsma, W.; Lindorff-Larsen, K. “Encore: Software for quantitative ensemble comparison”, *PLoS Computational Biology*, vol. 11–10, 2015, pp. e1004415.

[Tra04] Tramontano, A. “Integral and differential form of the protein folding problem”, *Physics of Life Reviews*, vol. 1–2, 2004, pp. 103–127.

[Tra07] Tramontano, A. “Protein structure prediction. concepts and applications.”, *Angewandte Chemie International Edition*, vol. 46–23, 2007, pp. 4213–4213.

[TSH07] Thachuk, C.; Shmygelska, A.; Hoos, H. H. “A replica exchange monte carlo algorithm for protein folding in the hp model”, *BMC Bioinformatics*, vol. 8–1, 2007, pp. 342.

[TTH06] Trebst, S.; Troyer, M.; Hansmann, U. H. E. “Optimized parallel tempering simulations of proteins”, *Journal of Chemical Physics*, vol. 124–17, 2006, pp. 174903.

[UM93] Unger, R.; Moult, J. “Genetic algorithms for protein folding simulations”, *Journal of Molecular Biology*, vol. 231–1, 1993, pp. 75–81.

[UUAD08] Urbic, T.; Urbic, T.; Avbelj, F.; Dill, K. A. “Molecular simulations find stable structures in fragments of protein g”, *Acta Chimica Slovenica*, vol. 2008–55, 2008, pp. 385–395.

[VGB90] Van Gunsteren, W. F.; Berendsen, H. J. C. “Computer simulation of molecular dynamics: Methodology, applications, and perspectives in chemistry”, *Angewandte Chemie*, vol. 29–9, 1990, pp. 992–1023.

[VRS03] Vila, J. A.; Ripoll, D. R.; Scheraga, H. A. “Atomically detailed folding simulation of the b domain of staphylococcal protein a from random structures”, *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100–25, 2003, pp. 14812–14816.

[VS12] Verma, A.; Schug, A. “Hamiltonian replica exchange simulations to enhance sampling for protein folding”, *Biophysical Journal*, vol. 102–3, 2012, pp. 456a.

[VV06] Voet, D.; Voet, J. G. “Bioquímica”. Porto Alegre: Artmed, 2006, 1616p.

[VVVT+04] Vermeulen, W.; Vanhaesebrouck, P.; Van Troys, M.; Verschueren, M.; Fant, F.; Goethals, M.; Ampe, C.; Martins, J. C.; Borremans, F. A. M. “Solution structures of the c-terminal headpiece subdomains of human villin and advillin, evaluation of headpiece f-actin-binding requirements”, *Protein Science*, vol. 13–5, 2004, pp. 1276–1287.

[VW09] Verma, A.; Wenzel, W. “A free-energy approach for all-atom protein simulation”, *Biophysical Journal*, vol. 96–9, 2009, pp. 3483–3494.

[WAA+14] Weiner, B. E.; Alexander, N.; Akin, L. R.; Woetzel, N.; Karakas, M.; Meiler, J. “Bcl: Fold-protein topology determination from limited nmr restraints”, *Proteins: Structure, Function and Bioinformatics*, vol. 82–4, 2014, pp. 587–595.

[WL03] Whisstock, J. C.; Lesk, A. M. “Prediction of protein function from protein sequence and structure”, *Quarterly Reviews of Biophysics*, vol. 36–3, 2003, pp. 307–340.

[XM08] Xu, W.; Mu, Y. “Ab initio folding simulation of trpcage by replica exchange with hybrid hamiltonian”, *Biophysical Chemistry*, vol. 137–2–3, 2008, pp. 116–125.

[XX00] Xu, Y.; Xu, D. “Protein threading using prospect: Design and evaluation”, *Proteins: Structure, Function and Genetics*, vol. 40–3, 2000, pp. 343–354.

[XYZ15] Xue, X.; Yongjun, W.; Zhihong, L. “Folding of sam-ii riboswitch explored by replica-exchange molecular dynamics simulation”, *Journal of Theoretical Biology*, vol. 365–0, 2015, pp. 265–269.

[XZ12] Xu, D.; Zhang, Y. “Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field”, *Proteins: Structure, Function and Bioinformatics*, vol. 80–7, 2012, pp. 1715–1735.

[YCK12] Yuan, C.; Chen, H.; Kihara, D. “Effective inter-residue contact definitions for accurate protein fold recognition”, *BMC Bioinformatics*, vol. 13–1, 2012, pp. 292.

[YFZZ11] Yang, Y.; Faraggi, E.; Zhao, H.; Zhou, Y. “Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates”, *Bioinformatics*, vol. 27–15, 2011, pp. 2076–2082.

[YP03] Young, M. R.; Pande, V. S. “Multiplexed-replica exchange molecular dynamics method for protein folding simulation”, *Biophysical Journal*, vol. 84–2 I, 2003, pp. 775–786.

[YSG09] Yang, L.; Shao, Q.; Gao, Y. Q. “Thermodynamics and folding pathways of trpzip2: An accelerated molecular dynamics simulation study”, *Journal of Physical Chemistry B*, vol. 113–3, 2009, pp. 803–808.

[YZ08] Yang, Y.; Zhou, Y. “Specific interactions for ab initio folding of protein terminal regions with secondary structures”, *Proteins: Structure, Function, and Bioinformatics*, vol. 72–2, 2008, pp. 793–803.

[ZAH05] Zhu, J.; Alexov, E.; Honig, B. “Comparative study of generalized born models: Born radii and peptide folding”, *Journal of Physical Chemistry B*, vol. 109–7, 2005, pp. 3008–3022.

[ZAS05] Zhang, Y.; Arakaki, A. K.; Skolnick, J. R. “Tasser: An automated method for the prediction of protein tertiary structures in casp6”, *Proteins-Structure Function and Bioinformatics*, vol. 61, 2005, pp. 91–98.

[ZB07] Zvelebil, M.; Baum, J. “Understanding Bioinformatics”. New York: Garland Science, 2007, 772p.

[ZDY+11] Zhou, Y.; Duan, Y.; Yang, Y.; , E.; Lei, H. “Trends in template/fragment-free protein structure prediction”, *Theoretical Chemistry Accounts*, vol. 128–1, 2011, pp. 3–16.

[Zem03] Zemla, A. “Lga: A method for finding 3d similarities in protein structures”, *Nucleic Acids Research*, vol. 31–13, 2003, pp. 3370–3374.

[Zho04] Zhou, R. “Exploring the protein folding free energy landscape: Coupling replica exchange method with p3me/respa algorithm”, *Journal of Molecular Graphics and Modelling*, vol. 22–5, 2004, pp. 451–463.

[ZLC+07] Zhang, J.; Lin, M.; Chen, R.; Liang, J.; Liu, J. S. “Monte carlo sampling of near-native structures of proteins with applications”, *Proteins: Structure, Function and Genetics*, vol. 66–1, 2007, pp. 61–68.

[ZS04a] Zhang, Y.; Skolnick, J. “Automated structure prediction of weakly homologous proteins on a genomic scale”, *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101–20, 2004, pp. 7594–7599.

[ZS04b] Zhang, Y.; Skolnick, J. “Scoring function for automated assessment of protein structure template quality”, *Proteins: Structure, Function, and Bioinformatics*, vol. 57–4, 2004, pp. 702–710.

[ZS11] Zhou, H.; Skolnick, J. “Goap: A generalized orientation-dependent, all-atom statistical potential for protein structure prediction”, *Biophysical Journal*, vol. 101–8, 2011, pp. 2043 – 2052.

[ZS15] Zhang, Y.; Sagui, C. “Secondary structure assignment for conformationally irregular peptides: Comparison between dssp, stride and kaksi”, *Journal of Molecular Graphics and Modelling*, vol. 55–0, 2015, pp. 72–84.

[ZSSP02] Zagrovic, B.; Snow, C. D.; Shirts, M. R.; Pande, V. S. “Simulation of folding of a small alpha-helical protein in atomistic detail using worldwide-distributed computing”, *Journal of Molecular Biology*, vol. 323–5, 2002, pp. 927–937.

[ZWD05] Zhang, W.; Wu, C.; Duan, Y. “Convergence of replica exchange molecular dynamics”, *Journal of Chemical Physics*, vol. 123–15, 2005, pp. 154105.

[ZZ02] Zhou, H.; Zhou, Y. “Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction”, *Protein Science*, vol. 11–11, 2002, pp. 2714–2726.

[ZZ10] Zhang, J.; Zhang, Y. “A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction”, *PLOS ONE*, vol. 5–10, 10 2010, pp. 1–13.

APÊNDICE A – PROTOCOLO DE MAPEAMENTO SISTEMÁTICO

Este mapeamento seguiu os preceitos estipulados por Pereira *et al.* [PPLB07], e Kitchenham [Kit07].

1. FORMULAÇÃO DA QUESTÃO:

   (a) Questões foco: O foco de interesse fica por conta de sumarizar a informação existente no que se trata do tratamento do problema da predição de estrutura tridimensionais de proteína no âmbito das diferentes técnicas utilizadas para amostrar a superfície de energia, ressaltando as abordagens utilizadas e os resultados alcançados até então. O segundo objetivo do mapeamento é o de identificar lacunas na pesquisa que sugiram novos rumos para a pesquisa na área.

   (b) Qualidade e amplitude da questão:

      i. Problema: Predição de estrutura de proteínas através de Dinâmica Molecular, Monte Carlo e Replica Exchange Molecular Dynamics

      ii. Questão: O que foi feito até então se utilizando de DM, MC ou REMD para o problema da predição de estrutura de proteínas?

      iii. Palavras-Chave: Para a predição de proteínas: *"Protein Structure Prediction*" e "*Protein Folding*". Para as técnicas de amostragem: "*Molecular Dynamics*" , "Monte Carlo" e "*Replica Exchange Molecular Dynamics*".

      iv. Intervenção: Será observado o tipo de abordagem de amostragem utilizada pelo trabalho e o tipo de método utilizado (*ab initio*, *de novo*, baseado em *templates*, etc)), além de levar em conta as características específicas de cada trabalho como a maneira de representação das proteínas, linguagem utilizada, tamanho das proteínas e resultados.

      v. Efeito: Descobrir como as diferentes técnicas tem sido utilizadas até então.

      vi. Medidas de saída: RMSD [GDT] e tamanho das proteínas testadas, tempo de execução e método.

      vii. Aplicação: A estrutura terciária de uma proteína está diretamente ligada a sua função, pois pode permitir a identificação de domínios conhecidos, como sítios catalíticos, sítios de modificação alostérica e outros [Les08]. Tendo em vista que a grande maioria dos fármacos atualmente no mercado atuam interagindo com proteínas, o estudo da relação estrutura-função mostra-se vital para a criação de novos fármacos e a bioinformática possui o importante papel de acelerar o processo de evolução deste conhecimento [ZB07]. A solução do problema PSP, ou avanços no seu tratamento, nos permitirá obter estruturas 3D de proteínas importantes, com aplicações relevantes na indús-

tria biofarmacêutica. Ela nos permitirá compreender a estrutura de proteínas envolvidas em processos vitais, incluindo doenças como o câncer [DK01].

2. SELEÇÃO DE FONTES:

   (a) Critério de seleção de fontes: Para a execução das pesquisas foi utilizada a ferramenta StArt [FHT+12].

   (b) Linguagem: Inglês

   (c) Identificação de fontes:

      i. Métodos de busca de fontes: Embora seja utilizada a ferramenta StArt (conforme explicado em 2.a), a mesma não permite que as buscas sejam automatizadas. Além disso, uma importante característica da busca por palavras-chave é a de que todas as buscas foram feitas em todo o documento (não somente no *abstract*/palavras-chave).

      ii. *Strings* de busca: Foram utilizadas combinações de palavras chave entre as duas grandes áreas alvo da pesquisa: Simulação Multi-agente e Técnica de Amostragem. Para cada grande área foram escolhidas diferentes palavras chave, em inglês: Para a predição de proteínas: "*Protein Structure Prediction*" e "*Protein Folding*". Para as técnicas de amostragem: "*Molecular Dynamics*", "Monte Carlo" e "*Replica Exchange Molecular Dynamics*". Uma vez que não se pretendia relacionar, por hora, diferentes métodos de amostragem, geraram-se 6 strings de busca, 3 para cada método, verificando como os mesmos estavam inseridos no contexto da predição de estruturas.

         - Para Dinâmica Molecular: "*Protein Structure Prediction*" AND "*Molecular Dynamics*" e "*Protein Folding*" AND "*Molecular Dynamics*".
         - Para Monte Carlo: "*Protein Structure Prediction*" AND "Monte Carlo" e "*Protein Folding*" AND "Monte Carlo".
         - Para REMD: "*Protein Structure Prediction*" AND "*Replica Exchange Molecular Dynamics*" e "*Protein Folding*" AND "*Replica Exchange Molecular Dynamics*"

      iii. Um conjunto de 5 bases de dados (exposto na Tabela A.1) foi utilizado para buscas visando a identificação dos estudos. Vale ressaltar, no entanto, que as bases de dados utilizadas são somente aquelas que passaram pelo aval do especialista na etapa de verificação de referências descrita em iii.e :

   (d) Seleção de fontes pós-avaliação: Nada a declarar.

   (e) Verificação de referências: Segundo Biolchini et al. em [BMC+05], a verificação da lista de bases de dados deve feita por especialista, com objetivo de retirar ou adicionar fontes . A verificação da lista de base de dados fica então, a cargo do especialista de domínio prof. Dr Osmar Norberto de Souza.

3. SELEÇÃO DE ESTUDOS Uma vez que as bases de dados estão definidas é necessário definir o processo e os critérios para seleção e avaliação dos estudos

   (a) Definição de estudos:

      i. Definição de critérios de inclusão e exclusão:Dada a grande quantidade de bases de dados alvo da pesquisa, a pesquisa utilizando-se das palavras chaves descritas em 1.b.iii encontrou um número demasiadamente grande de artigos não relacionados à questão de pesquisa do mapeamento sistemático, tornando necessária a definição de critérios bem definidos para a inclusão/exclusão de trabalhos. Os critérios passaram por um teste inicial para ter certeza de que eram capazes de classificar (incluir/não incluir) os trabalhos corretamente, chamaremos esse teste de piloto criterial. Para evitar que o viés do pesquisador afete a revisão, seguem os seguintes critérios:
      Critério 1: Serão incluídos artigos tanto de natureza qualitativa quanto quantitativa.
      Critério 2: Todo tipo de trabalho pode ser incluído, não apenas artigos.
      Critério 3: Os artigos devem passar pelos procedimentos de seleção descritos em 3.a.iii para serem considerados parte efetiva do conjunto de artigos que a revisão sistemática analisará.

      ii. Definição de Tipos de Estudos: Os estudos foram divididos de acordo com o tipo de abordagem de exploração conformacional que utilizam e o nível estrutural de proteínas que possuem como alvo.

      iii. Procedimentos para Seleção de Estudos: A seleção de estudos foi um processo composto por vários estágios. Como o conjunto inicial de trabalhos foi obtido de forma automática, muitos dos resultados que acataram as palavras-chave procuradas não tinham relação com o que procurávamos. Para descobrir quais artigos deveriam ser levados em conta, foi criado um procedimento de seleção. Primeiramente, partindo-se dos resultados obtidos através da pesquisas da *string* de busca nas referidas bases de dados iniciou-se o processo de retirada de duplicatas. Posteriormente, com o conjunto de trabalhos restantes, iniciou-se o processo de filtragem dos resultados afim de descobrir quais dos artigos realmente acatavam os interesses. A filtragem foi feita lendo-se os *abstracts*/palavras-chave de cada trabalho e excluindo os trabalhos que fossem julgados totalmente fora do escopo. Passamos então à fase de leitura da introdução dos trabalhos, o que caracteriza a 6a etapa da metodologia utilizada. Os trabalhos julgados fora do escopo foram retirados do conjunto de trabalhos sob análise e os restantes foram lidos por completo (7a etapa). Os trabalhos que passaram pela 7a etapa sem serem descartados foram aqueles estudados a fundo.

         A. Primeira etapa: Escolha das palavras chave.

B. Segunda etapa: Escolha das bases de dados.
C. Terceira etapa: Pesquisa.
D. Quarta etapa: Retirada de duplicatas.
E. Quinta etapa: 1o Filtro: Leitura de *abstracts* / palavras-chave.
F. Sexta etapa: 2o Filtro: Leitura da introdução.
G. Sétima etapa: 3o Filtro: Leitura do artigo completo.

Tabela A.1 – Lista de bases de dados

Nome da Base
ACM
IEEE
Pubmed / Medline (NLM)
Scopus
Web of Science (ISI)

4. RESULTADOS DO MAPEAMENTO:

Ao término da terceira etapa, um total de 3064 artigos foram capturados, das diversas fontes. A Tabela A.2 apresenta a contribuição de cada base na pesquisa.

Tabela A.2 – Contribuição por base de dados

Fonte	Quantidade de Artigos	%
ACM	11	≈ 0
IEEE	1002	≈ 33
PubMed	3	≈ 0
Scopus	1507	≈ 49
Web of Science	541	≈ 18

Seguindo a metodologia do software StArt, a 5a etapa foi aplicada, aplicando-se então o primeiro filtro nos artigos capturados. Após a leitura de *abstracts* e palavras-chave, parte dos artigos foi considerada irrelevante para o trabalho. A Figura 4 contabiliza a quantidade de artigos "aceitos", “rejeitados” e marcados como “duplicados”. Os artigos marcados como “aceitos” passaram então à próxima fase, a chamada Fase de Extração.

Na fase de extração, os artigos passam pelas etapas 6 e 7, ou seja, por mais dois filtros. Ao final da etapa 7, temos a quantidade final de artigos considerados relevantes ao Mapeamento Sistemático.

A Figura 4 apresenta a quantidade final de artigos considerados relevantes para o Mapeamento Sistemático em questão. Embora artigos duplicados tenham sido encontrados anteriormente, o software StArt o faz de forma automatizada e, assim sendo,

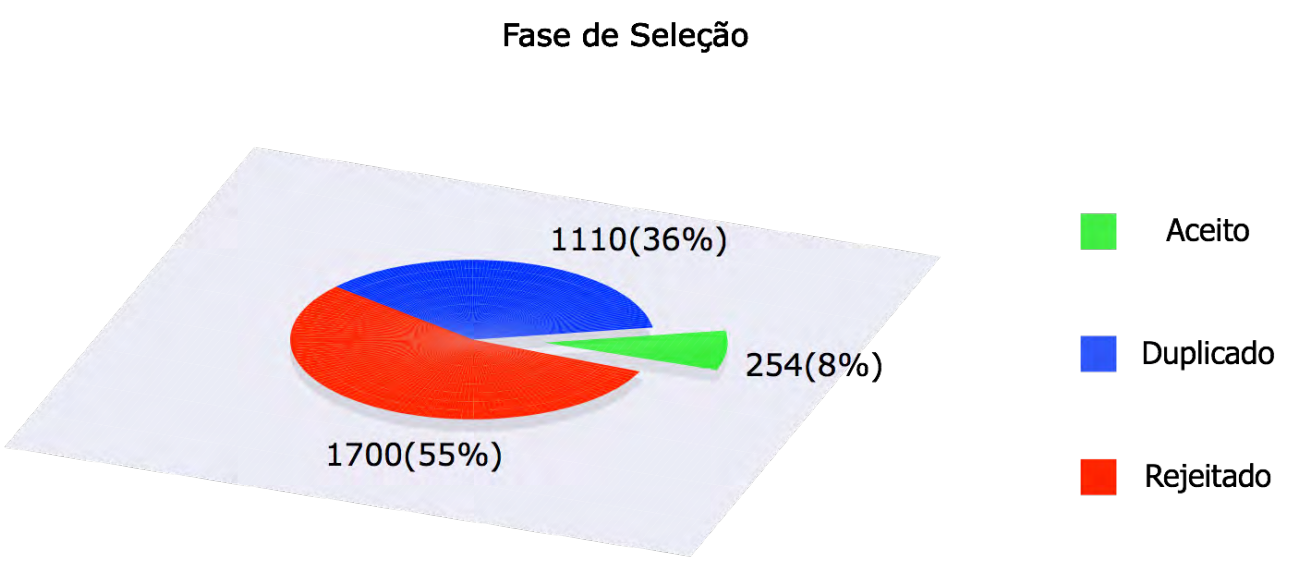

Figura A.1 – Resultados da fase de seleção de artigos. Em verde a quantidade de artigos aceitos para a fase de extração, em vermelho os rejeitados e, em azul, os duplicados.

pode ocorrer de artigos duplicados não serem considerados como tal. Isso explica a quantidade de "duplicados"encontrados na fase de extração.

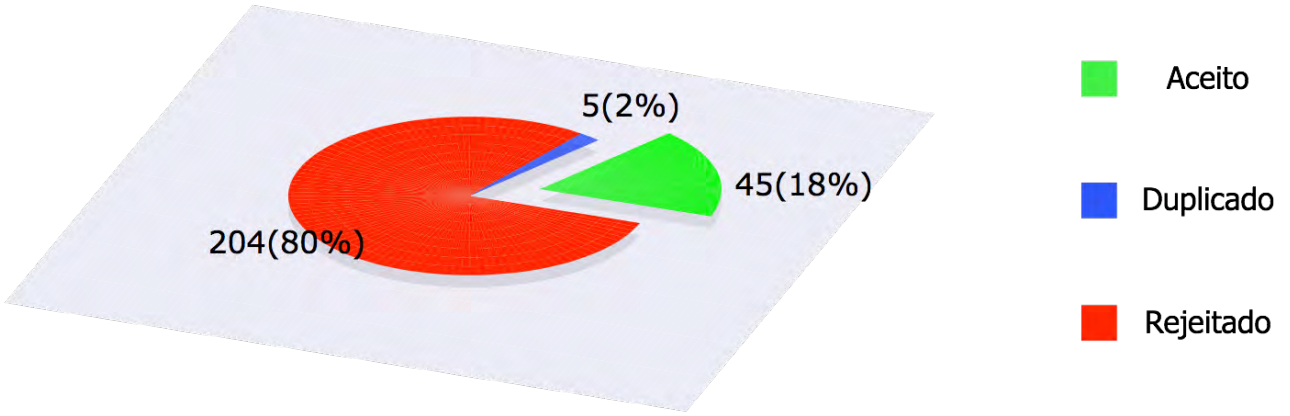

Figura A.2 – Resultados da fase de extração de artigos. Em verde a quantidade de artigos aceitos, em vermelho os rejeitados e, em azul, os duplicados.

As Tabelas A.3, A.3, A.4, A.5, A.6, A.7 e A.8 expõem os 45 artigos capturados como resultado do Mapeamento Sistemático executado para esta tese. Levando em consideração a pesquisa inicial, 45 simboliza menos de 2,3 % dos artigos capturados já desconsiderando os 1115 artigos duplicados.

Tabela A.3 – Artigos aceitos na fase de extração: parte 1. A prioridade de cada artigo e dada com base no *abstract* lido na fase anterior. *Score* é calculado de forma automática baseando-se em palavras-chave.

Titulo	Autores	Prioridade	Score	Ano	Periódico
Ab Initio Folding of Proteins with All-Atom Discrete Molecular Dynamics	Ding, F. and Tsao, D. and Nie, H. and Dokholyan, N.V.	BAIXA	63	2008	Structure
Ab initio folding simulation of Trpcage by replica exchange with hybrid Hamiltonian	Xu, W. and Mu, Y.	BAIXA	57	2008	Biophysical Chemistry
Accelerating molecular simulations of proteins using {Bayesian} inference on weak information.	Perez, Alberto and MacCallum, Justin L. and Dill, Ken A.	ALTA	12	2015	Proceedings of the National Academy of Sciences of the United States of America
Application of biasing-potential replicaexchange simulations for loop modeling and refinement of proteins in explicit solvent	Kannan, S. and Zacharias, M.	BAIXA	57	2010	Proteins: Structure, Function and Bioinformatics
Assessment of the utility of contact-based restraints in accelerating the prediction of protein structure using molecular dynamics simulations.	Raval, Alpan and Piana, Stefano and Eastwood, Michael P. and Shaw, David E.	MUITO_ALTA	27	2015	Protein science : a publication of the Protein Society
Combining coarse-grained protein models with replica-exchange all-atom molecular dynamics	Wabik, J. and Kmiecik, S. and Gront, D. and Kouza, M. and Kolinski, A.	MUITO_ALTA	92	2013	International Journal of Molecular Sciences
Effect of short- and long-range interactions on protein folding	Anderson, J.S. and Scheraga, H.A.	BAIXA	30	1982	Journal of Protein Chemistry
Enhanced sampling of peptide and protein conformations using replica exchange simulations with a peptide backbone biasing-potential	Kannan, S. and Zacharias, M.	BAIXA	68	2007	Proteins: Structure, Function and Genetics

Tabela A.4 – Artigos aceitos na fase de extração: parte 2. A prioridade de cada artigo e dada com base no *abstract* lido na fase anterior. *Score* é calculado de forma automática baseando-se em palavras-chave.

Titulo	Autores	Prioridade	Score	Ano	Periódico
Exploring the energy landscape of protein folding using replica-exchange and conventional molecular dynamics simulations	David A.C. Beck and George W.N. White and Valerie Daggett	BAIXA	114	2007	Journal of Structural Biology
Exploring the protein folding free energy landscape: Coupling replica exchange method with P3ME/RESPA algorithm	Zhou, R.	BAIXA	105	2004	Journal of Molecular Graphics and Modelling
Folding free-energy landscape of villin headpiece subdomain from molecular dynamics simulations	Lei, H. and Wu, C. and Liu, H. and Duan, Y.	MUITO_ALTA	60	2007	Proceedings of the National Academy of Sciences of the United States of America
Folding of fourteen small proteins with a residue-specific force field and replica-exchange molecular dynamics	Jiang, F. and Wu, Y.-D.	BAIXA	129	2014	Journal of the American Chemical Society
Folding of small proteins using constrained molecular dynamics	Balaraman, G.S. and Park, I.-H. and Jain, A. and Vaidehi, N.	MUITO_ALTA	86	2011	Journal of Physical Chemistry B
Folding of Trp-cage mini protein using temperature and biasing potential replica-exchange molecular dynamics simulations	Kannan, S. and Zacharias, M.	MUITO_ALTA	165	2009	International Journal of Molecular Sciences
Folding processes of the B domain of protein A to the native state observed in all-atom ab initio folding simulations	Lei, H. and Wu, C. and Wang, Z.-X. and Zhou, Y. and Duan, Y.	ALTA	105	2008	Journal of Chemical Physics
Folding simulation of Trp-cage utilizing a new AMBER compatible force field with coupled main chain torsions	Mou, L. and Jia, X. and Gao, Y. and Li, Y. and Zhang, J.Z.H. and Mei, Y.	BAIXA	9	2014	Journal of Theoretical and Computational Chemistry

Tabela A.5 – Artigos aceitos na fase de extração: parte 3. A prioridade de cada artigo e dada com base no *abstract* lido na fase anterior. *Score* é calculado de forma automática baseando-se em palavras-chave.

Titulo	Autores	Prioridade	Score	Ano	Periódico
Folding simulations for proteins with diverse topologies are accessible in days with a physics-based force field and implicit solvent	Nguyen, H. and Maier, J. and Huang, H. and Perrone, V. and Simmerling, C.	MUITO_ALTA	6	2014	Journal of the American Chemical Society
Folding simulations of Trp-cage mini protein in explicit solvent using biasing potential replica-exchange molecular dynamics simulations	Kannan, S. and Zacharias, M.	BAIXA	72	2009	Proteins: Structure, Function and Bioinformatics
Folding very short peptides using molecular dynamics	Ho, B.K. and Dill, K.A.	BAIXA	51	2006	PLoS Computational Biology
Fragment replica-exchange method for efficient protein conformation sampling	Suzuki, M. and Okuda, H.	MUITO_BAIXA	49	2008	Molecular Simulation
Hamiltonian Replica Exchange Simulations to Enhance Sampling for Protein Folding	Abhinav Verma and Alexander Schug	BAIXA	45	2012	Biophysical Journal
Hamiltonian replica-exchange simulations with adaptive biasing of peptide backbone and side chain dihedral angles	Ostermeir, K. and Zacharias, M.	MUITO_ALTA	200	2014	Journal of Computational Chemistry
Hydrophobic aided replica exchange: An efficient algorithm for protein folding in explicit solvent	Liu, P. and Huang, X. and Zhou, R. and Berne, B.J.	ALTA	59	2006	Journal of Physical Chemistry B
Insights into the folding pathway of the Engrailed Homeodomain protein using replica exchange molecular dynamics simulations	Koulgi, S. and Sonavane, U. and Joshi, R.	ALTA	191	2010	Journal of Molecular Graphics and Modelling

Tabela A.6 – Artigos aceitos na fase de extração: parte 4. A prioridade de cada artigo e dada com base no *abstract* lido na fase anterior. *Score* é calculado de forma automática baseando-se em palavras-chave.

Titulo	Autores	Prioridade	Score	Ano	Periódico
Investigation of protein folding by coarse-grained molecular dynamics with the UNRES force field	Maisuradze, G.G. and Senet, P. and Czaplewski, C. and Liwo, A. and Scheraga, H.A.	MUITO_ALTA	129	2010	Journal of Physical Chemistry A
Microsecond scale replica exchange molecular dynamic simulation of villin headpiece: An insight into the folding landscape	Jani, V. and Sonavane, U.B. and Joshi, R.	MUITO_BAIXA	96	2011	Journal of Biomolecular Structure and Dynamics
MMTSB Tool Set: Enhanced sampling and multiscale modeling methods for applications in structural biology	Feig, M. and Karanicolas, J. and Brooks III, C.L.	BAIXA	63	2004	Journal of Molecular Graphics and Modelling
Molecular Simulations Find Stable Structures in Fragments of Protein G	Urbic, Tjasa and Urbic, Tomaz and Avbelj, Franc and Dill, Ken A.	MUITO_BAIXA	0	2008	Acta chimica Slovenica
Multiplexed-replica exchange molecular dynamics method for protein folding simulation	Young, M.R. and Pande, V.S.	ALTA	112	2003	Biophysical Journal
On the Hamiltonian replica exchange method for efficient sampling of biomolecular systems: Application to protein structure prediction	Fukunishi, H. and Watanabe, O. and Takada, S.	MUITO_ALTA	84	2002	Journal of Chemical Physics
Predicting Three-Dimensional Conformations of Peptides Constructed of Only Glycine, Alanine, Aspartic Acid, and Valine	Oda, A. and Fukuyoshi, S.	MUITO_BAIXA	15	2015	Origins of Life and Evolution of Biospheres

Tabela A.7 – Artigos aceitos na fase de extração: parte 5. A prioridade de cada artigo e dada com base no *abstract* lido na fase anterior. *Score* é calculado de forma automática baseando-se em palavras-chave.

Titulo	Autores	Prioridade	Score	Ano	Periódico
Protein folding by zipping and assembly	Ozkan, S.B. and Wu, G.A. and Chodera, J.D. and Dill, K.A.	MUITO_ALTA	60	2007	Proceedings of the National Academy of Sciences of the United States of America
Protein folding simulations by generalized-ensemble algorithms	Yoda, T. and Sugita, Y. and Okamoto, Y.	ALTA	100	2014	Advances in Experimental Medicine and Biology
Protein folding simulations combining self-guided langevin dynamics and temperature-based replica exchange	Lee, M.S. and Olson, M.A.	BAIXA	45	2010	Journal of Chemical Theory and Computation
Protein Folding with the Parallel Replica Exchange Molecular Dynamics Method	Zhou, R.	MUITO_BAIXA	45	2005	Parallel Computing for Bioinformatics and Computational Biology: Models, Enabling Technologies, and Case Studies
Protein structure prediction and refinement using folding mechanism-informed replica exchange methods	Shell, M.S. and Ozkan, S.B.	BAIXA	30	2008	AIChE Annual Meeting, Conference Proceedings
Protein structure prediction by tempering spatial constraints	Gront, D. and Kolinski, A. and Hansmann, U.H.E.	BAIXA	64	2005	Journal of Computer-Aided Molecular Design
Recent advances in implicit solvent-based methods for biomolecular simulations	Chen, J. and Brooks III, C.L. and Khandogin, J.	BAIXA	29	2008	Current Opinion in Structural Biology
REMD and umbrella sampling simulations to probe the energy barrier of the folding pathways of engrailed homeodomain.	Jani, Vinod and Sonavane, Uddhavesh B. and Joshi, Rajendra	BAIXA	63	2014	Journal of molecular modeling

Tabela A.8 – Artigos aceitos na fase de extração: parte 6. A prioridade de cada artigo e dada com base no *abstract* lido na fase anterior. *Score* é calculado de forma automática baseando-se em palavras-chave.

Titulo	Autores	Prioridade	Score	Ano	Periódico
Reordering hydrogen bonds using Hamiltonian replica exchange enhances sampling of conformational changes in biomolecular systems	Vreede, J. and Wolf, M.G. and De Leeuw, S.W. and Bolhuis, P.G.	BAIXA	110	2009	Journal of Physical Chemistry B
Replica-exchange molecular dynamics simulations for a small-sized protein folding with implicit solvent	A. Suenaga	BAIXA	81	2003	Journal of Molecular Structure: {THEOCHEM}
Reproducible Polypeptide Folding and Structure Prediction using Molecular Dynamics Simulations	M. Marvin Seibert and Alexandra Patriksson and Berk Hess and David van der Spoel	BAIXA	59	2005	Journal of Molecular Biology
The temperature intervals with global exchange of replicas empirical accelerated sampling method: Parameter sensitivity and extension to a complex molecular system	Li, X. and Latour, R.A.	MUITO_ALTA	146	2011	Journal of Computational Chemistry
Towards fast, rigorous and efficient conformational sampling of biomolecules: Advances in accelerated molecular dynamics	Doshi, U. and Hamelberg, D.	MUITO_ALTA	104	2015	Biochimica et Biophysica Acta - General Subjects
Trends in template/fragment-free protein structure prediction	Zhou, Y. and Duan, Y. and Yang, Y. and Faraggi, E. and Lei, H.	MUITO_ALTA	21	2011	Theoretical Chemistry Accounts

REFERÊNCIAS BIBLIOGRÁFICAS

[*BMC*+05] Biolchini, J.; Mian, P.; Conte, T.; Natali, A.; Travassos, G. “A systematic review process for software engineering”. In: 2*nd* *Experimental Software Engineering Latin American Workshop*, 2005, pp. 2049-2075

[*DK*01] Duan, Y.; Kollman, P. A. "Computational protein folding: from lattice to all-atom", *IBM Systems Journal*, vol. 40, 2001, pp. 297-309.

[*Kit*07] Kitchenham, B. A. "Guidelines for performing Systematic Literature Reviews in Software", Technical Report, Keele University, 2007, 65p.

[*Les*08] Lesk, A. M. “Introduction to bioinformatics”. New York: Oxford University Press, 2008, 3 ed., 474p.

[*PPLB*07] Pedreira, O.; Piattini, M.; Luaces, M. R.; Brisaboa, N. R. “A systematic review of software process tailoring”, *SIGSOFT Software Engineering Notes*, vol. 32, 2007, pp. 1–6.

[*ZB*07] Zvelebil, M.; Baum, J. "Understanding Bioinformatics". New York: Garland Science, 2007, 772p.

APÊNDICE B – DESCRIÇÃO DETALHADA DOS PARÂMETROS DAS SIMULAÇÕES

Neste capítulo serão explicados os parâmetros do AMBER para simulações por DM e REMD, juntamente com com o respectivo valor de cada parâmetro utilizado pelas simulações CuT-REMD, com base no manual do AMBER, versão 14.0 [CBB+14].

Parâmetros Gerais

- imin=0,
  sem minimização

- ntx=1
  Opção para ler as coordenadas do arquivo “inpcrd”. Somente as opções 1 e 2 são suportadas nesta versão. Outras opções farão com que o pbsa emita um aviso, embora não afete o cálculo da energia. = 1 X é lido formatado sem informação de velocidade inicial. Padrão.

- *irest*=1
  Flag para reiniciar uma simulação. Um irest = 0 (padrão) não reinicia a simulação; Em vez disso, executa como uma nova simulação. As velocidades no arquivo de coordenadas de entrada, se houver, serão ignoradas e a contagem de *timesteps* será definida como 0 (a menos que substituída por t; veja abaixo). Um irest = 1 reinicia a simulação, lendo as coordenadas e as velocidades de um arquivo de reinício previamente salvo. As informações de velocidade são necessárias ao reiniciar, de modo que ntx deve ser 4 ou superior se irest = 1.

- *ntxo*=2,
  Formato das coordenadas finais, velocidades e tamanho da caixa (se a simulação utiliza volume ou pressão constantes) gravados no arquivo “restrt”.
  = 1 (padrão), formato ASCII.
  =2, NetCDFfile .

- *ntpr*=1000,
  Cada *ntpr* passos, informações de energia seão impressas em forma legível para os arquivos “mdout” e “mdinfo”. “Mdinfo” é fechado e reaberto cada vez, de modo que sempre contém a mais recente energia e temperatura.
  = 50, (padrão).

- *ntave*=0,
  A cada *ntave* passos de dinâmica, médias em tempo real das energias e as flutuações sobre os últimos *ntave* passos são impressas. Um valor de 0 desativa esta impressão. Definir *ntave* para um valor 1/2 ou 1/4 de *nstlim* fornece uma maneira simples de observar a convergência durante a simulação.
  = 0 (padrão), desativado.

- *ntwr*=100000
  A cada *ntwx* passos, as coordenadas são gravadas no arquivo mdcrd.
  = 0 (padrão), nenhum arquivo de trajetória de coordenadas será gravado.

- *iwrap*=0
  Se *iwrap* = 1, as coordenadas escritas nos arquivos de reinício e de trajetórias serão "envolvidas" (*wrapped*) em uma caixa primária. Isto significa que para cada molécula, a sua imagem periódica mais próxima do meio da "caixa primária" (com coordenadas x entre 0 e a, coordenadas y entre 0 e b e coordenadas z entre 0 e c) será escrita no arquivo de saída. Isso muitas vezes faz com que as estruturas resultantes pareçam melhores visualmente, mas não tem nenhum efeito sobre energia ou forças. Executar tal envolvimento, no entanto, pode atrapalhar difusão e outros cálculos.
  Se *iwrap* = 0, nada disso será feito, caso em que é típico usar cpptraj como programa de pós-processamento para converter moléculas de volta para a caixa principal. Para execuções muito longas, a configuração de *iwrap* = 1 pode ser necessária para evitar que as coordenadas de saída provoquem *overflow* prejudicando toda a trajetória sendo gravada e os arquivos de reinicialização, especialmente se as trajetórias estiverem escritas em formato ASCII e não NetCDF.
  = 0 (padrão), desativado.

- *ntwx*=1000
  A cada *ntwx* passos, as coordenadas serão gravadas para no arquivo mdcrd.
  = 0 (padrão), nenhum arquivo de trajetória de coordenadas será gravado.

- *ntwv*=0,
  A cada *ntwv* passos, as velocidades serão gravadas no arquivo mdvel.
  = 0 (padrão), nenhum arquivo de trajetória de velocidades será gravado.
  = -1, as velocidades serão gravadas em mdcrd, que então se torna um arquivo combinado coordenada/trajetória de velocidades, no intervalo definido por *ntwx*. Esta opção está disponível apenas para saída binária NetCDF (*ioutfm* = 1).
  A maioria dos usuários não terá necessidade de um arquivo de trajetória de velocidade e, portanto, poderá deixar seguramente *ntwv* no padrão. Observe que escrevendo velocidades com freqüência, assim como forças ou coordenadas, irão introduzir sobrecargas de comunicação de E/S potencialmente significativas, prejudicando tanto o desempenho como a paralelização.

- *ntwe*=0
  Cada *ntwe* passos, as energias e as temperaturas serão escritas no arquivo “mden” em uma forma compacta.
  = 0 (padrão), nenhum arquivo mden será escrito.
  Observe que as energias no arquivo mden não são sincronizadas com coordenadas ou velocidades nos arquivos mdcrd ou mdvel. Assumindo valores *ntwe* e *ntwx* idênticos, as energias são um passo de tempo antes das coordenadas (bem como as velocidades que são sincronizadas com as coordenadas). Conseqüentemente, um arquivo mden raramente é escrito.

- *ioutfm*=1
  O formato dos arquivos de trajetória de coordenadas e velocidade (mdcrd, mdvel e inptraj). A partir do AMBER, o formato binário utilizado em versões anteriores não é mais suportado; A saída binária está agora no formato de trajetória NetCDF. Embora não seja a opção padrão, os arquivos de trajetória binária têm muitas vantagens: são menores, possuem maior precisão, muito mais rápidos de ler e gravar e são capazes de aceitar uma faixa mais ampla de valores de coordenadas (ou velocidades) do que os arquivos de trajetória formatados.
  = 0, (padrão) trajetória ASCII formatada.
  = 1, trajetória NetCDF binária.

Átomos Congelados ou Restringidos

- *ibelly*
  = 0 (padrão), desativado.
  = 1, um subconjunto dos átomos no sistema será autorizado a se mover e as coordenadas dos restantes serão congeladas. Os átomos em movimento são especificados por uma máscara de *ibelly*. Esta opção não está disponível quando *igb* > 0. Observe também que esta opção não fornece nenhuma vantagem em termos de desempenho significativa e é mantida basicamente para compatibilidade com versões anteriores do AMBER. A maioria das aplicações deve usar a variável *ntr* ao invés de restringir partes do sistema com o objetivo de fazê-las permanecerem próximas de alguma configuração inicial.

- ntr=0
  Flag para restringir átomos especificados no espaço cartesiano usando um potencial harmônico.
  = 0 (padrão), desativado.
  > 0, os átomos restritos são determinados pela *string* resttramask. A constante de

força é dada pela restrição _wt. As coordenadas são lidas no formato “restrt” a partir do arquivo “refc”.

Dinâmica Molecular

- *nstlim*=100000
  Número de passos de DM a serem executados.
  =1 (padrão).

- *nscm*=1000,
  Flag para a remoção do movimento de translação e de rotação do centro de massa em intervalos regulares (o padrão é 1000). Para simulações não-periódicas, após cada *nscm* passos, movimentos de translação e rotação são removidos. Para sistemas periódicos, apenas o movimento de translação do centro de massa será removido. Este parâmetro é ignorado para simulações com *belly*. Para a dinâmica de Langevin, a posição do centro de massa da molécula é reposta em zero em cada passo de *nscm*, mas as velocidades não são afetadas. Por conseguinte, não há qualquer alteração nas componentes de translação ou de rotação dos momentos (fazer qualquer outra coisa destruiria a maneira pela qual a temperatura é regulada em um sistema de dinâmica de Langevin). A única razão para redefinir as coordenadas é impedir que a molécula acabe tão longe da origem que suas coordenadas ocasionem *overflow* devido ao formato dos arquivos utilizado na reinicialização ou na criação das trajetórias.

- *t*=0.0
  O tempo no início (ps). Tal parâmetro é para ser utilizado como referência pelo usuário e não é crítico. A hora de início é obtida do arquivo de entrada de coordenadas se *irest* = 1.
  = 0 (padrão).

- *dt*=0.001 para *cut < 6.0* e 0.002 caso contrário
  Passo de integração (ps). O máximo recomendado pelo manual do AMBER é .002 se SHAKE é utilizado, ou .001 se não for. Observe que para temperaturas acima de 300K, o tamanho do passo deve ser reduzido uma vez que temperaturas maiores significam velocidades maiores e maior distância percorrida entre cada avaliação de força, o que pode levar a energias anormalmente altas e à explosão do sistema. Impacta diretamente a rapidez das simulações.
  =0.001 (padrão).

- *nrespa*=1,
  Esta variável permite que o usuário avalie os termos de variação lenta no campo de força com menor frequência. Para PME, “variando lentamente” (agora) significa a

soma recíproca. Para simulações com GB as forças de “variação lenta” são aquelas que envolvem derivadas com relação aos raios efetivos, e interações de pares, cujas distâncias são maiores que o ponto de corte “interno”, atualmente ligado por cabo a 8 Å. Se NRESPA> 1 essas forças de variação lenta são avaliadas cada passo nrespa. As forças são ajustadas apropriadamente, levando a um impulso nesse passo. Se nrespa * dt for menor ou igual a 4 fs, a conservação de energia não é seriamente comprometida. No entanto se nrespa * dt> 4 fs a simulação torna-se menos estável. Note que as energias e as quantidades relacionadas são acessíveis somente a cada passo nrespa, já que os valores noutros momentos não têm sentido.

Regulação de Temperatura

- *ntt*=1,
  Desvio para escala de temperatura. Observe que a configuração *ntt*=0 corresponde ao *ensemble* micro-canônico (NVE) (que deve se aproximar do canônico para número de graus de liberdade elevado). Alguns aspectos do “*ensemble* de acoplamento fraco” ou *weak-coupling ensemble* (*ntt*=1) foram examinados e interpolam-se grosseiramente entre os *ensembles* micro-canônico e canônico [Mor00,MC04]. As opções *ntt*= 2 e 3 correspondem ao *ensemble* canônica (T constante).
  = 1, temperatura constante, usando o algoritmo de acoplamento fraco [BPvG+84]. Um único fator de escala é usado para todos os átomos. Note que este algoritmo apenas garante que a energia cinética total seja apropriada para a temperatura desejada; Ele não faz nada para garantir que a temperatura seja a mesma sobre todas as partes da molécula. As colisões atômicas tenderão a garantir uma distribuição uniforme da temperatura, mas isso não é garantido e há muitos problemas sutís que podem surgir com o fraco acoplamento de temperatura [HTC98]. O uso de *ntt*=1 é especialmente perigoso para simulações por *Generalized Born*, onde não há colisões com solvente para auxiliar na termalização. Em vez disso, devem ser usadas outras opções de acoplamento de temperatura (especialmente *ntt*=3).

- tempi=10.0,
  Temperatura inicial. Para a execução inicial da dinâmica (*ntx* < 3), as velocidades são atribuídas a partir de uma distribuição de Maxwell em TEMPI K. Se TEMPI = 0.0, as velocidades serão calculadas a partir das forças. TEMPI não tem efeito se *ntx* > 3.
  = 0 (padrão).

- *temp0*=XXXXX,
  Temperatura de referência em que o sistema deve ser mantido, se *ntt* > 0. Note que para temperaturas acima de 300K, o tamanho do degrau deve ser reduzido, uma vez que o aumento da distância percorrida entre avaliações pode levar a SHAKE e outros

problemas.
= 300 (padrão).

- *ig*=RANDOM_NUMBER,
  A semente aleatória ou número semente para o gerador de números pseudo-aleatórios. A velocidade de partida da DM depende desse valor se *ntx* > 3 e TEMPI != 0.0. O valor desta semente também afeta o conjunto de valores pseudo-aleatórios usados para dinâmica de Langevin ou acoplamento de Andersen (*Andersen coupling*) e, portanto, deve ser ajustado para um valor diferente em cada reinício se *ntt* = 2 ou 3.
  = 71277 (padrão).
  Se *ig* = -1, a semente aleatória será baseada na data e hora atuais e, portanto, será diferente para cada execução. Recomenda-se que, a menos que você deseje especificamente reprodutibilidade (caso do trabalho presente nesta tese, por exemplo), que você defina *ig* = -1 para todas as execuções envolvendo *ntt* = 2 ou 3.

- *tautp*=5.0,
  = 1 (padrão), constante de tempo (em ps) para acoplamento do banho de térmico ao o sistema, se *ntt* = 1.
  Geralmente, os valores para *tautp* devem estar na faixa de 0,5-5,0 ps, com um valor menor proporcionando um acoplamento mais justo ao banho térmico e, assim, resultando em um aquecimento mais rápido e uma trajetória menos natural. Valores menores de *tautp* resultam em flutuações menores na energia cinética, mas flutuações maiores na energia total. Valores muito maiores do que o comprimento da simulação resultam em um retorno a condições de energia constantes.

- *gamma*_ln=0

  A frequência de colisão *gamma* (em ps^{-1}), quando *ntt* = 3. Um integrador Leapfrog simples é utilizado para propagar a dinâmica, com a energia cinética ajustada para ser correta para o caso do oscilador harmônico [PBS88,LBP92]. Note que não é necessário que *gamma* se aproxime da frequência de colisão física, que é aproximadamente 50 ps^{-1} para água líquida. Na verdade, é frequentemente vantajoso, em termos de amostragem ou estabilidade de integração, utilizar valores muito menores, cerca de 2 a 5 ps^{-1} [LBP92,ICWS01].
  = 0 (padrão).

- vlimit=-1,
  Se não for igual a 0.0, então qualquer componente da velocidade que seja maior que *vlimit* será reduzido a *vlimit* (preservando o sinal). Isto pode ser utilizado para evitar instabilidades ocasionais na execução de DMs. O *vlimit* geralmente deve ser ajustado para um valor como 20 (o padrão), que está bem acima da velocidade mais

provável em uma distribuição de Maxwell-Boltzmann à temperatura ambiente. Uma mensagem de aviso será impressa sempre que as velocidades forem modificadas. As execuções que demonstrem mais do que apenas alguns desses avisos devem ser cuidadosamente examinadas.

Regulação de Pressão

- *ntp*=0
  Flag para dinâmicas a pressão constante. Esta opção deve ser definida como 1 ou 2 quando as condições de contorno periódicas de pressão constante são utilizadas.
  = 0, (padrão) sem escala de pressão.

- *pres0*=1.0
  Pressão de referência (em unidades bar, onde 1 bar $\approx$ 0,987 atm) em que o sistema é mantido (quando *ntp* > 0).
  = 1.0 (padrão.)

- *comp*=44.6
  Compressibilidade do sistema quando *ntp* > 0. As unidades estão em 1,0 * 10^{-6} bar^{-1}; Um valor de 44.6 (padrão) é apropriado para a água.

- *taup*=1.0
  Tempo de relaxação da pressão (em ps), quando *ntp* > 0. O valor recomendado está entre 1.0 e 5.0.
  = 1.0 (padrão), no entanto valores maiores que 1.0 podem às vezes serem necessários (se suas trajetórias parecem instáveis).

Restrição de Comprimento de Ligação pelo Algoritmo *SHAKE*

- *ntc*=2

  Flag para SHAKE para executar restrições de comprimento de ligação [306]. A opção SHAKE deve ser utilizada para a maioria dos cálculos de DM. O tamanho do passo de tempo ou *timestep* da DM é determinado pelos movimentos mais rápidos no sistema. SHAKE remove a liberdade de estiramento de ligação, que é o movimento mais rápido, e consequentemente permite que um *timestep* maior seja utilizado. Para os modelos de água, é utilizado um algoritmo especial de “três pontos” [MK92]. Consequentemente, para empregar TIP3P estipule *ntf* = n*tc* = 2. Uma vez que SHAKE é um algoritmo baseado em dinâmica, o minimizador não está ciente do que SHAKE

está a fazer; Por este motivo, as minimizações geralmente devem ser realizadas sem SHAKE. Uma exceção são minimizações curtas cujo objetivo é remover contatos ruins antes que a dinâmica possa começar. Para versões paralelas do SANDER, somente os átomos intramoleculares podem ser limitados. Assim, tais átomos devem estar na mesma cadeia no arquivo PDB de origem.
= 1 (padrão), SHAKE não é executado
= 2 ligações envolvendo hidrogênio são limitadas.

- *tol*=0.00001
  Tolerância geométrica relativa para a reposição de coordenadas em *SHAKE*. Máximo recomendado: < 0.00005.
  = 0.00001 (padrão).

Parâmetros da Função de Potencial

- *ntf*=2
  Avaliação de força. Nota: Se *SHAKE* for utilizado, não é necessário calcular forças para as ligações restritas.
  = 1 (padrão), todas interações são calculadas.
  = 2 interações de ligações envolvendo átomos de hidrogênio são omitidas (utilização com *ntc* = 2).

- *ntb*=0
  Esta variável controla se são impostos ou não limites periódicos ao sistema durante o cálculo de interações não ligadas. Ligações abrangendo limites periódicos ainda não são suportadas. Não há mais necessidade de definir esta variável, pois pode ser determinada a partir dos parâmetros *igb* e *ntp*. O padrão "apropriado" para *ntb* é especificado (*ntb* = 0 quando *igb* > 0, *ntb* = 2 quando *ntp* > 0 e *ntb* = 1 caso contrário). Esse comportamento pode ser substituído pelo fornecimento de um valor explícito, embora isso seja desencorajado para evitar erros.

- *dielc*=1.0
  Constante dielétrica multiplicativa para as interações eletrostáticas. O padrão é 1.0. Observe que isto NÃO está relacionado às constantes dielétricas para cálculos de Generalized Born ou Poisson-Boltzmann. Deve ser utilizado apenas para simulações de quase vácuo, por exemplo quando se pretende ε = 4r; Neste caso, você também deve definir a variável *eedmeth*.

- *cut*=4.0
  Isso é utilizado para especificar o raio de corte não-ligado, em Ångstroms. Para PME,

o raio de corte é utilizado para limitar a somatória de espaço direto e 8,0 é normalmente um bom valor. Quando *igb* > 0, o raio de corte é utilizado para truncar pares não-ligados (em uma base átomo a átomo); Aqui um valor maior do que o padrão é geralmente exigido. Um parâmetro separado (*rgbmax*) controla a distância máxima entre pares de átomos que serão considerados na realização da soma para a par envolvida no cálculo dos raios de Born efetivos.
Quando *igb* > 0, o padrão é 9999.0 (efetivamente infinito)
Quando *igb* == 0, o padrão é 8.0.

- *nsnb*= 10
  Determina a frequência de atualizações de lista não não-ligados quando *igb* = 0 e *nbflag* = 0; Consulte a descrição de *nbflag* para obter mais informações. O padrão é 25.

- *igb*=1
  Flag Bandeira para utilização dos modelos de solventes implícitos Generalized Born ou Poisson-Boltzmann.

- intdiel=1.0
  Define a constante dielétrica interna da molécula de interesse. O padrão é 1.0. Outros valores não foram extensivamente testados.

- *extdiel*=78.5
  Define a constante dielétrica externa ou solvente. O padrão é 78.5.

- *rgbmax*=6.0

  Este parâmetro controla a distância máxima entre os pares de átomos que serão considerados na realização da somatória para a par envolvida no cálculo dos raios de Born efetivos. Átomos cujas esferas associadas estão mais distantes do que *rgbmax* para um certo átomo não contribuirão para o raio de Born efetivo desse átomo. Isto é implementado de uma forma “suave” (graças principalmente a W.A. Svrcek-Seiler), de modo que quando parte da esfera atômica do átomo está dentro do valor de corte *rgbmax*, essa parte contribui para a região de baixa-dielétrica, a qual determina o raio Born efetivo. O padrão é 25 Â, que é geralmente abundante para proteínas de domínio único de algumas centenas de resíduos. Valores ainda menores (de 10-15 Å) são entendidos como razoáveis, alterando um pouco a forma funcional da teoria Generalized Born, em troca de uma aceleração considerável na eficiência e sem introduzir artefatos como deslocamentos na energia total.
  Neste trabalho *rgbmax* foi definido em 6.0 para *cut* < 6.0 devido ao fato de considerarmos tais raios de corte muito baixos e 10.0 caso contrário.
  O parâmetro *rgbmax* afeta apenas os raios de Born efetivos (e as derivadas desses

valores em relação às coordenadas atômicas). O parâmetro *cut*, por outro lado, determina a distância máxima para os termos eletrostáticos, van der Waals e "fora da diagonal" da interação GB. O valor de *rgbmax* pode ser maior ou menor do que o de *cut*: estes dois parâmetros são independentes um do outro.

- *rbornstat*=0
  Se *rbornstat* = 1, as estatísticas dos raios efetivos de Born para cada átomo da molécula em toda a simulação de dinâmica molecular são relatadas no arquivo de saída. O padrão é 0.

- *offset*=0.09
  Os raios dielétricos para os cálculos de GB são diminuídos por um valor uniforme para retornar os "raios intrínsecos" utilizados na obtenção de raios de Born efetivos. O padrão é 0.09 Å.

- *gbsa*=1
  Opção para realização de simulações GB/SA (Generalized Born/Surface Area).
  = 0 (padrão), A área de superfície não será computada e não será incluída no termo de solvatação.
  = 1, a área superficial será calculada usando o modelo LCPO [WSS99].
  = 2, a área superficial será calculada aproximando-se recursivamente de uma esfera em torno de um átomo, a partir de um icosaedro. Observe que nenhuma força é gerada neste caso, portanto, *gbsa* = 2 só funciona para um único cálculo de energia pontual e destina-se principalmente à decomposição de energia no domínio de MM_GBSA.

- *surften*=0.005
  Tensão superficial usada para calcular a contribuição não-polar para a energia livre de solvatação (quando *gbsa* = 1), como Enp = *surften* * SA. O padrão é 0.005 kcal/mol / AA^2 [SSH94].

- *nmropt*=1
  = 1 As restrições de RMN e as alterações de peso serão lidas.

REFERÊNCIAS BIBLIOGRÁFICAS

[*BPvG*+84] Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. "Molecular dynamics with coupling to an external bath", *The Journal of Chemical Physics*, vol. 81–8, 1984, pp. 3684–3690.

[*CBB*+14] Case, D. A.; Babin, V.; Berryman, J. T.; Betz, R. M.; Cai, Q.; Cerutti, D. S.;Cheatham, T. E.; Darden, T. A.; Duke, R. E.; Gohlke, H.; Goetz, A. W.;Gusarov, S.; Homeyer, N.; Janowski, P.; Kaus, J.; Kolossváry, I.; Kovalenko, A.; Lee, T. S.; LeGrand, S.; Luchko, T.; Luo, R.; Madej, B.; Merz, K. M.; Paesani, F.; Roe, D. R.; Roitberg, A.; Sagui, C.; Salomon-Ferrer, R.; Seabra, G.; Simmerling, C. L.; Smith, W.; Swails, J.; Walker; Wang, J.; Wolf, R. M.; Wu, X.; Kollman, P. A. "Amber 14". San Francisco: University of California Press, 2014.

[*MC*04] Mudi, A.; Chakravarty, C. "Effect of the berendsen thermostat on the dynamical properties of water", *Molecular Physics*, vol. 102–7, 2004, pp. 681–685.

[*MK*92] Miyamoto, S.; Kollman, P. A. "Settle: An analytical version of the shake and rattle algorithm for rigid water models", *Journal of Computational Chemistry*, vol. 13–8, 1992, pp. 952–962.

[*Mor*00] Morishita, T. "Fluctuation formulas in molecular-dynamics simulations with theweak coupling heat bath", *The Journal of Chemical Physics*, vol. 113–8, 2000, pp. 2976–2982.

[*HTC*98] Harvey, S. C.; Tan, R. K.-Z.; Cheatham, T. E. "The flying ice cube: Velocity rescaling in molecular dynamics leads to violation of energy equipartition", *Journal of Computational Chemistry*, vol. 19–7, 1998, pp. 726–740.

[*ICWS*01] Izaguirre, J. A.; Catarello, D. P.; Wozniak, J. M.; Skeel, R. D. "Langevin stabilization of molecular dynamics", *The Journal of Chemical Physics*, vol. 114, 2001, pp. 2090–2098.

[*LBP*92] Loncharich, R. J.; Brooks, B. R.; Pastor, R. W. "Langevin dynamics of peptides: The frictional dependence of isomerization rates of nacetylalanyln'methylamide", *Biopolymers*, vol. 32–5, 1992, pp. 523–535.

[*PBS*88] Pastor, R.W.; Brooks, B. R.; Szabo, A. "An analysis of the accuracy of langevin and molecular-dynamics algorithms", *Molecular Physics*, vol. 65–6, 1988, pp. 1409–1419.

[*SSH*94] Sitkoff, D.; Sharp, K. A.; Honig, B. "Accurate calculation of hydration free energies using macroscopic solvent models", *The Journal of Physical Chemistry*, vol. 98–7, 1994, pp. 1978–1988.

[*WSS*99] Weiser, J.; Shenkin, P. S.; Still, W. C. “Approximate atomic surfaces from linear combinations of pairwise overlaps (lcpo)”, *Journal of Computational Chemistry*, vol. 20–2, 1999, pp. 217–230.

APÊNDICE C – COEFICIENTES DE CORRELAÇÃO ENTRE AS SUPERFÍCIES DE ENERGIA AMOSTRADAS PELAS SIMULAÇÕES E O ESPERADO TEORICAMENTE DE UMA DISTRIBUIÇÃO DE BOLTZMANN

As Tabelas a seguir exibem os coeficientes de correlação para cada par de temperaturas. Cada par de temperaturas e seus respectivos pontos na distribuição de energia, a partir do cálculo baseado na Equação 5.1 são ajustados a uma reta que representa seu comportamento. De tal reta pode-se obter ainda um coeficiente angular da memsa, o qual é comparado com a declividade ou *slope* teórico de uma distribuição de Boltzmann, chegando-se então a coeficientes de correlação entre as duas retas.

Tabela C.1 – Coeficientes de correlação para todos os pares de temperaturas, para os protocolos A, B, C, D, E e F. Etapa 1.

Temps (K)/ Método	Coef.Ang Teórico	ETAPA 1 A	B	C	D	E	F
269,50-285,22	0,103	0,600	0,587	0,721	0,713	0,535	0,455
285,22-301,66	0,096	0,561	0,581	0,655	0,679	0,525	0,439
301,66-318,76	0,089	0,492	0,517	0,646	0,622	0,553	0,395
318,76-336,61	0,084	0,469	0,530	0,659	0,634	0,586	0,288
336,61-355,26	0,078	0,532	0,529	0,655	0,608	0,462	0,370
355,26-374,74	0,074	0,461	0,494	0,612	0,614	0,486	0,365
374,74-395,07	0,069	0,436	0,441	0,593	0,59	0,473	0,330
395,07-416,32	0,065	0,448	0,433	0,624	0,605	0,457	0,315
416,32-438,50	0,061	0,419	0,444	0,565	0,596	0,361	0,294
438,50-461,67	0,058	0,419	0,391	0,581	0,542	0,424	0,297
461,67-485,87	0,054	0,373	0,362	0,504	0,539	0,491	0,326
485,87-511,14	0,051	0,390	0,434	0,572	0,576	0,456	0,318
511,14-537,54	0,048	0,446	0,453	0,602	0,607	0,444	0,243
Média	-	0,465	0,477	0,615	0,61	0,481	0,341
Desvio Padrão	-	0,066	0,069	0,055	0,048	0,059	0,061

Tabela C.2 – Coeficientes de correlação para todos os pares de temperaturas, para os protocolos A, B, C, D, E e F. Etapa 2.

Temps (K)/ Método	Coef.Ang Teórico	ETAPA 2 A	B	C	D	E	F
269,50-285,22	0,103	0,818	0,841	0,922	0,906	0,667	0,940
285,22-301,66	0,096	0,852	0,903	0,929	0,89	0,949	0,903
301,66-318,76	0,089	0,828	0,872	0,908	0,354	0,667	0,941
318,76-336,61	0,084	0,789	0,858	0,915	0,919	0,955	0,925
336,61-355,26	0,078	0,848	0,844	0,892	0,871	0,648	0,922
355,26-374,74	0,074	0,856	0,832	0,908	0,867	0,941	0,924
374,74-395,07	0,069	0,810	0,908	0,923	0,895	0,620	0,882
395,07-416,32	0,065	0,813	0,878	0,911	0,892	0,954	0,907
416,32-438,50	0,061	0,878	0,865	0,908	0,887	0,542	0,924
438,50-461,67	0,058	0,818	0,831	0,883	0,85	0,935	0,946
461,67-485,87	0,054	0,830	0,849	0,892	0,908	0,622	0,919
485,87-511,14	0,051	0,835	0,841	0,867	0,891	0,925	0,908
511,14-537,54	0,048	0,760	0,813	0,867	0,848	0,63	0,903
Média	-	0,752	0,781	0,845	0,798	0,715	0,803
Desvio Padrão	-	0,03	0,028	0,02	0,149	0,166	0,018

Tabela C.3 – Coeficientes de correlação para todos os pares de temperaturas, para os protocolos A, B, C, D, E e F. Etapa 3.

Temps (K)/ Método	Coef.Ang Teórico	ETAPA 3 A	B	C	D	E	F
269,50-285,22	0,103	0,907	0,906	0,942	0,957	0,669	0,955
285,22-301,66	0,096	0,908	0,921	0,937	0,935	0,908	0,94
301,66-318,76	0,089	0,915	0,871	0,954	0,924	0,666	0,926
318,76-336,61	0,084	0,935	0,915	0,945	0,936	0,913	0,892
336,61-355,26	0,078	0,931	0,934	0,946	0,935	0,643	0,917
355,26-374,74	0,074	0,932	0,894	0,934	0,941	0,816	0,910
374,74-395,07	0,069	0,934	0,925	0,941	0,947	0,611	0,906
395,07-416,32	0,065	0,911	0,876	0,922	0,919	0,899	0,912
416,32-438,50	0,061	0,914	0,886	0,941	0,949	0,54	0,909
438,50-461,67	0,058	0,889	0,886	0,932	0,958	0,875	0,901
461,67-485,87	0,054	0,913	0,916	0,921	0,942	0,615	0,930
485,87-511,14	0,051	0,889	0,907	0,932	0,911	0,898	0,929
511,14-537,54	0,048	0,892	0,876	0,933	0,913	0,613	0,904
Média	-	0,915	0,903	0,937	0,938	0,754	0,919
Desvio Padrão	-	0,017	0,021	0,009	0,016	0,141	0,017

Tabela C.4 – Coeficientes de correlação para todos os pares de temperaturas, para os protocolos A, B, C, D, E e F. Etapa 4.

Temps (K)/ Método	Coef.Ang Teórico	ETAPA 4 A	B	C	D	E	F
269,50-285,22	0,103	0,924	0,892	0,945	0,947	0,657	0,93
285,22-301,66	0,096	0,944	0,914	0,938	0,939	0,934	0,931
301,66-318,76	0,089	0,916	0,900	0,948	0,92	0,664	0,899
318,76-336,61	0,084	0,929	0,910	0,933	0,955	0,932	0,940
336,61-355,26	0,078	0,903	0,906	0,946	0,326	0,634	0,933
355,26-374,74	0,074	0,898	0,913	0,935	0,946	0,907	0,928
374,74-395,07	0,069	0,895	0,895	0,948	0,950	0,612	0,921
395,07-416,32	0,065	0,922	0,911	0,940	0,935	0,940	0,898
416,32-438,50	0,061	0,908	0,895	0,922	0,935	0,518	0,926
438,50-461,67	0,058	0,878	0,899	0,927	0,925	0,891	0,877
461,67-485,87	0,054	0,895	0,897	0,923	0,929	0,616	0,878
485,87-511,14	0,051	0,905	0,896	0,928	0,915	0,89	0,905
511,14-537,54	0,048	0,884	0,88	0,926	0,933	0,627	0,889
Média	-	0,910	0,902	0,936	0,885	0,766	0,914
Desvio Padrão	-	0,019	0,010	0,010	0,169	0,159	0,022

Tabela C.5 – Coeficientes de correlação para todos os pares de temperaturas, para os protocolos A, B, C, D, E e F. Etapa 5.

Temps (K)/ Método	Coef.Ang Teórico	ETAPA 5 A	B	C	D	E	F
269,50-285,22	0,103	0,911	0,932	0,945	0,965	0,668	0,921
285,22-301,66	0,096	0,924	0,925	0,956	0,948	0,954	0,955
301,66-318,76	0,089	0,943	0,897	0,953	0,934	0,672	0,899
318,76-336,61	0,084	0,942	0,903	0,935	0,959	0,935	0,907
336,61-355,26	0,078	0,918	0,915	0,925	0,954	0,641	0,925
355,26-374,74	0,074	0,916	0,922	0,951	0,936	0,942	0,901
374,74-395,07	0,069	0,923	0,924	0,934	0,942	0,606	0,947
395,07-416,32	0,065	0,920	0,912	0,947	0,950	0,890	0,941
416,32-438,50	0,061	0,920	0,916	0,938	0,927	0,545	0,842
438,50-461,67	0,058	0,887	0,885	0,943	0,938	0,936	0,915
461,67-485,87	0,054	0,906	0,894	0,939	0,908	0,619	0,883
485,87-511,14	0,051	0,878	0,884	0,902	0,934	0,903	0,905
511,14-537,54	0,048	0,849	0,855	0,927	0,926	0,617	0,930
Média	-	0,911	0,905	0,938	0,940	0,764	0,913
Desvio Padrão	-	0,026	0,022	0,014	0,015	0,161	0,030

Tabela C.6 – Coeficientes de correlação para todos os pares de temperaturas, para os protocolos A, B, C, D, E e F. Etapa 6.

Temps (K)/ Método	Coef.Ang Teórico	ETAPA 6 A	B	C	D	E	F
269,50-285,22	0,103	0,987	0,985	0,98	0,993	0,677	0,99
285,22-301,66	0,096	0,989	0,977	0,986	0,989	0,98	0,986
301,66-318,76	0,089	0,986	0,990	0,985	0,989	0,691	0,992
318,76-336,61	0,084	0,982	0,986	0,979	0,989	0,991	0,993
336,61-355,26	0,078	0,986	0,992	0,990	0,983	0,655	0,984
355,26-374,74	0,074	0,983	0,988	0,986	0,989	0,985	0,983
374,74-395,07	0,069	0,987	0,990	0,981	0,983	0,628	0,983
395,07-416,32	0,065	0,983	0,989	0,982	0,984	0,989	0,988
416,32-438,50	0,061	0,979	0,984	0,989	0,985	0,558	0,990
438,50-461,67	0,058	0,987	0,988	0,981	0,989	0,983	0,987
461,67-485,87	0,054	0,989	0,982	0,978	0,985	0,634	0,992
485,87-511,14	0,051	0,982	0,984	0,982	0,985	0,982	0,971
511,14-537,54	0,048	0,982	0,982	0,983	0,989	0,64	0,976
Média	-	0,985	0,986	0,983	0,987	0,799	0,986
Desvio Padrão	-	0,003	0,004	0,004	0,003	0,181	0,006

APÊNDICE D – ANÁLISE COMPARATIVA ENTRE CUT-REMD E REMD CONVENCIONAL NA FORMAÇÃO E ESTABILIZAÇÃO INDIVIDUAL DAS TRÊS HÉLICES QUE COMPÕEM A PROTEÍNA *VILLIN HEADPIECE*

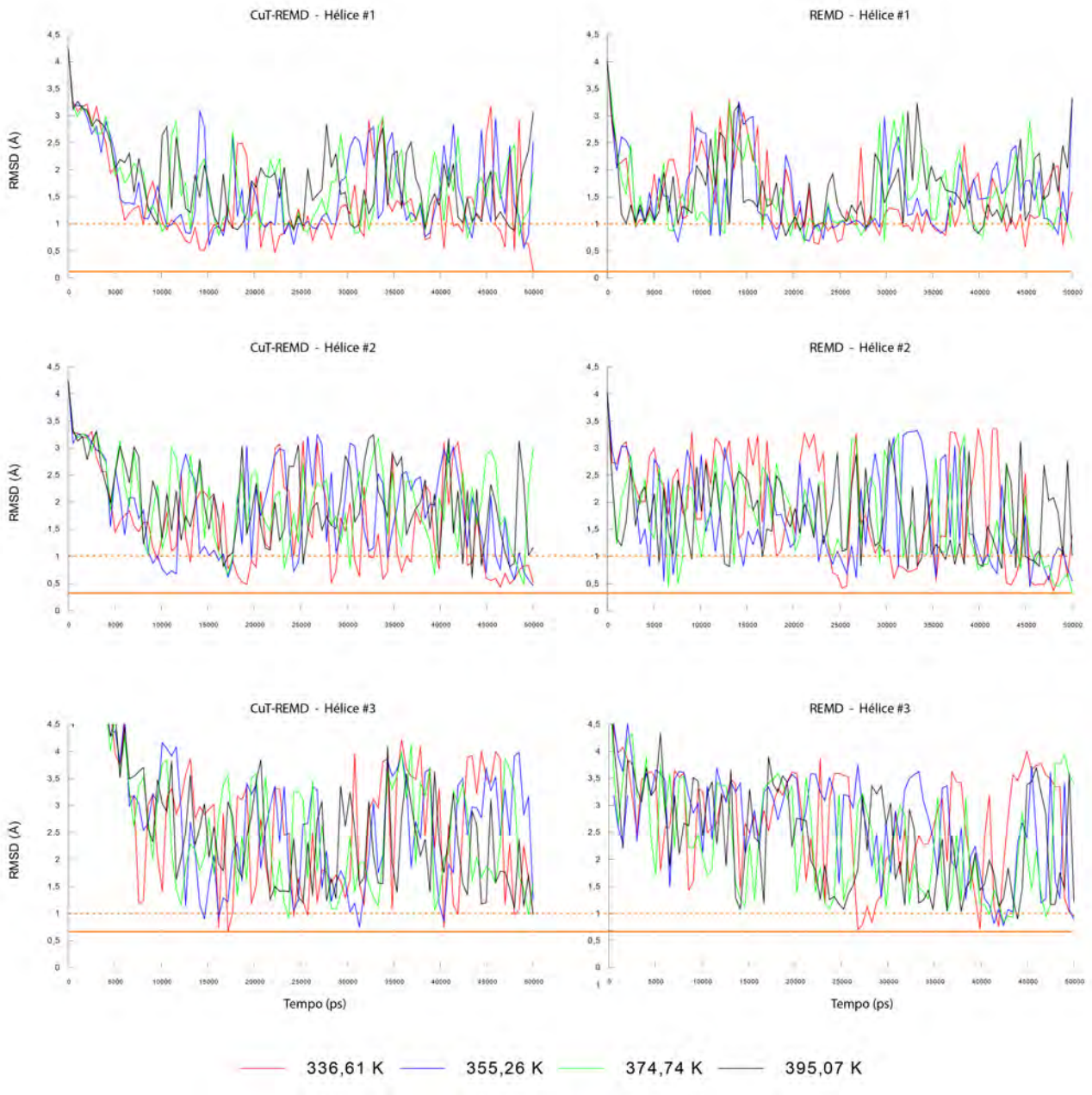

Figura D.1 – Análise comparativa entre CuT-REMD e REMD concencional na formação individual das três hélices que compõem a proteína *villin headpiece*. Resultados para as temperaturas 336,61 K, 355,26 K, 374,74 K e 395,07 K. Em laranja, as linhas pontilhadas e contínuas representam, respectivamente, o limiar de 1 Å e o menor valor de RMSD (considerando a suavização da linha).

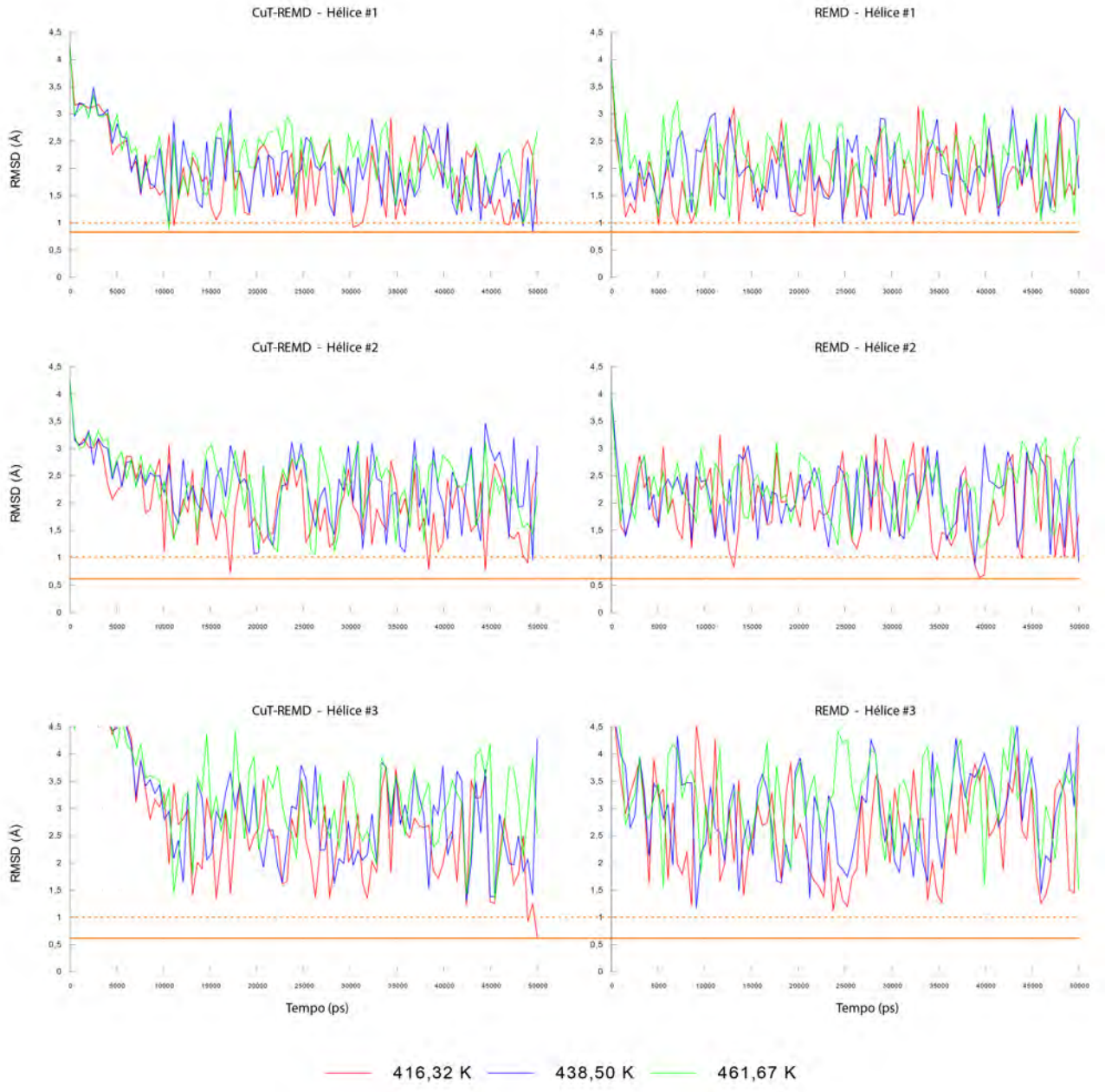

Figura D.2 – Análise comparativa entre CuT-REMD e REMD concencional na formação individual das três hélices que compõem a proteína *villin headpiece*. Resultados para as temperaturas 416,32 K, 438,50 K e 461,67 K. Em laranja, as linhas pontilhadas e contínuas representam, respectivamente, o limiar de 1 Å e o menor valor de RMSD (considerando a suavização da linha).

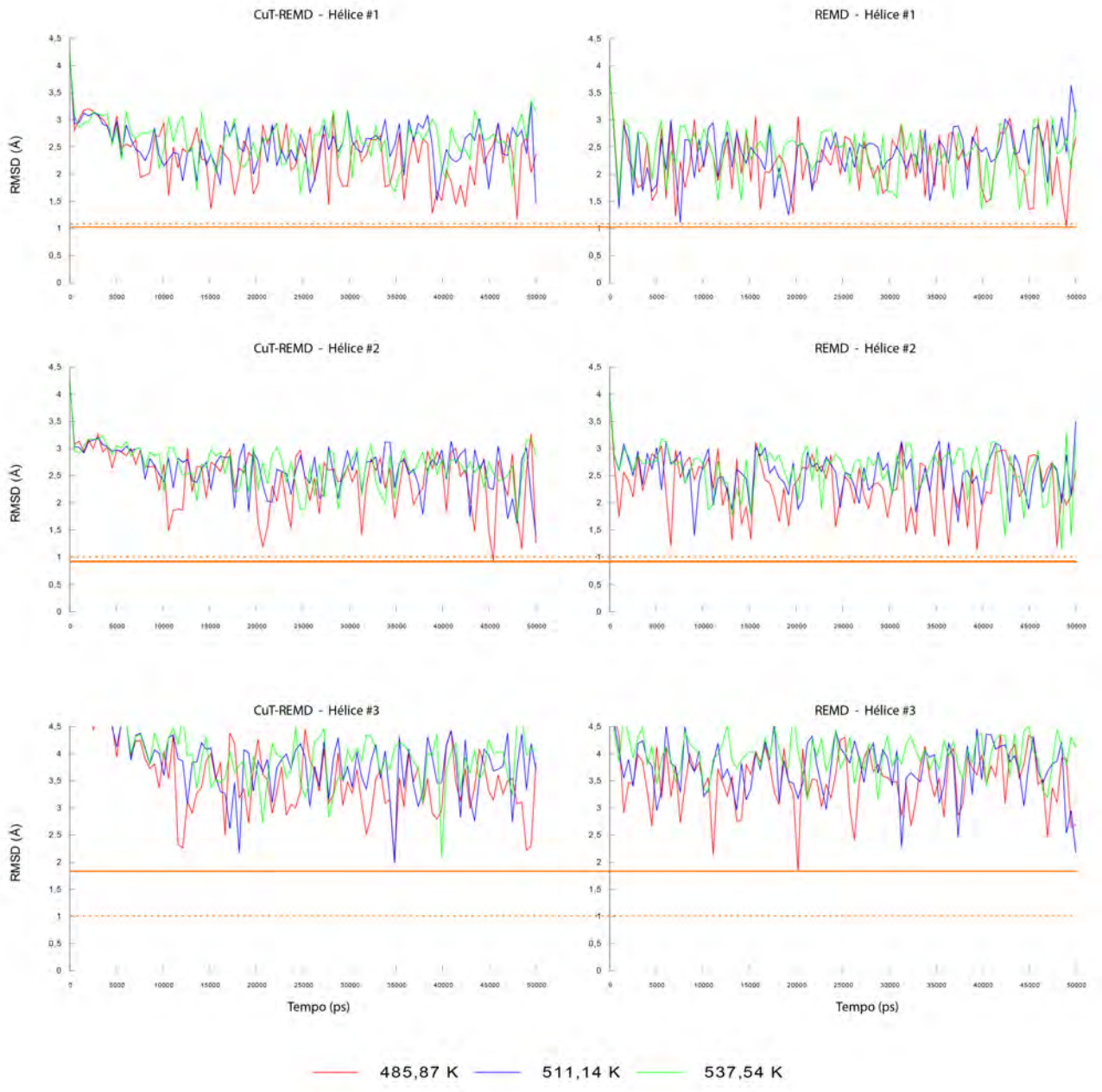

Figura D.3 – Análise comparativa entre CuT-REMD e REMD concencional na formação individual das três hélices que compõem a proteína *villin headpiece*. Resultados para as temperaturas 485,87 K, 511,14 K e 537,54 K. Em laranja, as linhas pontilhadas e contínuas representam, respectivamente, o limiar de 1 Å e o menor valor de RMSD (considerando a suavização da linha).