

PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL

FACULDADE DE INFORMÁTICA  
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIAS DA COMPUTAÇÃO

**PRO-SMART: PREDIÇÃO DE ESTRUTURAS  
TERCIÁRIAS DE PROTEÍNAS UTILIZANDO  
SISTEMAS MULTIAGENTE**

THIAGO LIPINSKI PAES

Dissertação apresentada como  
requisito parcial à obtenção do  
grau de Mestre em Ciência da  
Computação pela Pontifícia  
Universidade Católica do Rio  
Grande do Sul.

Orientador: Prof. Dr. Osmar Norberto de Souza

Porto Alegre  
2013

## Dados Internacionais de Catalogação na Publicação (CIP)

P126p Paes, Thiago Lipinski  
Pro-smart: predição de estruturas terciárias de proteínas  
utilizando sistemas multiagente / Thiago Lipinski Paes. –  
Porto Alegre, 2013.  
133 f.

Diss. (Mestrado em Ciência da Computação) – Faculdade  
de Informática, PUCRS.  
Orientação: Prof. Dr. Osmar Norberto de Souza.

1. Informática. 2. Sistemas Multiagentes. 3. Proteínas.  
4. Biologia Computacional. I. Souza, Osmar Norberto de.  
II. Título.

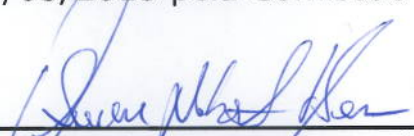
CDD 006.39

**Ficha Catalográfica elaborada pelo  
Setor de Tratamento da Informação da BC-PUCRS**



## TERMO DE APRESENTAÇÃO DE DISSERTAÇÃO DE MESTRADO

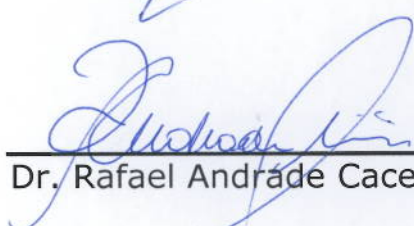
Dissertação intitulada "Pro-Smart: Predição de Estruturas Terciárias de Proteínas Utilizando Sistemas Multiagente" apresentada por Thiago Lipinski Paes como parte dos requisitos para obtenção do grau de Mestre em Ciência da Computação, Bioinformática e Computação Bioinspirada, aprovada em 15/03/2013 pela Comissão Examinadora:

  
Prof. Dr. Osmar Norberto de Souza –  
Orientador

PPGCC/PUCRS

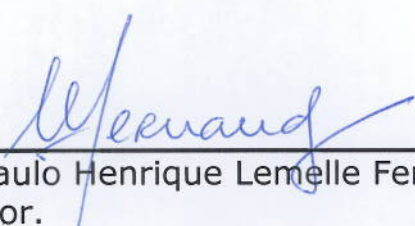
  
Prof. Dr. Rafael Heitor Bordini –

PPGCC/PUCRS

  
Dr. Rafael Andrade Caceres –

Pesquisador - FACIN

Homologada em 27/08/2013, conforme Ata No. 015 pela Comissão Coordenadora.

  
Prof. Dr. Paulo Henrique Lemelle Fernandes  
Coordenador.

**PUCRS**

**Campus Central**

Av. Ipiranga, 6681 – P32– sala 507 – CEP: 90619-900

Fone: (51) 3320-3611 – Fax (51) 3320-3621

E-mail: [ppgcc@pucrs.br](mailto:ppgcc@pucrs.br)

[www.pucrs.br/facin/pos](http://www.pucrs.br/facin/pos)



*Success, acknowledgement, fame, glory.  
Many of us fight for reasons like that, but they don't build a  
good name from one day to the next.  
It is necessary to work hard even if there are stumbles and falls  
It is necessary to overcome obstacles.  
It is necessary to have motivation, to persevere and insist.  
Life is a succession of battles.*

The Gladiator

## AGRADECIMENTOS

Na caminhada destes dois anos de mestrado encontrei muitos obstáculos, porém o número de incentivos que recebi, se contabilizados, somariam um valor muito maior que o dobro ou triplo - foram incontáveis. Venho por meio deste agradecer àqueles que contribuíram para que este trabalho fosse concluído com êxito. Aos amigos forjados durante o tempo em que dividia meu tempo entre a pesquisa e o trabalho como Bernardo Estácio, Eduardo Spies e Samuel Souza: sei que apesar dos pesares aprendemos muito, em pouco tempo e com muito suor, acerca do que queremos e do que não queremos para nossas vidas. Passados os primeiros três semestres de aulas, pesquisas e trabalho, obtive a oportunidade de me dedicar apenas à pesquisa e, a partir do momento que comecei a frequentar o Laboratório de Bioinformática, Modelagem e Simulação de Biosistemas (LABIO), fui muito bem recepcionado pelos seus membros e, em especial, venho agradecer a Luis Fernando Saraiva Macedo Timmers, Mirocem Fernandes de Oliveira e Rafael Andrade Caceres pelos inúmeros conselhos e ensinamentos que obtive, além da amizade formada. Agradeço também ao professor Osmar Norberto de Souza pela paciência e preocupação com meu projeto, além - é claro - de todas as elucidações e norteamentos que obtive nas diversas reuniões que se sucederam. Aprendi e sigo aprendendo muito não somente sobre como ser um cientista mas também sobre assuntos diversos que vão desde ética a gastronomia. Aproveito a oportunidade para agradecer também aos meus amigos das mais diferentes “bandas”, origens e cidades, os quais prefiro não elencar por medo de esquecer algum. Por fim, agradeço às pessoas mais importantes e também principais responsáveis por eu ter chego até aqui: minha família. Pai e mãe, obrigado por tudo, sei que ainda preciso de muito mais que isso para poder agradecer o que vocês me proporcionaram e proporcionam. Mano, Jessica, Gabi, Vini, Tia Marisa, Tio Dudu e Tio Paulinho, vocês também fazem parte desta conquista. Muito Obrigado.

## **PRO-SMART: PREDIÇÃO DE ESTRUTURAS TERCIÁRIAS DE PROTEÍNAS UTILIZANDO SISTEMAS MULTIAGENTE**

### **RESUMO**

Atualmente existem aproximadamente 16 milhões de sequências únicas de proteínas (não redundantes) no GenBank. Entretanto, no PDB, podemos encontrar apenas cerca de 85.000 estruturas tridimensionais (3D) de proteínas das quais apenas 1.393 possuem dobramento SCOP diferentes. Existe então uma grande lacuna entre nossas atuais habilidades no que se trata de gerar sequências de proteínas e nossas habilidades em resolver estruturas 3D de proteínas com novos dobramentos. Essa lacuna vem sendo reduzida com a ajuda da bioinformática estrutural por meio do endereçamento do problema de como uma proteína alcança sua estrutura 3D partindo-se apenas de sua sequência de aminoácidos. Esse é conhecido como o Problema PSP, do inglês *Protein Structure Prediction Problem*. Considerações termodinâmicas apresentadas por Christian Anfinsen e colaboradores em 1973 deram início a uma forma de abordar o problema que hoje é conhecida como a hipótese de Anfinsen, a qual afirma que a estrutura nativa de uma proteína é aquela que minimiza sua energia global livre. Podemos então, tratar o problema PSP como um problema de minimização, tendo em mente ser um problema de complexidade NP-Completo. Neste são utilizados conceitos advindos da inteligência artificial até hoje não muito explorados na bioinformática. Mais especificamente, propomos um arcabouço baseado em uma abordagem *ab initio*, envolvendo um sistema multi-agente hierarquicamente cooperativo e guiado por um esquema baseado no método de Monte Carlo e de Arrefecimento Simulado, a fim de obter-se a otimização de uma função de energia. O sistema multi-agente tem como entrada apenas a sequência de aminoácidos das proteínas. Cada aminoácido é representado por dois agentes: O agente C-Alfa (correspondendo o átomo C alfa) e um agente C-Beta (correspondendo ao centroide da cadeia lateral do aminoácido). Esses agentes aminoácidos interagem entre si. Existem dois outros tipos de agentes: um coordena os agentes Aminoácidos (C-Alfa e C-Beta) e outro coordena o sistema por inteiro. O sistema multi-agente foi criado utilizando a plataforma NetLogo. Um protocolo de clusterização foi desenvolvido para a obtenção da estrutura modelo de cada simulação e os resultados foram comparados com a literatura no que se trata de PSP e multi-agentes e se mostraram promissores.

**Palavras-Chave: Predição de Estrutura 3D de Proteínas, Sistemas Multi-Agente, Método de Monte Carlo, Arrefecimento Simulado.**

## **PRO-SMART: PROTEIN STRUCTURE PREDICTION BY A MULTI-AGENT TOOL**

### **ABSTRACT**

There currently are approximately 16 million of unique (non-redundant) protein sequences in the GenBank. In the PDB, we can only find about 89,000 three-dimensional (3-D) protein structures and only 1,393 different SCOP protein folds. Thus, there is a huge gap between our ability to generate protein sequences and that of solving 3-D structures of proteins with unique, novel folds. This gap has been reduced with the aid from structural bioinformatics by addressing the problem of how a protein reaches its 3-D structure starting only from its amino acid sequence. This is called the protein structure prediction (PSP) problem. Thermodynamics considerations presented by Christian Anfinsen and co-workers in 1973 have it that a protein native structure is the one that minimizes its global free energy. Hence, we can treat the PSP problem as a minimization one within an NP-complete class of computation complexity. Several techniques have been used to predict the 3-D structure of proteins. In this work we supplement these techniques by adding artificial intelligence concepts still not much exploited in bioinformatics. More specifically, we propose a framework, based on an ab initio approach, of a cooperative hierarchical multi-agent system guided by a Simulated Annealing and a Monte Carlo scheme to address the PSP problem. Our multi-agent system has as input the protein amino acid sequence. Amino acids are represented by two agents: The C-Alpha agent (in lieu of the C alpha carbon atom) and the C-Beta agent (in lieu of the side chain centroid). These Amino Acid agents can interact with each other. There are two other agents: one coordinates the Amino Acid agents; the other coordinates the protein system. The multi-agent system was created using the NetLogo platform. A clustering protocol was implemented for obtaining each simulation representant model. The results were compared with published papers regarding similar methodology and the use of Multi-Agent Systems to address the Protein Structure Prediction Problem. We present partial results which are encouraging for mini proteins.

**Keywords: Protein Structure Prediction, Multi-agent Systems, Monte Carlo, Simulated Annealing.**



## LISTA DE FIGURAS

Figura 1 - Estrutura química de dois aminoácidos, onde R representa as cadeias laterais. A estrutura dos aminoácidos tem uma característica comum: a presença de um grupamento carboxilato (COO <sup>-</sup> ) e um grupamento amino (H <sub>3</sub> N <sup>+</sup> ) ligados ao mesmo átomo de carbono (o carbono $\alpha$ ). Os aminoácidos diferem entre si por suas cadeias laterais, também conhecidos como grupos R, que se ligam também a seus respectivos carbonos $\alpha$ . .....	23
Figura 2 - Estrutura secundária de uma proteína. Hélices $\alpha$ e fitas de folhas $\beta$ estão coloridas de vermelho e azul, respectivamente. Voltas e alças são as linhas retas conectando essas ES regulares. Figura obtida de [81]. .....	25
Figura 3 - Estrutura terciária da proteína acilfosfatase de <i>Eschaerichia Coli</i> . PDB ID: 2GV1. Hélices $\alpha$ e a folha $\beta$ , 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 <i>cartoon</i> [38]. .....	25
Figura 4 - Estrutura quaternária da proteína PNP de <i>Mycobacterium Tuberculosis</i> , PDB ID: 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 <i>cartoon</i> [38]. .....	26
Figura 5 - Proteínas possuem um funil no que se trata da 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 [25]. .....	27
Figura 6 - Diagrama ilustrando o problema do mínimo global unidimensional, adaptado de [73]. 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 <sub>1</sub> , chegará até A. Entretanto, se começar em P <sub>2</sub> , logo a direita do ponto de máximo (barreira) existente entre A e B, o calculo nos levará ao mínimo B. ....	32
Figura 7 - Neste exemplo, duas estruturas possuem topologias similares nas regiões do núcleo da proteína. com TM-Score igual a 0,70 e 0,67 respectivamente. Entretanto, as variações nas regiões N e C terminais resultam em uma significativa diferença em termos de RMSD (de 1.9 Å para 10.5 Å). Esta figura foi obtida da Figura 5 de [79]. .....	36
Figura 8 - Mapa de Ramachandran: região mais favorável em vermelho, região permitida em amarelo, região ainda aceitável em amarelo claro e região não permitida em branco. O canto superior em vermelho trata-se de região favorável para folhas $\beta$ e no centro direito e esquerdo em vermelho para hélices $\alpha$ , respectivamente. Modelo adotado por Thornton e colaboradores [44]. .....	37
Figura 9 - Definições dos estados conformacionais no mapa de Ramachandran segundo A. V. Efimov [46]. .....	38
Figura 10 - Exemplo de modelo com representação reduzida: No modelo proposto por Berrera <i>et. al</i> em [10] cada aminoácido é composto por apenas duas esferas, uma representa seu carbono alfa e outra representa o centróide de sua cadeia lateral. A distância entre dois átomos de carbono alfa consecutivos foi definida como 3.8 Å. Imagem criada pelo software ChemDraw [36]. .....	44
Figura 11 - <i>Bend angle</i> – Figura adaptada de [14]. .....	44
Figura 12 - <i>Torsion angle</i> – Figura adaptada de [14]. .....	44
Figura 13 - Exemplo do arquivo de clusterização gerado. O usuário deve escolher a estrutura correspondente ao centro do maior cluster encontrado. ....	50
Figura 14 - Esquema geral do PRO-SMART. Enquanto dentro do NetLogo os agentes cooperam para alcançar a melhor conformação, scripts e banco de dados, de fora do NetLogo, contribuem no processamento de informações e cálculos. ....	51

Figura 15 - Hierarquia dos agentes. Agentes de maior nível controlam os de menor. O número de agentes C-Alfa e C-Beta depende do tamanho da cadeia protéica.....	52
Figura 16 - Estratégia de movimentação - Agentes C-Alfa: Movimentam-se ao longo de uma circunferência que mantém constante a distância entre C-Alfas adjacentes.....	55
Figura 17 - Estratégia de movimentação - Agentes C-Alfa: Quando estão no início ou fim da sequência de aminoácidos, o agente verifica a distância D dele para o agente adjacente e é criada uma esfera de raio D na volta do agente adjacente. Posteriormente o agente escolhe sua nova localização em algum ponto (com igual probabilidade) da superfície da esfera.....	56
Figura 18 - Pseudocódigo, funcionamento dos agentes C-Alfa e C-Beta.....	57
Figura 19 - Movimento <i>Crankshaft</i> : Primeiramente escolhe dois aminoácidos A e B da cadeia (a uma distancia de três aminoácidos). Posteriormente aplica rotação aos aminoácidos que separam A e B escolhendo um ângulo aleatório, tendo como eixo a reta AB.....	58
Figura 20 - Movimento <i>Pivot</i> : Escolhe um ponto (aminoácido pivô) e gira uma parte da cadeia em torno desse ponto, utilizando um ângulo aleatório.....	58
Figura 21 - Pseudocódigo, funcionamento do agente Diretor.....	59
Figura 22 - Pseudocódigo, funcionamento do agente Ambiente.....	60
Figura 23 - Interface I. Parte da interface na qual o usuário configura a proteína a ser modelada e controla a execução da simulação.....	64
Figura 24 - Visualização 3D. Por intermédio de uma janela separada é possível verificar a conformação atual da proteína que está sendo modelada e, em tempo real, acompanhar as modificações estruturais que estão ocorrendo em virtude das movimentações dos agentes.....	65
Figura 25 - Interface II. Parte da interface na qual o usuário pode visualizar em tempo real gráficos referentes ao comportamento da simulação.....	65
Figura 26 - Interface III. Parte da interface na qual o usuário pode configurar a estratégia de busca da conformação de menor energia, assim como o tipo de movimentação que os agentes devem adotar.....	66
Figura 27 - Exemplos de gráficos gerados automaticamente pelo PRO-SMART. Em (A) podemos verificar a relação de Energia x RMSD ao longo da simulação. Em (B) podemos notar as flutuações de RMSD do início ao fim da simulação e, em (C), (D), (E) e (F) as flutuações da Energia durante a simulação, dando ênfase ao tamanho das mudanças em termos de energia em diferentes fases da simulação.....	69
Figura 28 - Estruturas 3D das proteínas alvo de simulação.....	72
Figura 29 - Diagrama de Ramachandran contendo a análise estereoquímica das estruturas escolhidas como representantes das simulações 7, 8 e 9. Proteína de PDB ID: 1EDP.....	75
Figura 30 - Estutura experimental (em cinza) e estrutura predita (em preto) pela simulação 9, proteína de PDB ID: 1EDP.....	76
Figura 31 - Estutura experimental (em cinza) e estrutura predita (em preto) pela simulação 10, proteína de PDB ID: 1PG1. Na sobreposição apenas dos resíduos que compõem as fitas temos: 1,29 Å para a primeira fita e 0,92 Å para a segunda.....	77
Figura 32 - Diagrama de Ramachandram do modelo obtidos na simulação 10. Proteína PDB ID: 1PG1. Cooperação via estrutura secundária desativada. Triângulos simbolizam glicinas.....	78
Figura 33 - Diagrama de Ramachandram dos modelos obtidos nas simulações 4 e 10. Proteína PDB ID: 1LE0. Cooperação via estrutura secundária desativada.....	79
Figura 34 - Estutura experimental (em cinza) e estrutura predita (em preto) pela simulação 10, proteína de PDB ID: 1LE0.....	80
Figura 35 - Diagrama de Ramachandram dos modelos obtidos nas simulações 4 e 7. Proteína PDB ID: 1ZDD. Cooperação via estrutura secundária desativada.....	81
Figura 36 - Estutura experimental (em cinza) e estrutura predita (em preto) pela simulação 7, proteína de PDB ID: 1ZDD.....	82

Figura 37 - Diagrama de Ramachandram dos modelos obtidos nas simulações 1, 2 e 6. Proteína PDB ID: 1KVG. Cooperação via estrutura secundária desativada. ....	82
Figura 38 - Estrutura experimental (em cinza) e estrutura predita (em preto) pela simulação 3, proteína de PDB ID: 1KVG. ....	84
Figura 39 - Diagrama de Ramachandram dos modelos obtidos nas simulações 2 e 3. Proteína PDB ID: 1LE3. Cooperação via estrutura secundária desativada. ....	84
Figura 40 - Estrutura experimental (em cinza) e estrutura predita (em preto) pela simulação 3, proteína de PDB ID: 1LE3. Cooperação via estrutura secundária desativada. ....	86
Figura 41 – Diagrama de Ramachandram dos modelos obtidos nas simulações 2, 5 e 8. Proteína PDB ID: 1VII. Cooperação via estrutura secundária desativada. ....	87
Figura 42 - Estrutura experimental (em cinza) e estrutura predita (em preto) pela simulação 5, proteína de PDB ID: 1VII. Cooperação via estrutura secundária desativada. O RMSD levando-se em conta uma sobreposição somente os 10 resíduos que formam a terceira hélice tem o valor de 3,70 Å. ....	88
Figura 43 - Diagrama de Ramachandran dos modelos obtidos nas simulações 1, 2, 3 e 7. Proteína de PDB ID: 2GP8. Cooperação via estrutura secundária desativada. ....	89
Figura 44 - Estrutura experimental (em cinza) e estrutura predita (em preto) pela simulação 1, proteína de PDB ID: 2GP8. Cooperação via estrutura secundária desativada. ....	90
Figura 45 - Diagrama de Ramachandram dos modelos obtidos nas simulações 4, 8 e 9. Proteína PDB ID: 1ED0. Cooperação via estrutura secundária desativada. Triângulos simbolizam glicinas. ....	90
Figura 46 - Estruturas experimental (em cinza e amarelo) e estruturas preditas (em preto) pela simulação 4, proteína de PDB ID: 1ED0. Cooperação via estrutura secundária desativada. Em A temos a melhor sobreposição entre a estrutura experimental e a predita. Em B a sobreposição apenas dos resíduos que compõem a primeira hélice e, em C, a sobreposição apenas dos resíduos que compõem a segunda hélice. ....	92
Figura 47 - Diagrama de Ramachandram dos modelos obtidos nas simulações 2, 6, 9 e 10. Proteína PDB ID: 1EDP. Cooperação via estruturas secundárias ativada, com peso = 1. ....	96
Figura 48 - Estrutura experimental (em cinza) e estrutura predita (em preto) pela simulação 6, proteína de PDB ID: 1EDP. Cooperação via estrutura secundária ativada com peso = 1. Pontes dissulfídicas destacadas em amarelo. ....	97
Figura 49 - Diagrama de Ramachandram dos modelos obtidos nas simulações 4 e 10. Proteína PDB ID: 1VII. Cooperação via estruturas secundárias ativada, com peso = 1. ....	98
Figura 50 - Estrutura experimental (em cinza) e estrutura predita (em preto) pela simulação 6, proteína de PDB ID: 1VII. Cooperação via estrutura secundária ativada com peso = 1. ....	99
Figura 51 - Diagrama de Ramachandram dos modelos obtidos na simulação 1. Proteína PDB ID: 1ZDD. Cooperação via estruturas secundárias ativada, com peso = 1. ....	100
Figura 52 - Estrutura experimental (em cinza) e estrutura predita (em preto) pela simulação 1, proteína de PDB ID: 1ZDD. Cooperação via estrutura secundária ativada com peso = 1. ....	101
Novamente tendo como alvo a proteína 2GP8 de estrutura já descrita anteriormente, com a adição da cooperação via estruturas secundárias com peso equivalente a 1, os modelos originários das simulações 1, 8 e 10 não demonstraram significativa melhora, alcançando resultados do mesmo calão em termos de RMSD, entretanto, no que se trata de estereoquímica, foi possível verificar melhorias tanto no que se trata do número de resíduos em regiões não permitidas quanto em termos da disposição dos resíduos pertencentes às estruturas secundárias. A seguir está disposta a estrutura 3D do modelo 186, originário da simulação 1, o qual possui mais resíduos dispostos na região de hélices $\alpha$ , segundo Figura 53. Figura 53 - Diagrama de Ramachandram dos modelos obtidos nas simulações 1, 8 e 10. Proteína PDB ID: 2GP8. Cooperação via estruturas secundárias ativada, com peso = 1. ...	102
Figura 54 - Estrutura experimental (em cinza) e estrutura predita (em preto) pela simulação 1, proteína de PDB ID: 2GP8. Cooperação via estrutura secundária ativada com peso = 1. ....	103

Figura 55 - Diagrama de Ramachandram do modelo obtido na simulação 10. Proteína PDB ID: 1EDP. Cooperação via estruturas secundárias ativada, com peso = energia/10.....	104
Figura 56 - Estrutura experimental (em cinza) e estrutura predita (em preto) pela simulação 1, proteína de PDB ID: 1EDP. Cooperação via estrutura secundária ativada com peso = energia/10.. Pontes bissulfídricas em amarelo.....	105
Figura 57 - Diagrama de Ramachandram dos modelos obtidos nas simulações 3 e 6. Proteína PDB ID: 1ZDD. Cooperação via estruturas secundárias ativada, com peso = energia/10.....	106
Figura 58 - Diagrama de Ramachandram dos modelos obtidos nas simulações 3, 4 e 6. Proteína PDB ID: 1VII. Cooperação via estruturas secundárias ativada, com peso = energia/10.....	107
Figura 59 - Estrutura experimental (em cinza) e estrutura predita (em preto) pela simulação 6, proteína de PDB ID: 1ZDD. Cooperação via estrutura secundária ativada com peso = energia/10.....	108
Figura 60 - Estrutura experimental (em cinza) e estrutura predita (em preto) pela simulação 6, proteína de PDB ID: 1VII. Cooperação via estrutura secundária ativada com peso = energia/10.....	109
Figura 61 - Diagrama de Ramachandram dos modelos obtidos nas simulações 1 e 7. Proteína PDB ID: 2GP8. Cooperação via estruturas secundárias ativada, com peso = energia/10.....	110
Figura 62 - Estrutura experimental (em cinza) e estrutura predita (em preto) pela simulação 1, proteína de PDB ID: 2GP8. Cooperação via estrutura secundária ativada com peso = energia/10.....	111
Figura 63 - Estrutura experimental (em cinza), estrutura predita (em preto) pela simulação 1 e estrutura refinada (em azul), proteína de PDB ID: 1EDP. Cooperação via estrutura secundária ativada com peso = energia/10. Pontes bissulfídricas em amarelo.....	119
Figura 64 - Mapa de Ramachandran para a estrutura refinada proveniente da simulação 1, proteína de PDB ID: 1EDP. Cooperação via estrutura secundária ativada com peso =energia/10.....	120

## LISTA DE TABELAS

Tabela 1 - Conjunto de Proteínas alvo de simulações, junto de seu PDB ID, do artigo publicado que a descreve, número de aminoácidos que possui e classe. ....	72
Tabela 2 - Conjunto de variáveis e funcionalidades configuradas nas simulações com cooperação via estruturas secundária desativada. ....	73
Tabela 3 - Resultados obtidos para as diferentes medidas de similaridade em relação à proteína de PDB ID: 1EDP, sem cooperação via estrutura secundária. ....	74
Tabela 4 - Resultados obtidos para as diferentes medidas de similaridade em relação à proteína de PDB ID: 1PG1, sem cooperação via estrutura secundária. ....	77
Tabela 5 - Resultados obtidos para as diferentes medidas de similaridade em relação à proteína de PDB ID: 1LE0, sem cooperação via estrutura secundária. ....	78
Tabela 6 - Resultados obtidos para as diferentes medidas de similaridade em relação à proteína de PDB ID: 1ZDD, sem cooperação via estrutura secundária. ....	80
Tabela 7 - Resultados obtidos para as diferentes medidas de similaridade em relação à proteína de PDB ID: 1KVG, sem cooperação via estrutura secundária. ....	83
Tabela 8 - Resultados obtidos para as diferentes medidas de similaridade em relação à proteína de PDB ID: 1LE3, sem cooperação via estrutura secundária. ....	85
Tabela 9 - Resultados obtidos para as diferentes medidas de similaridade em relação à proteína de PDB ID: 1VII, sem cooperação via estrutura secundária. ....	86
Tabela 10 - Resultados obtidos para as diferentes medidas de similaridade em relação à proteína de PDB ID: 2GP8, sem cooperação via estrutura secundária. ....	88
Tabela 11 - Resultados obtidos para as diferentes medidas de similaridade em relação à proteína de PDB ID: 1ED0, sem cooperação via estrutura secundária. ....	91
Tabela 12 - Valores (em Å) Comparação em termos de RMSD entre o PROSMART e o <i>framework</i> de Bortolussi <i>et al.</i> Cooperação via estrutura secundária desativada. ....	93
Tabela 13 - Comparação em termos de energia entre o PROSMART e o <i>framework</i> de Bortolussi <i>et al.</i> Cooperação via estrutura secundária desativada. ....	93
Tabela 14 - Conjunto de atributos e funcionalidades configuradas nas simulações com cooperação via estruturas secundárias ativada com peso=1. ....	94
Tabela 15 - Sequências referentes à estrutura secundária das proteínas simuladas. Única informação utilizada pelo PRO-SMART. A letra C simboliza resíduos que, segundo o preditor, não formam estruturas secundárias (estes resíduos são chamados <i>coils</i> ), a letra H simboliza resíduos que, segundo o preditor, formam hélices. ....	95
Tabela 16 - Simulações com cooperação via estrutura secundária. Proteína de PDB ID: 1EDP. Peso = 1. ....	96
Tabela 17 - Simulações com cooperação via estrutura secundária. Proteína de PDB ID: 1VII. Peso = 1. ....	98
Tabela 18 - Simulações com cooperação via estrutura secundária. Proteína de PDB ID: 1ZDD. Peso = 1. ....	100
Tabela 19 - Simulações com cooperação via estrutura secundária. Proteína de PDB ID: 2GP8. Peso = 1. ....	102
Tabela 20 - Simulações com cooperação via estrutura secundária. Proteína de PDB ID: 1EDP. Peso = energia/10. ....	104
Tabela 21 - Simulações com cooperação via estrutura secundária. Proteína de PDB ID: 1ZDD. Peso = energia/10. ....	106

Tabela 22 - Simulações com cooperação via estrutura secundária. Proteína de PDB ID: 1VII. Peso = energia/10.....	108
Tabela 23 - Simulações com cooperação via estrutura secundária. Proteína de PDB ID: 2GP8. Peso = energia/10.....	109
Tabela 24 – Valores (em Å) da comparação em termos de RMSD entre o PROSMART e o <i>framework</i> de Bortolussi <i>et al.</i> Cooperação via estruturas secundárias ativada com peso = 1.....	111
Tabela 25 - Comparação em termos de energia entre o PROSMART e o <i>framework</i> de Bortolussi <i>et al.</i> Cooperação via estruturas secundárias ativada com peso = 1.....	111
Tabela 26 - Valores (em Å) da comparação em termos de RMSD entre o PROSMART e o <i>framework</i> de Bortolussi <i>et al.</i> Cooperação via estruturas secundárias ativada com peso estipulado em 1/10 da energia.....	111
Tabela 27 - Comparação em termos de energia entre o PROSMART e o <i>framework</i> de Bortolussi <i>et al.</i> Cooperação via estruturas secundárias ativada com peso estipulado em 1/10 da energia.....	112
Tabela 28 - Desempenho dos diferentes tipos de cooperação via estrutura secundária. Proteína de PDB ID: 1EDP. Melhores resultados em negrito.....	114
Tabela 29 - Desempenho dos diferentes tipos de cooperação via estrutura secundária. Proteína de PDB ID: 1VII. Melhores resultados em negrito.....	115
Tabela 30 – Desempenho dos diferentes tipos de cooperação via estrutura secundária. Proteína de PDB ID: 1ZDD. Melhores resultados em negrito.....	115
Tabela 31 – Desempenho dos diferentes tipos de cooperação via estrutura secundária. Proteína de PDB ID: 2GP8. Melhores resultados em negrito.....	116

## LISTA DE SIGLAS

3D	Tridimensional
AAMAS	<i>International Joint Conference on Autonomous Agents and Multiagent Systems</i>
AG	Algoritmo Genético
BE	Bioinformática Estrutural
CG	<i>Coarse-grained</i>
DM	Dinâmica Molecular
ER	Estruturas Regulares
ES	Estruturas Secundárias
IA	Inteligência Artificial
K	Temperatura Kelvin
MABS	<i>Multi-Agent-Based Simulation</i>
MAS-BIOMED	<i>Multi-Agent Systems for Medicine, Computational Biology, and Bioinformatics</i>
MC	Monte Carlo
PDB	<i>Protein Data Bank</i>
PSP	<i>Protein Structure Prediction</i>
RMSD	<i>Root Mean Square Deviation</i> – Desvio médio quadrático
SMA	Sistemas Multi-agentes





## SUMÁRIO

1. INTRODUÇÃO .....	19
1.1 Motivação .....	19
1.2 Objetivos .....	20
1.2.1 Objetivo Geral .....	20
1.2.2 Objetivos Específicos .....	21
1.3 Metodologia .....	21
1.4 Organização da Dissertação .....	22
2. FUNDAMENTAÇÃO TEÓRICA .....	23
2.1 Proteínas e sua composição .....	23
2.2 Predição de Estruturas 3D de Proteínas .....	26
2.3 Métodos Computacionais para Predição de Estruturas 3D de Proteínas .....	28
2.3.2 Reconhecimento de Padrões ou <i>Folding Recognition</i> .....	28
2.3.3 Predição <i>Ab initio</i> e <i>De novo</i> .....	28
2.4 Simulação Baseada em Sistemas Multi-agentes .....	29
2.4.1 Sistemas Multi-agentes .....	29
2.4.2 Simulação Computacional .....	29
2.5 Simulação Termodinâmica e Otimização Global .....	31
2.6 Critérios de Avaliação .....	34
2.6.1 RMSD .....	34
2.6.2 MaxSub .....	36
2.6.3 GDT .....	36
2.6.4 Diagrama de Ramachandran .....	37
3. TRABALHOS RELACIONADOS .....	39
3.1 Mapeamento Sistemático: Protocolo .....	39
3.2 Trabalhos Encontrados .....	39
4. PRO-SMART: ASPECTOS CONCEITUAIS .....	43
4.1 Nível de Abstração .....	43
4.2 Função de Energia .....	45
4.3 Cooperação .....	47
4.4 Clusterização de Estruturas .....	49
5. PRO-SMART: IMPLEMENTAÇÃO .....	51
5.1 Esquema Geral .....	51
5.2 Agentes .....	52

5.2.1 Agentes tipo C-Alfa – Nível 1.....	52
5.2.2 Agentes tipo C-Beta – Nível 1 .....	56
5.2.3 Agente tipo Diretor – Nível 2.....	57
5.2.4 Agente tipo Ambiente – Nível 3 .....	59
5.3 Requisitos.....	61
5.4 Interface.....	61
5.5 Execução .....	66
6. RESULTADOS.....	71
6.1 Conjunto de Proteínas Alvo .....	71
6.2 Simulações Sem Cooperação via Estruturas Secundárias.....	73
6.2.1 Configuração .....	73
6.2.2 Resultados Obtidos.....	73
6.2.3 Comparação.....	92
6.3 Simulações com Cooperação via Estruturas Secundária.....	93
6.3.1 Configuração .....	93
6.3.2 Resultados Obtidos.....	95
6.3.3 Comparação.....	111
6.4 Desempenho da Cooperação via Estruturas Secundárias.....	113
7. CONSIDERAÇÕES FINAIS.....	117
7.1 Principais Contribuições .....	118
7.2 Trabalhos Futuros.....	118
REFERÊNCIAS.....	121
APÊNDICE A .....	127
APÊNDICE B .....	133

## 1. INTRODUÇÃO

Proteínas são polímeros formados por sequências de 20 diferentes resíduos (19 de aminoácidos e um de iminoácido) que, ao interagir físico-quimicamente, formam uma estrutura tridimensional (3D) única [47]. O aumento na velocidade de geração de sequências de DNA nos últimos tornou disponível um grande número de sequências proteicas (também conhecidas como estruturas primárias de proteínas). Estas sequências, traduzidas do DNA, podem ser obtidas do *GenBank* [8]. As estruturas 3D de proteínas, por sua vez, podem ser obtidas do *Protein Data Bank* ou PDB [9]. Atualmente, há cerca de 16 milhões de sequências de proteínas únicas, não redundantes. Entretanto, no PDB, encontramos aproximadamente 89 mil estruturas 3D de proteínas. Ao eliminarmos a redundância, filtrando estruturas muito similares, obtemos apenas 1.393 formas de dobramento espaciais diferentes (*PDB Statistics*, acessado em 5 de abril de 2013). Portanto, existe uma enorme lacuna entre a nossa capacidade de produzir sequências proteicas e a nossa capacidade de resolver estruturas 3D de novas proteínas, com dobramentos diferentes dos já existentes [59]. Esta lacuna tem sido reduzida com o auxílio da bioinformática estrutural.

A bioinformática estrutural (BE) conceitua a biologia em termos de moléculas, no sentido físico-químico, e, por meio da informática, empregando estatística, matemática, biologia, física, engenharia biomédica, genética, bioquímica, química aliada à ciência da computação, permite-nos armazenar, organizar e compreender, em larga escala, esta explosão de dados biológicos [53]. É na BE que se procura, entre outros aspectos, entender como uma proteína atinge a sua estrutura terciária ou 3D a partir apenas da sua sequência primária. Este é denominado o problema da predição da estrutura 3D de proteínas ou problema PSP (do inglês *Protein Structure Prediction* ou *PSP Problem*). Existem várias técnicas para predição de estruturas 3D de proteínas. Este trabalho está focado em uma abordagem *ab initio* baseada em sistemas multi-agente e no construto descrito por Bortolussi *et al.* [12], [13]. Este construto ou modelo multi-agente utiliza como entrada um arquivo contendo a sequência de aminoácidos da proteína. Ela é carregada pelo sistema multi-agente que trata cada aminoácido como um agente diferente e simula a interação entre eles.

### 1.1 Motivação

Uma recente revisão sobre predição de estrutura de proteínas, publicada na revista *Science* (dia 23 de novembro de 2012) referencia justamente os 50 anos do “nascimento de um dos grandes desafios da ciência básica”, o problema PSP. A revisão ainda enfatiza os

avanços consideráveis obtidos no entendimento do problema e destaca o considerável valor em termos da pesquisa de métodos precisos para a predição de estruturas a partir de seqüências [25]. O problema PSP surgiu na década de 60 e até hoje sua solução continua sendo uma das principais pendências da biologia molecular. 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. 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 [26]. Considerando 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. Dado que um tratamento envolvendo predição 3D de uma proteína requer, por definição, considerar a adaptabilidade do sistema em tempo real (no caso a sucessiva modificação de parâmetros como a temperatura do banho térmico em que a proteína está inserida), torna-se evidente a necessidade de uma simulação que possua tal característica, ou seja, experimentos *in virtuo* (experimentos a serem executados por simulação computacional, porém com a característica adicional de serem suscetíveis a perturbações durante a execução, *on the fly*). Enquanto é uma propriedade típica de toda simulação computacional (experimentos *in silico*) a fácil modificação de parâmetros que a caracterizam, é uma propriedade específica de simulações baseadas em sistemas multi-agente a fácil modificação da estrutura do experimento em si, originando experimentos *in virtuo* [3], [71]. É importante observar ainda que experimentos *in virtuo* podem ser realizados sem a utilização de agentes, entretanto, a utilização de sistemas multi-agente provê um suporte tecnológico particularmente adequado para tarefas desta natureza [3].

## 1.2 Objetivos

### 1.2.1 Objetivo Geral

O objetivo geral do trabalho foi a criação de um ambiente baseado em multi-agentes capaz de, a partir de apenas a estrutura primária ou seqüência de aminoácidos de uma

proteína, chegar a estruturas tridimensionais potencialmente capazes de representar sua estrutura nativa.

### 1.2.2 Objetivos Específicos

Os objetivos específicos dessa dissertação foram:

- Definir uma Função de Energia a ser utilizada para verificar o potencial da ferramenta.
- Definir as dimensões do ambiente a ser modelado, o limite de tamanho das proteínas.
- Escolher a coleção de proteínas a ser testada.
- Definir a quantidade de tipos diferentes de agentes contidos no sistema e a relação hierárquica existente entre eles.
- Definir a heurística de otimização e controle do ambiente / Parâmetros do arrefecimento simulado.
- Analisar comparativamente os resultados obtidos em relação à literatura.

## 1.3 Metodologia

A metodologia empregada para a realização do trabalho tem 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 [5]. Para isso foi utilizada um função de energia com termos baseados em leis físicas e químicas de interação entre aminoácidos, os quais são representados cada um como um agente inteligente autônomo disposto em um ambiente multia-gente criado no software NetLogo. Os agentes, seguindo a nomenclatura de Garcia e Sichman [32], possuem interação de simbiose, cooperando para alcançar um objetivo comum. É utilizada uma abordagem definida por diferentes níveis para classificar diferentes tipos de agente em uma hierarquia onde aqueles pertencentes a posições mais altas de hierarquia podem controlar aqueles que estiverem em posições de hierarquia mais baixas [63]. As alterações conformacionais da proteína são avaliadas pelo método de Monte Carlo. Sendo nosso problema relativo à otimização de uma função de energia, é preciso levar em conta o tamanho das barreiras de energia ou mínimos locais passíveis de superação nos diferentes momentos da simulação e, para que o sistema tenha maior êxito na tarefa de alcançar mínimos globais, foi utilizada uma técnica chamada arrefecimento simulado (seção 2.5)

Foi criada uma coleção de proteínas alvo de simulação com base naquelas utilizadas por trabalhos já publicados. Os resultados foram analisados com base nos critérios utilizados no CASP10 (*Critical Assessment of Techniques for Protein Structure Prediction*), uma competição internacional que ocorre a cada dois anos com o objetivo de mensurar objetivamente as capacidades atuais da predição de estruturas de proteínas. Os resultados obtidos foram confrontados com os obtidos por Bortolussi *et al.* e seu arcabouço.

#### **1.4 Organização da Dissertação**

Esta dissertação está organizada em sete capítulos:

- O primeiro capítulo conta com a introdução ao problema, junto da motivação para seu tratamento, a metodologia utilizada e os objetivos da dissertação;
- O segundo capítulo contém a fundamentação teórica necessária para o entendimento do trabalho. É onde o conceito de proteínas é introduzido juntamente com o problema da predição de estruturas 3D e os métodos utilizados para o tratamento do problema, além de introduções à sistemas multi-agentes e métodos de otimização termodinâmica;
- No terceiro capítulo estão dispostos os trabalhos relacionados, com ênfase para o trabalho que serviu de fundamento para parte do PRO-SMART;
- Nos capítulos 4 e 5 o PRO-SMART é apresentado, primeiramente com seus aspectos conceituais e posteriormente com os aspectos de implementação, onde os diferentes tipos de agentes são introduzidos;
- Em seguida, no capítulo 6, os resultados alcançados são expostos, sempre que possível comparando-os com trabalhos anteriores e
- Por fim, no último capítulo, de número 7, são feitas as considerações finais, com as principais contribuições da dissertação seguido dos trabalhos futuros.

## 2. FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta uma base teórica referente aos principais conceitos que envolvem o trabalho. Primeiramente é abordado o conceito de proteínas, seguido do problema alvo de nossa abordagem, o problema da predição estrutural de proteínas. Em seguida são expostos os principais métodos computacionais de predição de proteínas, os conceitos referentes a simulações baseadas em sistemas multi-agente seguidos de conceitos relacionados à simulação termodinâmica e otimização global (os quais ajudam a reger o modelo multi-agente) e, por fim, os critérios de avaliação utilizados pelo 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 20 resíduos (19 aminoácidos e 1 iminoácido) que se ligam covalentemente em uma sequência linear [15], [47].

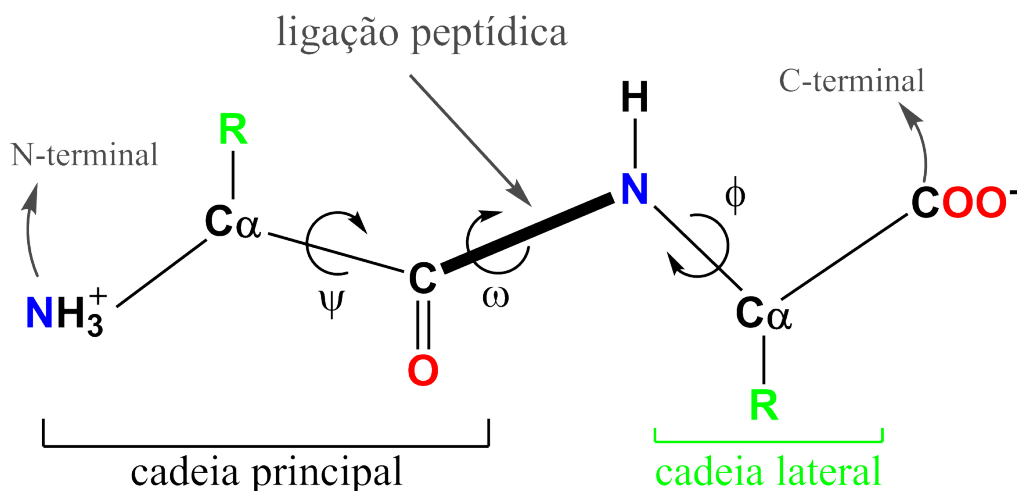


Figura 1 - Estrutura química de dois aminoácidos, onde R representa as cadeias laterais. A estrutura dos aminoácidos tem uma característica comum: a presença de um grupamento carboxilato (COO<sup>-</sup>) e um grupamento amino (H<sub>3</sub>N<sup>+</sup>) ligados ao mesmo átomo de carbono (o carbono  $\alpha$ ). Os aminoácidos diferem entre si por suas cadeias laterais, também conhecidos como grupos R, que se ligam também a seus respectivos carbonos  $\alpha$ .

Os aminoácidos diferem entre si por suas cadeias laterais, também conhecidos como grupos R, que se ligam também a seus respectivos carbonos  $\alpha$ . Os grupos R variam em se tratando de estrutura e carga elétrica, além de tamanho, podendo contar com de 1 a 18 átomos [15]. Um peptídeo é uma molécula composta por dois ou mais aminoácidos unidos por uma ligação peptídica (Figura 1) e possui três ângulos de torção em sua cadeia principal, chamados phi ( $\phi$ ), psi ( $\psi$ ) e ômega ( $\omega$ ).

Os grupos peptídicos, com poucas exceções, assumem uma configuração tal que os átomos  $C\alpha$  sucessivos ficam em lados opostos da ligação peptídica que os une. Essa e outras observações indicam que o esqueleto de uma proteína compõe-se de uma sequência de grupos peptídicos planares rígidos e ligados [75]. 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- $C\alpha$  e  $C\alpha$ -C (exceto prolina). O enovelamento de uma proteína é dado pelos ângulos diedrais  $\phi$  (phi) e  $\psi$  (psi) dessas ligações e pelo ângulo  $\omega$  (ômega) de rotação em torno da ligação peptídica [47]. Os ângulos  $\phi$ ,  $\psi$  e  $\omega$  da cadeia principal representam de forma única a conformação de uma proteína. Das combinações entre os 20 tipos de resíduos uma gama imensa de proteínas pode ser formada e assim, diferentes organismos podem então fazer uso de diferentes produtos. Algumas proteínas realmente contêm resíduos que não os 20 acima referidos, todavia esses são produzidos por modificações químicas pós-síntese ou pela introdução de uma selenocisteína durante a tradução, como na glutathione peroxidase [46]. Entre a gama de proteínas existentes podemos 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 [15].

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) 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 hélice  $\alpha$  e folhas  $\beta$ .





Figura 2 - Estrutura secundária de uma proteína. Hélices  $\alpha$  e fitas de folhas  $\beta$  estão coloridas de vermelho e azul, respectivamente. Voltas e alças são as linhas retas conectando essas ES regulares. Figura obtida de [81].

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 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 4) [81].

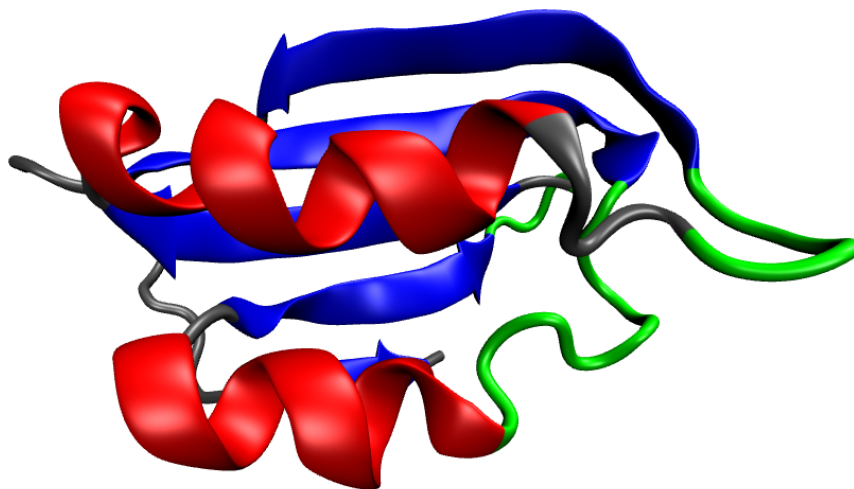


Figura 3 - Estrutura terciária da proteína acilfosfatase de *Escherichia Coli*. PDB ID: 2GV1. Hélices  $\alpha$  e a folha  $\beta$ , contendo cinco fitas, estão coloridas de vermelho e azul, respectivamente. As alças estão em cinza e as voltas em verde. Imagem criada pelo software VMD, representação do tipo *cartoon* [38].

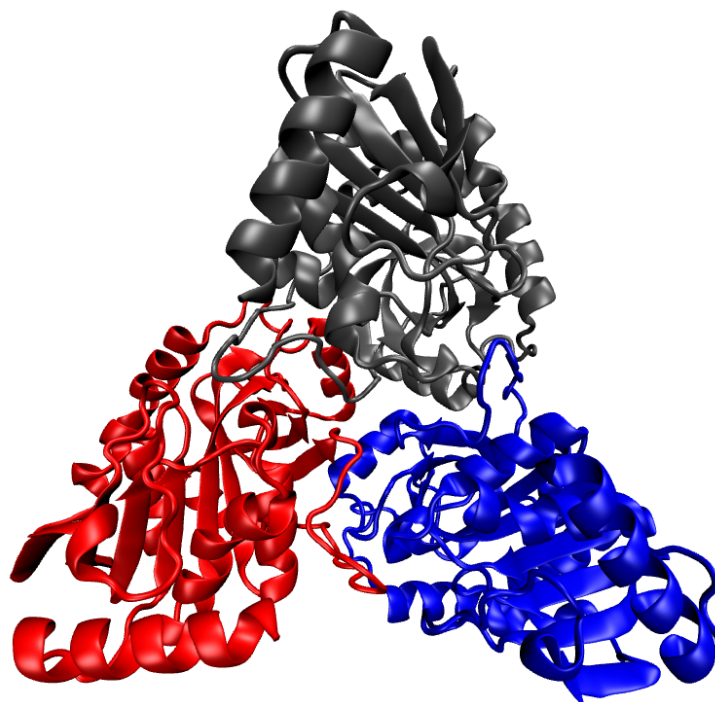


Figura 4 - Estrutura quaternária da proteína PNP de *Mycobacterium tuberculosis*, PDB ID: 1G2O. Formada pela interação de três subunidades diferentes, uma em azul, outra em cinza e outra em vermelho. Imagem criada utilizando o software VMD, representação do tipo *cartoon* [38].

## 2.2 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 a sua função, pois pode permitir a identificação de domínios conhecidos, como sítios catalíticos, sítios de modificação alostérica e outros [47]. A grande maioria dos fármacos atualmente no mercado atua interagindo com enzimas, logo o estudo da relação estrutura-função mostra-se vital para a criação de novas drogas e a bioinformática possui o importante papel de acelerar o processo de evolução deste conhecimento [81]. A abordagem aqui escolhida para a descoberta da estrutura 3D da proteína é pela da minimização de sua energia, uma abordagem *ab initio* baseada na hipótese termodinâmica de Anfinsen (1973) que diz que a conformação nativa adotada por uma proteína é aquela com a menor energia livre (Figura 5), o que representa o estado mais

estável. Entretanto, a predição dessa estrutura tridimensional é nada trivial e até mesmo abordagens simplificadas têm complexidade NP - Completa [22].

Com base no paradoxo de Levinthal [48], o número de conformações estruturais que uma proteína pode ter é enorme. Para uma cadeia com 100 aminoácidos, por exemplo, teremos ao menos  $2^{100}$  estados conformacionais, caracterizando um problema intratável [74]. O processo físico pelo qual um polipeptídeo se dobra em uma proteína funcional é uma questão antiga (revisado por Snow *et al.* [68]) 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. Enquanto o objetivo maior é o de prever a estrutura 3D a partir da estrutura primária, nossos atuais conhecimentos e poder computacional são simplesmente insuficientes para tratar um problema de complexidade tão alta [35].

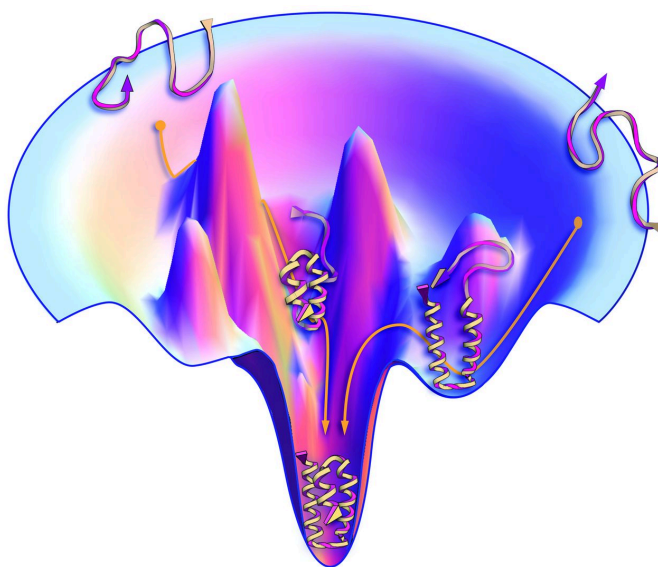


Figura 5 - Proteínas possuem um funil no que se trata da distribuição de energia, com vários picos e vales relacionados a estruturas não enoveladas e poucos vales com energia baixa e estruturas enoveladas. Figura obtida de [25].

## 2.3 Métodos Computacionais para Predição de Estruturas 3D de Proteínas

Os métodos computacionais para predição de estruturas de proteínas podem ser classificados em três grupos: (1) modelagem por homologia, (2) reconhecimento de padrões de enovelamento ou *fold recognition* e (3) métodos *ab initio* e *de novo*

### 2.3.1 Modelagem por Homologia

A modelagem comparativa se baseia no princípio de que se duas sequências de proteínas são relacionadas evolutivamente, elas possuem estruturas 3D similares [30]. 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 [54].

### 2.3.2 Reconhecimento de Padrões ou *Folding Recognition*

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. Nesse caso, as proteínas são ditas remotamente homólogas. O reconhecimento de padrões visa à identificação de estruturas remotamente homólogas por meio de uma coleção de enovelamentos candidatos [78]. 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 alinhamento [41]. Assim como na modelagem por homologia, nesse método só é possível prever estruturas que possuam sequências idênticas ou semelhantes armazenadas no PDB.

### 2.3.3 Predição *Ab initio* e *De novo*

As abordagens de predição *ab initio* e *de novo* são aquelas que não se baseiam em estruturas 3D e sim na termodinâmica estatística, mais especificamente na hipótese termodinâmica de Anfinsen [5], [66], [78]. 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

aos métodos citados anteriormente o fato de que, utilizando-a, é possível prever novas formas de enovelamento (se é que existem), devido ao fato de não ser baseado em proteínas com estruturas conhecidas [30]. Na verdade, em se tratando de métodos *ab initio* para a predição de estruturas de proteínas, podemos ainda fazer uma segunda classificação em: (i) os métodos que, para gerar suas soluções, utilizam além da sequência de aminoácidos, informações de bancos de dados como fragmentos de estruturas tridimensionais de proteínas (esses métodos são chamados métodos *de novo*) e (ii) aqueles que utilizam apenas as leis da física e a sequência primária da proteína (chamados métodos “verdadeiramente *ab initio*”).

## 2.4 Simulação Baseada em Sistemas Multi-agentes

### 2.4.1 Sistemas Multi-agentes

O estudo de Sistemas Multi-agentes (SMA) faz parte da área da Ciência da Computação intitulada Inteligência Artificial (IA) e se refere à modelagem de agentes autônomos que se relacionam em um universo em comum. Agentes são entidades computacionais que interagem com um ambiente e, basicamente, são guiados por objetivos, possuindo um corpo e uma localização no tempo e espaço. Um agente não existe sem um ambiente para atuar, e esse ambiente pode ser de diferentes tipos e complexidades. O modo como o agente reconhece o ambiente o qual habita depende fortemente das suas capacidades, por isso a classificação do ambiente deve ser feita do ponto de vista do agente. A complexidade do ambiente depende da complexidade do agente [21]. Normalmente, cada agente pode ser descrito por (i) um conjunto de capacidades comportamentais, as quais definem a sua competência, (ii) um conjunto de objetivos e (iii) a autonomia necessária para utilizar suas capacidades e alcançar seus objetivos. Um agente é uma entidade computacional autônoma, que decide suas próprias ações [2], [29], [76]. Um conjunto de agentes autônomos atuando em um ambiente caracteriza um SMA. A autonomia dos agentes significa que eles possuem uma existência própria, independente dos outros agentes, e seus próprios objetivos a alcançar [32].

### 2.4.2 Simulação Computacional

Uma simulação computacional é qualquer método criado por meio da computação para a exploração de propriedades de modelos matemáticos para os quais métodos analíticos são indisponíveis [39]. Seu uso como uma ferramenta auxiliar para tomadas de decisão é muito eficiente, pois torna possível o estudo de cenários das mais diferentes áreas. Especificamente,

no âmbito da biologia, a simulação é muito atrativa, pois permite que tarefas que demandariam elevado número de horas, sejam executadas em um tempo muito menor e pode prever as consequências de situações que envolvem altos riscos e custos. Normalmente, quando se faz referência à simulação computacional, se remete a fenômenos físicos como sendo fonte de exemplos e análises, entretanto, outras disciplinas podem ser levadas em consideração quando se tratando da definição de simulação computacional e de seu funcionamento. Deixando um pouco a física de lado e dando atenção à biologia, notamos que é difícil encontrarmos teorias compactas e elegantes no que se diz respeito a seus fenômenos e que, tipicamente, essas explicações são feitas por narrativas de linguagem natural e nem sempre são baseadas em paradigmas completos e bem fundamentados. A complexidade inerente a certos sistemas biológicos e a falta de teorias satisfatórias para explicá-los fazem a simulação em biologia alcançar propósitos maiores que a simulação em física, pois encontra a possibilidade de contribuir de forma central não somente na simulação do fenômeno em si, como também na construção do conhecimento teórico [3].

A Simulação Baseada em Sistemas Multi-agentes (MABS ou *Multi-Agent-Based Simulation*) é resultado da união de técnicas de Simulação Computacional e Sistemas Multi-agentes. Essa combinação gera uma nova área de pesquisa chamada Simulação Baseada em Sistemas Multi-agentes, a qual lida com problemas que envolvem múltiplos domínios [33]. Pesquisas anteriores destacam que abordagens baseadas em regras possibilitam a simulação e análise de classes de reações complexas que de outra forma seriam intratáveis [37]. A combinação de descrições de reações bioquímicas baseadas em regras juntamente com simulações computacionais baseadas em agentes nos leva a uma nova abordagem para explorar processos celulares complexos [7].

Em se tratando de SMAs aplicados no âmbito da predição de proteínas existem duas grandes vertentes: (i) a primeira vertente é abordada por Palopoli e Terracina [58] e se baseia na utilização de múltiplos preditores de proteínas atuando cada um como um agente que interage com os demais para alcançar a melhor predição possível, (ii) a segunda vertente, da qual este trabalho faz parte, visa uma modelagem de mais baixo nível, onde interações químicas entre agentes são simuladas.

#### 2.4.3 O Ambiente NetLogo

Para construirmos nossa simulação foi utilizado o ambiente NetLogo [72].

Sobre o NetLogo, segundo Lima *et al.* em [49]:

“É particularmente bem adaptado para modelar sistemas complexos que se desenvolvem ao longo do tempo. Os modeladores podem instruir centenas ou milhares de agentes, todos operando de forma independente. Isto torna possível explorar a conexão entre o comportamento no micronível de indivíduos e no macronível de padrões que emergem a partir da interação de muitos indivíduos. É simples o suficiente para permitir que estudantes possam facilmente executar ou até mesmo construir suas próprias simulações e avançado o suficiente para servir como uma poderosa ferramenta para pesquisadores de diversas áreas.”

Além disso, tendo em vista que o software foi desenvolvido em Java, torna-se possível a execução dos modelos nele criados em diversas plataformas operacionais. Basicamente, a estrutura do NetLogo é formada por três diferentes abas chamadas de “Interface”, “Info” e “Code”.

Na primeira aba está contida a interface do modelo de simulação, onde os valores dos parâmetros podem ser escolhidos e a simulação pode ser posta em execução. Podem também ser criados botões para fins de interatividade com o usuário. A aba contém também uma janela responsável pela visualização 3D do ambiente. Na segunda aba ficam contidas informações sobre o modelo, que respondem a perguntas como: “Para que serve?”, “Como funciona?”, além de dicas destinadas ao usuário. Na terceira aba está o código fonte. É onde os agentes são criados, seus atributos são estipulados e onde, no caso, toda abstração biologia-computação é feita. A programação em uma linguagem própria também intitulada NetLogo, que é uma variação da linguagem StarLogo [57], com a adição de novos recursos. A versão do software utilizada é a versão 5.0. Em comparação com as versões anteriores do NetLogo, nessa versão foi incorporada uma estrutura para simulação em três dimensões, até então não disponível.

## **2.5 Simulação Termodinâmica e Otimização Global**

Em nossa simulação, buscamos a conformação tridimensional que corresponde ao estado de menor energia livre, ou seja, o mínimo global de nossa função de energia. A função pode ser dividida em regiões e para cada região podemos ter um mínimo local diferente, na Figura 6 ilustramos uma hipotética função unidimensional, onde temos 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. Se começarmos do ponto  $P_1$ , por exemplo, chegaremos até A, enquanto começando de  $P_2$ , chegaremos 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 nos 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 [81]. Para tratar o problema de fugir de mínimos locais, utilizamos uma abordagem baseada no método de Monte Carlo e no método do arrefecimento simulado.

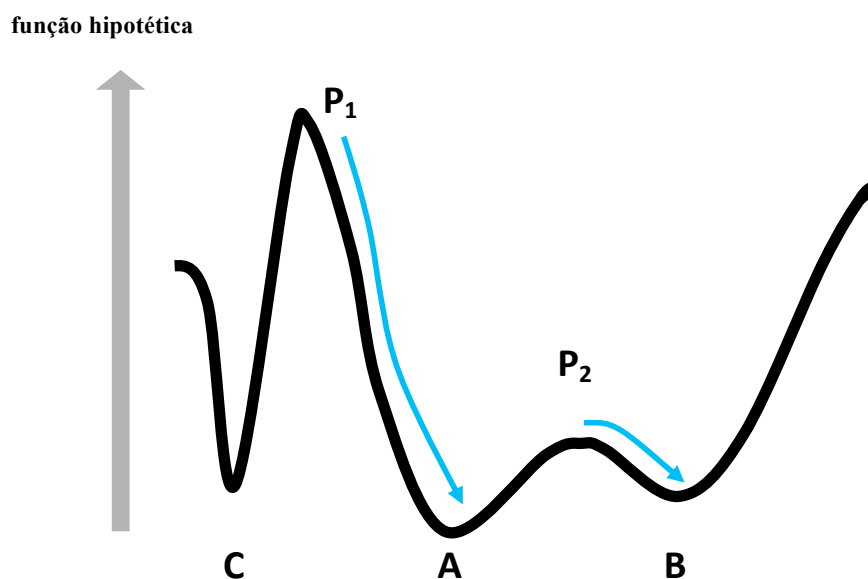


Figura 6 - Diagrama ilustrando o problema do mínimo global unidimensional, adaptado de [73]. A função mostrada contém três mínimos: A, B e C onde A é o mínimo global. O mínimo encontrado por uma otimização depende do ponto de início e da topologia da superfície. Se uma otimização é iniciada em  $P_1$ , chegará até A. Entretanto, se começar em  $P_2$ , logo a direita do ponto de máximo (barreira) existente entre A e B, o cálculo nos levará ao mínimo B.

### 2.5.1 Monte Carlo

O método de Monte Carlo permite que os movimentos sejam feitos em qualquer direção dentro de uma função, e especifica uma probabilidade para cada um desses movimentos. Por exemplo, se definirmos um estado 1 pela posição de todos os átomos do



sistema, teremos 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  $e^{-E_1/kT}$ , onde  $k$  é a constante de Boltzmann e  $T$  é a temperatura absoluta em Kelvins (K). A partir disso, se resolvermos comparar o estado 1 com um estado 2 considerando uma energia  $E_2$ , a relação de probabilidade seria dada pelo seguinte termo:

$$e^{-(E_2-E_1)/kT} = e^{-\Delta E_{21}/kT} \quad (1)$$

Partindo-se do estado 1, podemos facilmente determinar se o novo estado 2 é mais provável ou não de ocorrer em equilíbrio. Se  $\Delta E_{21}$  é negativo (estado 2 possui menor energia) o numerador terá um valor maior que um (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 uniforme no intervalo de  $[0,1]$  e, se esse número for menor que o número gerado pelo fator de Boltzmann (Equação 1) o movimento é aceito, caso contrário é 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 a frente), localizar a região do mínimo global energético, o qual seria o estado de melhor probabilidade. Vale frisar que, no método de MC a direção do movimento e a distância de deslocamento devem ser escolhidas aleatoriamente [81].

A amplitude máxima do movimento deve ser controlada e parametrizada de modo que aproximadamente 50% das tentativas de movimentação sejam aceitas, pelo menos no início da simulação. Se a amplitude de movimentação for muito pequena, então muitos movimentos serão aceitos, mas os estados serão demasiadamente similares e o espaço conformacional vai ser explorado lentamente. Valores muito altos para a amplitude de movimentação farão que muitas tentativas de movimentação sejam rejeitadas por levarem a conformações não favoráveis. A amplitude máxima pode ser ajustada automaticamente enquanto o programa executa para alcançar taxa de aceitação (*acceptance ratio*) desejada. Existe também a possibilidade de movimentar-se mais de um agente de uma só vez, com o intuito de percorrer o espaço conformacional mais eficientemente. Para isso, uma taxa de aceitação apropriada para esse tipo de movimento deve ser especificada [45].

### 2.5.2 Arrefecimento Simulado

O método de MC, em sua essência, não foi criado para otimização de funções e isso nos trás algumas desvantagens. No âmbito do problema PSP, por exemplo, a temperatura do sistema determina o tamanho dos mínimos locais que podem ser potencialmente transpostos pelo método. Se a temperatura é muito baixa, o método não conseguirá vaguar muito longe dos mínimos energéticos encontrados, o que é traduzido na obtenção de mínimos locais e não globais. O método de arrefecimento simulado (do inglês *Simulated Annealing*) é uma simples modificação que existe para o método de Monte Carlo que o transforma em um otimizador global. No começo da simulação o sistema está em a uma alta temperatura, o que permite ao sistema saltar sobre barreiras de energia razoavelmente altas. O sistema então começa a ter sua energia gradualmente resfriada, por fim causando que o sistema seja confinado a um único poço de energia. Em virtude da taxa gradativa (logarítmica) de resfriamento da temperatura, o sistema acaba passando mais tempo atuando sobre estados com energias mais baixas, há então uma boa chance de que o estado correspondente à energia mais baixa seja encontrado. Entretanto, não existe garantia de que isso ocorrerá [81]. Como vimos, a temperatura e a maneira como a mesma é diminuída é essencial para o desempenho do algoritmo de arrefecimento simulado. Tendo em vista que a convergência é garantida se a temperatura é reduzida logaritmicamente até zero, em nosso ambiente, a cada espaço de tempo a temperatura é diminuída gradativamente de acordo com a seguinte equação:

$$T_{K+1} = T_K * \alpha \quad \text{onde } \alpha = 0.98 \quad (2)$$

## 2.6 Critérios de Avaliação

### 2.6.1 RMSD

O desvio quadrático médio, do inglês *root-mean-square deviation* (RMSD) é a medida da distancia média entre os átomos de proteínas sobrepostas. É a medida mais comum no que se trata da comparação de estruturas de proteínas. A equação 10 mostra como o calculo de RMSD é feito.

$$RMSD = \sqrt{\frac{1}{N} \sum_{i=1}^N \delta_i^2} \quad (3)$$

Onde  $\delta$  é a distancia entre N pares de átomos equivalentes. Normalmente esses pares são formados por átomos referentes aos carbonos alfas (Cas) ou ao esqueleto da cadeia (C, N, O, C $\beta$ ). É comum também que, durante o calculo de RMSD, sejam efetuados 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 11 e o valor retornado é expresso em uma unidade de medida de distância, usualmente o Angström (Å), que equivale a  $10^{-10}$  m.

$$\begin{aligned} RMSD(v, w) &= \sqrt{\frac{1}{n} \sum_{i=1}^n \|v_{ix} - w_{ix}\|^2} \\ &= \sqrt{\frac{1}{n} \sum_{i=1}^n ((v_{ix} - w_{ix})^2 + (v_{iy} - w_{iy})^2 + (v_{iz} - w_{iz})^2)} \end{aligned} \quad (4)$$

### 2.6.2 TM-Score

Embora o cálculo de RMSD seja o mais conhecido método para comparação de estruturas de proteínas ele pode fornecer resultados não necessariamente condizentes com a qualidade da predição. Um exemplo é o caso onde estruturas alvo da comparação diferem na torção de um ângulo em um resíduo e isso torna o RMSD muito alto, quando na verdade um pequeno ajuste poderia levar a uma estrutura de RMSD muito baixo. Outro caso acontece quando uma estrutura adota o dobramento correto no centro da proteína ou *core* (onde as estruturas secundárias são mais estáveis) entretanto nas regiões N e C terminais a estrutura se distancia da estrutura nativa. Com isso em mente, os avaliadores presentes no encontro para experimentos de predição de estruturas de proteínas que acontece a cada dois anos, o CASP (Critical Assessment of Techniques for Protein Structure Prediction), passaram a utilizar métodos diferentes, o TM-Score ou *Template Modeling Score* é um deles. O TM-Score é mais sensível que o RMSD para avaliar a similaridade entre estruturas e isso pode ser notado na Figura 7. O TM-Score varia no intervalo (0,1] e, baseado em estatística, se um modelo possui um TM-Score por volta de 0,17 ou menor, isso significa que a predição possui similaridade randômica, entre 0,17 e 0,5 a predição possui certo grau de similaridade e um TM-Score de 0,5 a 1,0 significa que a estrutura possui o mesmo dobramento[77].

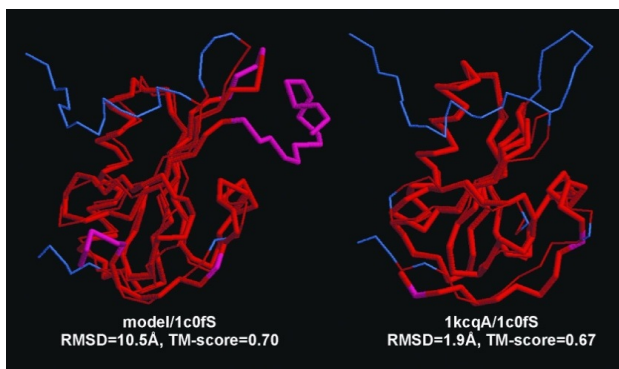


Figura 7 – Neste exemplo, duas estruturas possuem topologias similares nas regiões do núcleo da proteína, com TM-Score igual a 0,70 e 0,67 respectivamente. Entretanto, as variações nas regiões N e C terminais resultam em uma significativa diferença em termos de RMSD (de 1.9 Å para 10.5 Å). Esta figura foi obtida da Figura 5 de [79].

### 2.6.2 MaxSub

O MaxSub é um método que visa a identificação do maior subconjunto de átomos C alfa de um modelo que se sobrepõem “bem” sob uma estrutura experimental, e produz um valor ou *score* que representa a qualidade do modelo. O valor do MaxSub varia entre 0 e 1, onde 1 é um pareamento idêntico de estruturas [65].

### 2.6.3 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 *cutoff* [79]. O GDT é calculado por meio da Equação 12:

$$\text{GDT score} = (C1 + C2 + C3 + C4) / 4N \quad (5)$$

Onde C1 é o número de resíduos sobrepostos levando-se em consideração um raio de corte igual a  $(\text{threshold}/4)$ , C2 é o número de resíduos sobrepostos levando-se em consideração um raio de corte igual a  $(\text{threshold}/2)$ , C3 é o número de resíduos sobrepostos levando-se em consideração um raio de corte igual a  $(\text{threshold})$ , C4 é o número de resíduos sobrepostos levando-se em consideração um raio de corte igual a  $(\text{threshold} * 2)$  e N é o número total de resíduos. O valor resultante do cálculo de GDT possuirá valores variando de 0 até 1, onde valores de até 0,2 são tidos como a sobreposição aleatória de estruturas. No presente trabalho foi utilizado os softwares MaxCluster [65] e TM-Score [80] para o cálculo

do GDT, e foram utilizados dois limiares (*thresholds*), o primeiro de 4 Ångstroms (GDT\_TS ou GDT *Total Score*) e o segundo de 2 Ångstroms (GDT\_HA ou GDT *High Accuracy*).

#### 2.6.4 Diagrama de Ramachandran

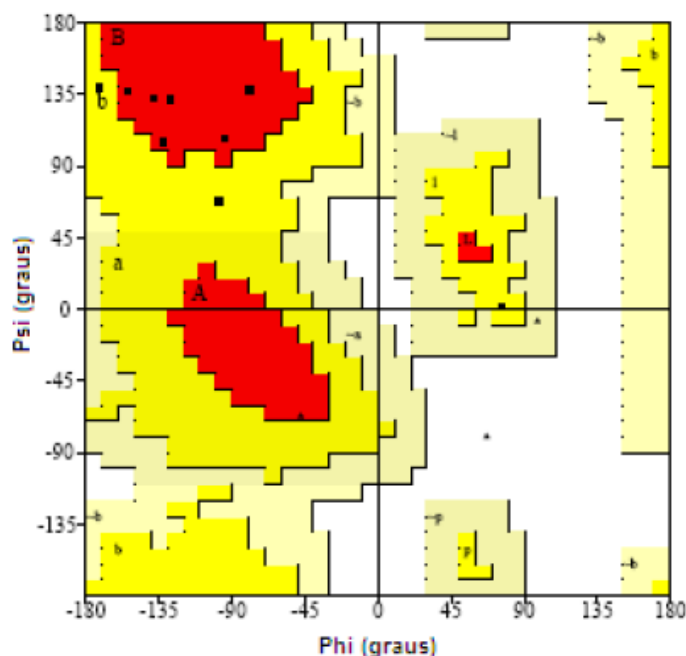


Figura 8 - Mapa de Ramachandran: região mais favorável em vermelho, região permitida em amarelo, região ainda aceitável em amarelo claro e região não permitida em branco. O canto superior em vermelho trata-se de região favorável para folhas  $\beta$  e no centro direito e esquerdo em vermelho para hélices  $\alpha$ , respectivamente. Modelo adotado por Thornton e colaboradores [44].

A conformação de uma proteína pode ser descrita, quantitativamente, em termos dos ângulos internos de rotação em torno das ligações entre os átomos da cadeia principal. As conformações proibidas estericamente são aquelas onde qualquer distância interatômica entre átomos não-ligados é menor que a distância de van der Waals correspondente. A Figura 8 demonstra que as regiões do mapa que identificam as conformações permitidas dependem do raio de van der Waals escolhido para calculá-las. Em termos de enovelamento, as regiões do mapa representam padrões de torção da cadeia polipeptídica para elementos da estrutura secundária como folhas  $\beta$  e hélices  $\alpha$  [75]. As regiões do mapa ou diagrama de Ramachandran também estão associadas a conformações de resíduos, isso é especificado pela nomenclatura de A. V. Efimov, disposto na Figura 9 [27]. Os diagramas de Ramachandran

contidos nesta dissertação foram obtidos utilizando-se o software PROCHECK, especificamente construído para destacar as regiões das proteínas que possuem geometria não usual e aquelas regiões que devem ser examinadas com mais cuidado [44]. Em virtude do modelo de abstração utilizado pelo PRO-SMART, foi necessário a utilização de dois programas auxiliares, um para completar a cadeia principal da proteína a partir dos carbonos alfas ([17]) e outro para completar as cadeias laterais dos aminoácidos[34].

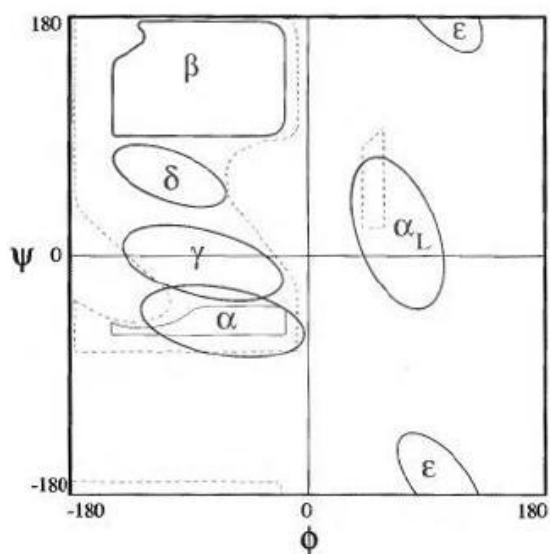


Figura 9 - Definições dos estados conformacionais no mapa de Ramachandran segundo A. V. Efimov [46].

### **3. TRABALHOS RELACIONADOS**

O capítulo a seguir será responsável por uma exposição de trabalhos encontrados na literatura os quais possuem relação com o foco da dissertação. O capítulo está dividido em duas seções. A primeira seção do capítulo expõe o protocolo de mapeamento sistemático executado visando, utilizando uma abordagem estruturada e suscetível à replicação, o alcance de uma maior gama de trabalhos relacionados. Na segunda seção do capítulo estão contidos os trabalhos encontrados, trazendo análises de vários artigos, com destaque específico para o artigo tido como “*seed*” para a dissertação.

#### **3.1 Mapeamento Sistemático: Protocolo**

Com intuito de alcançar melhores resultados no que se trata da descoberta do estado da arte na área específica acatada por esta dissertação, optou-se pela utilização de um protocolo estruturado para a execução da pesquisa 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 dissertação. O protocolo de mapeamento sistemático, criado com base em [60] está disposto no Apêndice A.

#### **3.2 Trabalhos Encontrados**

Dentre os trabalhos encontrados foi possível notar a grande diversidade de abordagens existente, entretanto, foi difícil encontrar abordagens que se enquadrassem no propósito desta dissertação (predição de estruturas 3D utilizando multi-agentes).

Jin e Kim [40] utilizaram uma abordagem envolvendo vários agentes simbolizando recursos disponíveis em várias plataformas diferentes e ataca o problema de coordenação desses recursos por meio de um mecanismo de planejamento automático aplicado junto a um sistema multi-agente.

Em [6] é descrita uma arquitetura genérica desenhada para dar suporte à implementação de aplicações visando o gerenciamento de informações de diferentes fontes. A arquitetura é baseada em: (i) Personalização, pois a informação é filtrada e organizada de acordo com os interesses do usuário; (ii) Adaptação, pois os perfis de usuário são melhorados e refinados ao longo do tempo por intermédio de técnicas sutis de adaptação e (iii) Cooperação, pois é baseado no gerenciamento de diferentes dados, provenientes de diferentes

fontes. As peculiaridades da arquitetura são destacadas em três diferentes casos de estudo onde um deles está relacionado à predição de estruturas secundárias de proteínas.

No trabalho de Armano *et al.* [6], por exemplo, há a descrição de um sistema multi-agente focado na predição de estruturas secundárias de proteínas. Em [1] há a apresentação de uma abordagem composta por um sistema multi-agente de aprendizado cooperativo, tendo como estudo de caso o problema da predição de estruturas secundárias de proteínas.

Em [24] nos é apresentada uma aplicação de métodos de inteligência de enxames na bioinformática, onde um enxame é tido como um grupo de agentes em cooperação, buscando atingir certo objetivo. Cada agente possui regras simples, e, das interações de todo o grupo, emerge uma inteligência coletiva. É importante notar que as características de inteligência de enxames citadas acima têm alta relação com o trabalho desenvolvido, pois, mesmo que enxames não sejam utilizados pela abordagem aqui descrita, tem-se como meta fazer que agentes, ao cooperarem (microscopicamente) gerem certas configurações grupais (macroscópicas). Embora seja sucinto, o trabalho ainda faz referências a aplicações no que trata da predição de estruturas secundárias e terciárias de proteína, sem demonstrações.

Cannata *et al.* [16] nos apresentam um *framework* conceitual para simulação de sistemas computacionais relacionados a comportamentos biológico modelado em termos de agentes e sociedades de agentes, já Ren *et al.*, em [62], apresentam uma plataforma de simulação de redes biológicas para o estudo de sistemas baseados em biologia e implementado utilizando agentes. Propõe a simulação de comportamentos de sistemas biológicos e sua modelagem em termos de entidades biológicas e de sociedade.

Em [3], Amigoni e Schiaffonati nos proveem de uma análise crítica bastante interessante em relação ao uso de sistemas multi-agentes para a execução de simulações de processos biológicos. Seu artigo destaca pontos positivos do uso de MAS no âmbito biológico, como a possibilidade de realização de experimentos dispostos de grande flexibilidade e a provocação de distúrbios no modelo a ser simulado. Entretanto, ressalta também pontos negativos como a lacuna que existe, dependendo do contexto biológico onde o sistema está inserido, no que se trata da validação dos resultados produzidos pela simulação. A análise aborda ainda um caminho para solucionar, em alguns casos, o problema de validação descrito acima fazendo uso de comparações dos dados gerados pela simulação com dados reais, abordagem factível em contextos como o da predição estrutural de proteínas, por exemplo.

Palopoli e Terracina [58] abordam o tema da predição de estruturas de proteínas utilizando multi-agentes de forma diferente, dando importância à grande quantidade de



técnicas de predição que vem surgindo nos últimos anos, fazendo que cada uma dessas técnicas seja representada por um agente diferente. A ideia base do trabalho parte da união de diferentes técnicas de predição e da combinação de resultados tendo como meta a melhoria de qualidade das predições.

Também foram analisadas revisões [52], [56] e [78]. Enquanto as duas primeiras citam, em se tratando de agentes e o problema PSP, apenas um artigo (“Protein Secondary Structure Prediction through a Cooperative MultiAgent Learning Approach.” de Addis e colaboradores [1]), a revisão de Jiang Ye ([78]) cita apenas o campo de pesquisa, caracterizando-o como uma técnica de programação evolutiva que pode ser utilizada para predição, controle e classificação. Os poucos artigos referenciados quando se trata do problema específico desta pesquisa evidenciam que, realmente, até hoje, um número reduzido de trabalhos abordaram o problema de tal forma.

Dentre todos os artigos encontrados, o que mais se destacou foram os artigos intitulados “Multi-Agent Simulation of Protein Folding” [12] e “Multiagent-Based Simulation” [13], de autoria de Luca Bortolussi, Agostino Dovier e Federico Fogolari, da Universidade de Udine, Itália. Bortolussi e colaboradores apresentam um arcabouço para a simulação *ab initio* da predição de estruturas de proteínas onde cada aminoácido é simbolizado por um agente, conceito que se aproxima da proposta desta dissertação e, por esse motivo, decidiu-se que o construto proposto por Bortolussi *et al.* serviria como artigo base e teria características incorporadas ao PRO-SMART como: (i) sua fundamentação baseada em Monte Carlo junto a arrefecimento simulado e (ii) uma arquitetura de agentes criada baseada na proposta de Roli e Milano [63], a qual se caracteriza por diferentes níveis de agentes configurados de maneira hierárquica. A descrição do trabalho preocupa-se ainda em manter a função de energia a ser utilizada em uma relação de ortogonalidade com as demais características da simulação, o que nos possibilita a escolha de qualquer função de energia para reger a simulação, o que também acontece no PRO-SMART. No artigo de Bortolussi *et al.* há ainda a demonstração de uma aplicação utilizando-se espaço de tuplas e a linguagem Prolog, o que destaca o potencial e a viabilidade de aplicação do construto.



## 4. PRO-SMART: ASPECTOS CONCEITUAIS

No capítulo a seguir serão apresentados aspectos conceituais específicos do PRO-SMART. Primeiramente será abordado o nível de abstração utilizado pela ferramenta nesta versão e em seguida a função de energia utilizada. Por fim será apresentado o esquema de cooperação implementado na ferramenta.

### 4.1 Nível de Abstração

Uma função de energia bem definida deve considerar todas as possíveis interações entre todos os átomos de cada aminoácido que compõe a proteína, entretanto, simulações que modelam todos os átomos presentes em cada aminoácido (*All-atom Simulations*) são tipicamente impraticáveis por serem demasiadamente dispendiosas em termos computacionais.

Representações de modelos reduzidos vem sendo objeto de interesse de pesquisadores no estudo teórico de simulações da estrutura e da dinâmica de proteínas [18], [20], [43], [73]. A primeira razão para tal é a de envolver esforços computacionais muito menores se comparado a 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 [51].

Com o objetivo de tornar a dinâmica mais fácil e rápida em nosso sistema foi utilizada uma abordagem simplificada, representando cada um por meio de apenas duas esferas, uma representando seu carbono alfa e outra representando o centroide de sua cadeia lateral. A distância entre dois átomos consecutivos foi definida como 3.8 Å. A Figura 10 exemplifica o modelo de representação abstrata da cadeia polipeptídica utilizado. Nesse nível de abstração, a geometria local é descrita por:

- *bend angles*: os ângulos formados entre duas ligações entre três átomos de carbonos consecutivos (Figura 11).
- *torsion angles*: os ângulos entre os planos formados pelos três primeiros e três últimos carbonos alfa em uma sequência de quatro consecutivos (Figura 12). Mais detalhadamente, o ângulo é obtido por duas normais, uma formada pelo plano criado por  $C_{ai}$ ,  $C_{ai+1}$  e  $C_{ai+2}$  e a normal formada por  $C_{ai+1}$ ,  $C_{ai+2}$  e  $C_{ai+3}$ .

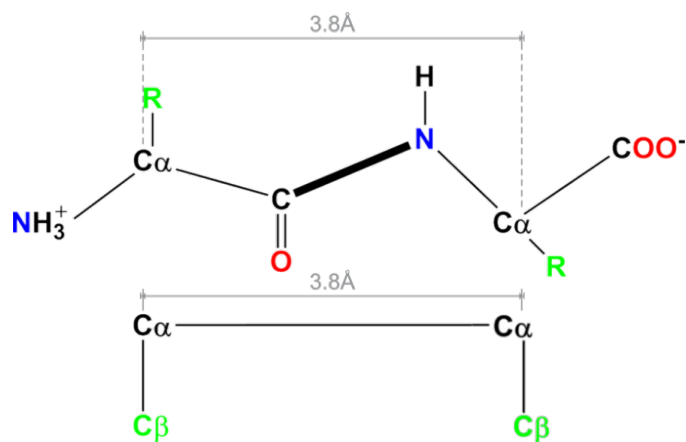


Figura 10 - Exemplo de modelo com representação reduzida: No modelo proposto por Berrera *et. al* em [10] cada aminoácido é composto por apenas duas esferas, uma representa seu carbono alfa e outra representa o centroide de sua cadeia lateral. A distância entre dois átomos de carbono alfa consecutivos foi definida como 3.8 Å. Imagem criada pelo software ChemDraw [36].

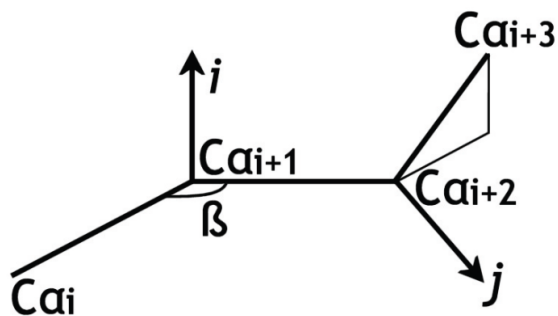


Figura 11 - *Bend angle* – Figura adaptada de [14].

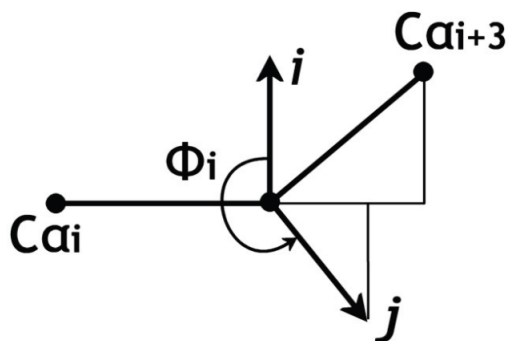


Figura 12 - *Torsion angle* – Figura adaptada de [14].

## 4.2 Função de Energia

Embora o PRO-SMART possua como característica o desprendimento em relação à função de energia escolhida, a mesma depende diretamente do nível de abstração escolhido para representar a proteína. Como nossa abstração leva em conta apenas carbonos alfa e o centroide da cadeia lateral (ou carbono beta, no caso de glicinas), nosso escopo de possíveis funções de energia que podem nortear a simulação é restrito. Respeitando essa premissa e visando uma melhor avaliação dos resultados encontrados nesta primeira versão do PRO-SMART, optou-se por utilizar uma função já testada pelo arcabouço de Bortolussi *et al.* [12], [13]. Criada por Berrera *et al.* em [10] e estendida por Bortolussi *et al.* em [14], é uma função empírica, onde as energias são calculadas por meio de informações estatísticas provenientes de um conjunto de estruturas de proteínas selecionadas de uma base de dados específica. Em termos teóricos, tal embasamento estatístico tem como objetivo refletir a predisposição de cada tipo de aminoácido de interagir com cada outro aminoácido (inclusive com aminoácidos de mesmo tipo). A função de energia é adimensional (não possui unidade de medida) e é formada por quatro termos: *bond distance* ( $E_b$ ), *bend angle* ( $E_a$ ), *torsion angle* ( $E_t$ ) e *contact interaction* ( $E_c$ ). Considerando  $x$  a disposição espacial do agente C-Alfa na cadeia proteica, a energia pode ser expressa pela equação (3):

$$E(x) = E(x)_b + E(x)_a + E(x)_t + E(x)_c \quad (6)$$

Em se tratando do termo *bond distance* ou distância de ligação, para cada par de aminoácidos consecutivos  $x_i, x_{i+1}$ , se tem um termo quadrático da forma da equação 4:

$$E(x)_b = \sum_{1 \leq i \leq n-1} (r(x_i, x_{i+1}) - r_0)^2 \quad (7)$$

Onde  $i$  representa o átomo alvo,  $n$  é o tamanho da cadeia,  $r(x_i, x_{i+1})$  representa a distância entre os  $C_\alpha$  dos dois aminoácidos e  $r_0$  é a distância típica de 3.8 Å entre dois  $C_\alpha$ . O termo tenta manter a distância entre dois agentes C-Alfa constante. A energia referente ao ângulo de deformação ou *bend angle* é associada ao ângulo de deformação formado por três  $C_\alpha$  consecutivos. É um potencial estatístico e, assim como o termo *torsional angle* ou ângulo de torção, tenta induzir o sistema a boas conformações geométricas locais favorecendo, por exemplo, a formação de estruturas secundárias.

A distribuição é bastante constante para todas as proteínas presentes no PDB e independente dos tipos de aminoácidos envolvidos. O perfil, ou modo como a distribuição ocorre (ver [31]) pode ser aproximado por uma combinação de duas Gaussianas, uma em

torno de 120 graus e outra mais precisaa, a cerca de 90 graus. A energia é obtida aplicando-se o oposto do logaritmo à função de distribuição:

$$E(x)_a = \sum_{i=1}^{n-2} -\log \left( \alpha_1 e^{-\left(\frac{\beta_i - \beta_1}{\sigma_1}\right)^2} + \alpha_2 e^{-\left(\frac{\beta_i - \beta_2}{\sigma_2}\right)^2} \right) \quad (8)$$

Onde os parâmetros  $\alpha_1$ ,  $\alpha_2$  estão relacionados às posições dos átomos,  $\beta_1$ ,  $\beta_2$  se referem aos *bend angles* e  $\sigma_1$  e  $\sigma_2$  aos perfis das Gaussianas.

O termo referente ao ângulo de torção é modelado utilizando informações de comportamentos de torção extraídos do PDB. O ângulo é influenciado tanto pelo tipo de aminoácidos envolvidos quanto por suas posições no espaço virtual. Segundo [14], a informação disponível no PDB não permite que se crie um bom perfil de distribuição para cada combinação de cadeias (consequência do pequeno número de proteínas não homólogas, tendo como parâmetro o corte ou *cutoff* de 25%). Como consequência, para tratar a energia potencial relacionada aos ângulos de torção, houve uma divisão dos aminoácidos que compartilham o mesmo comportamento em quatro classes, e então foram calculados perfis de distribuição para cada seqüência de classes consecutivas. O perfil resultante é aproximado pela soma de duas Gaussianas e a função possui a seguinte descrição:

$$E(x)_t = \sum_{i=1}^{n-3} -\log \left( \alpha_1 e^{\left(\frac{\phi_i - \phi_1}{\sigma_1 + \sigma_0}\right)^2} + \alpha_2 e^{\left(\frac{\phi_i - \phi_2}{\sigma_2 + \sigma_0}\right)^2} \right) \quad (9)$$

Onde os parâmetros  $\alpha_1$ ,  $\alpha_2$ ,  $\sigma_1$ ,  $\sigma_2$ ,  $\phi_1$ ,  $\phi_2$  dependem das classes dos quatro resíduos, e  $\sigma_0$  é utilizado para adaptar a distribuição a uma função de energia efetiva. Por fim, o termo  $E_c$  captura interações de longo alcance utilizando uma matriz estatística de contato desenvolvida por Berrera *et al.* [10]. Para cada par de aminoácidos não consecutivos:

$$E(x)_c = \sum_{i=1}^{n-3} \sum_{j=i+3}^n \left[ \begin{aligned} & |Pot(x_i, x_j)| \left( \frac{r_0(x_i, x_j)}{r(x_i, x_j)} \right)^{12} \\ & + Pot(x_i, x_j) \left( \frac{r_0(x_i, x_j)}{r(x_i, x_j)} \right)^6 \end{aligned} \right] \quad (10)$$

Onde  $r(x_i, x_j)$  é a distância entre  $C_\alpha$  de  $i$  e  $j$ , já  $r_0(x_i, x_j)$  é um parâmetro descrevendo o impedimento estérico (relativo à organização espacial dos átomos) entre um par de resíduos não consecutivos.

Um problema crucial quando lidando com funções de energia é estipular corretamente os parâmetros envolvidos, pois eles afetam dramaticamente o panorama energético, o que influencia diretamente o comportamento e resultados da simulação. Com isso em mente Bortolussi, Palú, Dovie e Fogolari, em [14], realizaram uma otimização utilizando como modelo 700 proteínas provenientes do PDB com menos de 25% de homologia. Foi calculada a soma dos quadrados das diferenças entre as energias computadas pelas conformações nativas e pela energia resultante da aplicação da função aqui descrita (aproximada proporcionalmente ao número de aminoácidos). Então a função foi minimizada usando-se uma abordagem utilizando-se conceitos de *Simulated Annealing* ou Arrefecimento Simulado (método o qual não foi descrito em detalhes pelos autores) chegando assim a parâmetros tidos como otimizados, capazes de gerar razoáveis valores de energia quando aplicando a função sobre a estrutura nativa de proteínas. Neste trabalho foram utilizados os mesmos valores para os parâmetros da função de energia utilizados por Bortolussi *et al.*, tornando possível a comparação dos métodos.

Um aspecto muito importante referente à função de energia refere-se à sua implementação dentro do sistema multi-agente: É no momento de calcular a energia que o sistema mais se utiliza da comunicação entre agentes, seja somente com aqueles que estão próximos o suficiente (cálculo da energia local realizado por agentes do tipo C-Alfa) ou levando em consideração todos os agentes presentes no ambiente (cálculo da energia global realizado pelo agente Diretor).

### 4.3 Cooperação

Outra característica do PRO-SMART é a possibilidade de se adicionar informações sobre a estrutura secundária da proteína alvo, a fim de favorecer tais conformações. A cooperação via estruturas secundárias, no PRO-SMART, acontece via penalizações em termos de energia para aquelas conformações que se distanciam em termos de RMSD (*Root Mean Square Deviation*, medida relacionada à distância média entre átomos de proteínas sobrepostas, é a medida mais comum quando tratamos da similaridade entre estruturas tridimensionais de proteínas) em relação às estruturas moldes da estrutura secundária que deveriam acatar. Aqui vale a ressalva de uma importante diferenciação em termos de abordagem utilizada quando em relação ao construto de Bortolussi *et al.* No construto de Bortolussi *et al.*, a cooperação via estruturas secundárias é feita utilizando-se de informações

provenientes do próprio arquivo PDB da proteína que se estava modelando, sendo assim, os moldes a serem utilizados no momento de calcular RMSD e penalizar ou não as estruturas que estão se formando eram exatamente as conformações experimentais (no artigo os autores expressam a vontade de buscar alternativas e enfatizam que a abordagem utilizada por eles é apenas uma abordagem inicial para vias de testes). Embora fosse cômodo, como nossa abordagem é puramente *ab initio* e se utilizar de informações sobre a exata conformação experimental das proteínas não é factível e, para tornar possível este aspecto dentro do PRO-SMART serão utilizados preditores disponíveis na internet o que, em teoria, deve diminuir o desempenho da cooperação, porém fará que o PRO-SMART possa utilizar informações sobre estruturas secundárias a partir de qualquer sequência de aminoácidos, esteja essa sequência no PDB ou não. Dada uma proteína, por exemplo, a entrada 2LR5 do PDB, representada pela seguinte cadeia:

GFGCPFNENECHAHCLSIGRKFGFCAGPLRATCTCGKQ

O usuário deverá informar ao programa uma sequência com informações sobre a estrutura secundária dessa proteína. Utilizando-se do método Porter, chegamos à seguinte sequência:

GFGCPFNENECHAHCLSIGRKFGFCAGPLRATCTCGKQ  
 CCCCCCCHHHHHHHHHHCCCCCECCCCCCCCCECCCC

Onde “C” significa *covil* ou volta, “H” significa *hélio* ou hélice e “E” significa *Extended Strand* ou folha  $\beta$ . Sobre qual método utilizar, na dissertação de mestrado intitulada “Um estudo sobre a predição da estrutura 3D aproximada de proteínas utilizando o método CReF com refinamento”, Dall’Agno e Norberto de Souza executaram uma avaliação dos métodos de predição de estruturas secundárias e, nos testes realizados, o método Porter [61] apresentou a melhor acurácia para a maioria das proteínas, exceto para proteínas com menos de 20 aminoácidos [23]. Quando tratamos de proteínas com menos de 20 aminoácidos, o método com melhor acurácia para predizer estruturas secundárias (em comparação resíduo a resíduo com a estrutura determinada experimentalmente) foi o SAM-T08 [42]. O usuário deve escolher qual método utilizar para, a partir da sequência da proteína, predizer a sua estrutura secundária. Entretanto, no PRO-SMART, este processo é automatizado. Um valor peso deve ser estipulado pelo usuário para definir a importância da cooperação via estruturas secundárias dentro da simulação (o quanto de fato alterará a energia). Assim sendo, quando a cooperação é ativada, os cálculos de energia terão um termo a mais (Equações 8 e 9):



$$Energia_{Cooperação} = Peso_{Cooperação} * RMSD_{Estrutura\ Secundária} \quad (11)$$

$$Energia = Energia + Energia_{Cooperação} \quad (12)$$

Fora escolhidos dois pesos diferentes para a cooperação via estruturas secundárias: (i) peso = 1 e (ii) peso = Energia / 10. Teste preliminares mostraram que a cooperação deve ser ativada somente depois que a simulação já transcorreu por certo tempo ao invés de ser ativada desde o início da simulação. A razão para isso é a que caso a cooperação seja ativa muito cedo o sistema ficará “preso” as conformações de estruturas secundárias e assim deixará de alcançar valores energéticos satisfatórios. Optou-se então pela ativação da cooperação via estruturas secundárias somente a partir do *tick* ou passo de temperatura 100.

#### 4.4 Clusterização de Estruturas

Analisando-se os resultados de simulações iniciais do PRO-SMART percebeu-se que a última estrutura que a simulação adota não é, na grande maioria das vezes, aquela com menor RMSD. Passou-se então a buscar abordagens alternativas para a determinação da estrutura tida melhor representante da simulação e então se decidiu que seria utilizado um esquema de clusterização de estruturas. O intuito da utilização de clusterização é encontrar os nichos mais acessados, os quais tem maior tendência a representarem uma estrutura nativa de uma proteína, entretanto é necessário enfatizar que, no PRO-SMART, a clusterização se utiliza apenas das estruturas alcançadas a partir do *tick* ou passo de tempo 100, o que nos dá a certeza de que a estrutura resposta terá como característica uma baixa energia. O algoritmo de clusterização utilizado é o do vizinho mais próximo ou *Nearest Neighbour* e o critério de avaliação é o GDT (ver seção 2.6.3 GDT) com limiar de 2 Å . Foi utilizado o software MaxCluster [65].

O passo a passo a seguir tem a intenção de deixar claro o protocolo utilizado para a escolha do modelo ou estrutura representante após uma simulação ter sido executada:

Passo 1 – Abrir o arquivo “CLUSTER-MENORES-ENERGIAS.txt” contido na pasta da simulação e

Passo 2 – Buscar a resposta do algoritmo de *Nearest Neighbour clustering*, conforme exemplifica a Figura 13.

A estrutura resposta será aquela que representar o centro do maior cluster, no caso a estrutura referente ao *tick* 170. Caso existam dois ou mais clusters do mesmo tamanho e estes sejam os maiores cluster, o usuário deverá escolher a estrutura que possuir a menor energia.

```
Processed 8380 of 8385 GDIs
6 INFO: CPU time = 1.33 seconds
7 INFO: =====
8 INFO: Nearest Neighbour clustering
9 INFO: =====
10 INFO: 1 : 71 37 0.113 /home/tpaes/netlogo-5.0/outputs/1PG1-Num10-170.pdb
11 INFO: 2 : 86 20 0.278 /home/tpaes/netlogo-5.0/outputs/1PG1-Num10-185.pdb
12 INFO: 3 : 116 12 0.347 /home/tpaes/netlogo-5.0/outputs/1PG1-Num10-215.pdb
13 INFO: 4 : 29 9 0.000 /home/tpaes/netlogo-5.0/outputs/1PG1-Num10-128.pdb
14 INFO: 5 : 48 8 0.347 /home/tpaes/netlogo-5.0/outputs/1PG1-Num10-147.pdb
15 INFO: 6 : 107 6 0.000 /home/tpaes/netlogo-5.0/outputs/1PG1-Num10-206.pdb
16 INFO: =====
```

Figura 13 - Exemplo do arquivo de clusterização gerado. O usuário deve escolher a estrutura correspondente ao centro do maior cluster encontrado.

## 5. PRO-SMART: IMPLEMENTAÇÃO

Este capítulo apresenta detalhes sobre a implementação do ambiente multi-agente PRO-SMART feito na linguagem e software NetLogo v5.0. Primeiramente é apresentado o esquema geral do sistema e em seguida são apresentados os tipos de agentes e suas características, os requisitos do sistema, a interface e, por fim, informações sobre a execução de simulações e coleta de resultados.

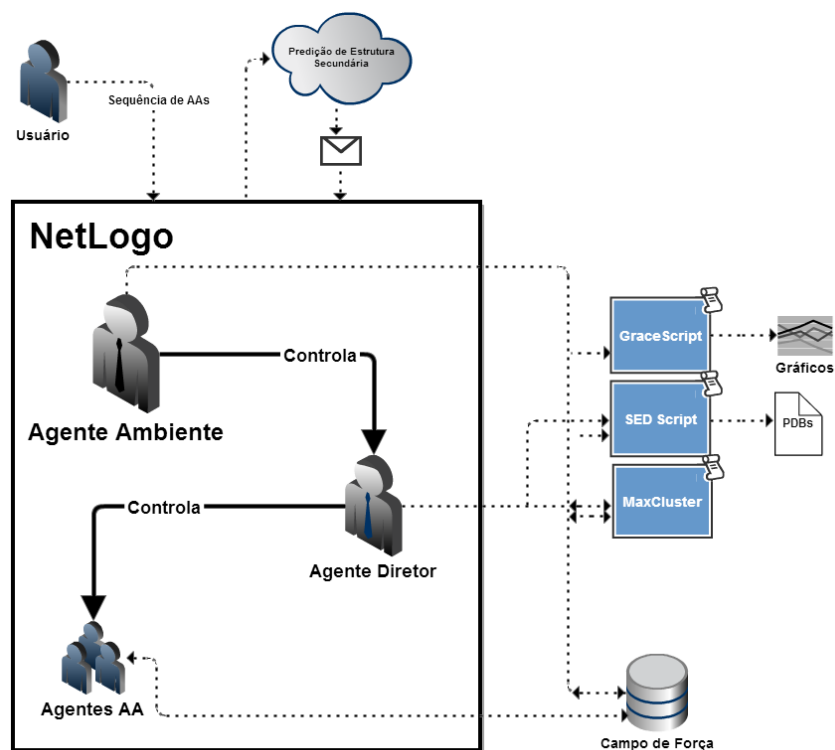


Figura 14 - Esquema geral do PRO-SMART. Enquanto dentro do NetLogo os agentes cooperam para alcançar a melhor conformação, scripts e banco de dados, de fora do NetLogo, contribuem no processamento de informações e cálculos

### 5.1 Esquema Geral

A Figura 14 mostra o esquema geral de funcionamento do PRO-SMART. A única entrada que o programa necessita é a sequência de aminoácidos da proteína e, caso o usuário opte por utilizar-se da cooperação por estruturas secundárias, uma sequência simbolizando-a. A Figura 14 mostra também os scripts existentes junto ao PRO-SMART, os quais são abstraídos do usuário. Existe um script em Python para a automatização do processo da

predição de estruturas secundárias, um script em SED (*Stream Editor*) para a manipulação de arquivos PDBs e um script na linguagem do Sistema Operacional (chamado *GraceScript*) que auxilia na criação de gráficos. O nome *GraceScript* foi escolhido pelo fato dos gráficos gerados serem no formato XMGRACE.

## 5.2 Agentes

Existem vários tipos de agentes e há uma divisão por tipos baseada em hierarquia, aplicando o que foi proposto por Roli e Milano em [63], onde os agentes se dividem em níveis diferentes e os agentes de nível mais elevado têm a função de coordenar as ações dos agentes de nível mais baixo. Nossa simulação conta com quatro tipos de agentes: C-Alfa e C-Beta, os quais podemos chamar de agentes Aminoácidos, o agente Diretor e o agente Ambiente. A Figura 15 nos trás o esquema geral dos agentes e suas hierarquias.

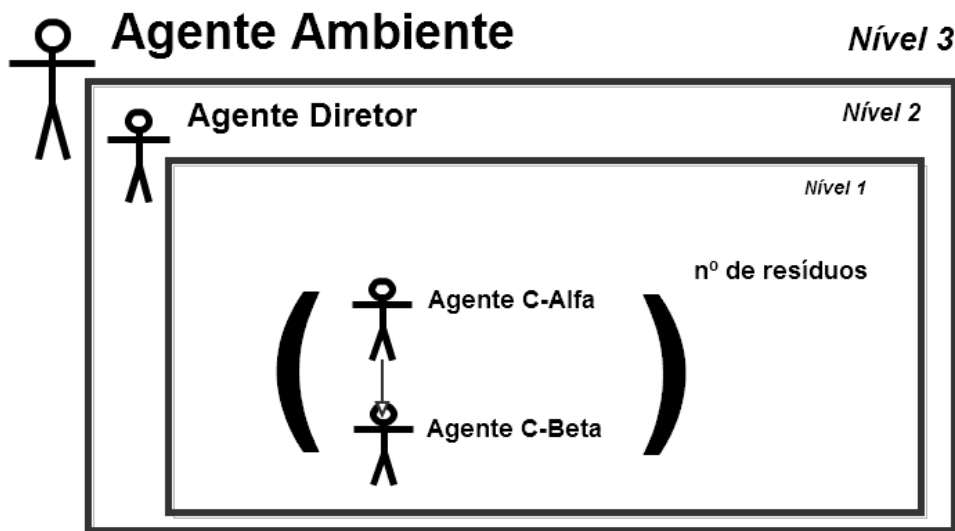


Figura 15 - Hierarquia dos agentes. Agentes de maior nível controlam os de menor. O número de agentes C-Alfa e C-Beta depende do tamanho da cadeia proteica.

### 5.2.1 Agentes tipo C-Alfa – Nível 1

*Definição:* Tem como missão a exploração do espaço conformacional. Para cada aminoácido presente na proteína, está associado um agente C-Alfa. A posição dos agentes C-Alfas (e também C-Betas) é expressa por coordenadas cartesianas, implicando nos movimentos de cada agente serem locais, ou seja, não afetam a posição dos outros agentes.

Sobre os atributos/atributos que definem o agente, alguns estão relacionados ao controle da simulação e outros estipulados de acordo com o padrão para átomos do PDB [9]. São eles:

- *who*: número do átomo no arquivo;
- *color*: cor do átomo, modifica-se de acordo com o aminoácido: cinza para não polares, amarelo para polares sem carga, vermelho para polares ácidos e azul para polares básicos;
- *heading*, *pitch* e *roll*: ângulos relacionados com a direção para qual o agente aponta, sua inclinação cima-baixo e sua torção no espaço 3D;
- *xcor*: posição do átomo em relação ao eixo x;
- *ycor*: posição do átomo em relação ao eixo y;
- *zcor*: posição do átomo em relação ao eixo z;
- *shape*: relacionado com a forma do agente, no caso sempre será “circle”;
- *label*: é o nome do átomo, de acordo com os padrões PDB;
- *size*: o tamanho do átomo em relação ao universo 3D em que se situa;
- *residue-name*: contém o nome do resíduo ao qual o átomo pertence, dentre os 20 tipos de aminoácidos;
- *chain*: a que cadeia pertence (inicialmente definido o valor “A” como default, até que se tenham heurísticas para reger a diferenciação entre cadeias);
- *residue-nr*: relacionado ao número do resíduo em dada sequência de aminoácidos;
- *occupancy*: parâmetro utilizado pelo PDB para a ocupância do átomo, aqui tendo o valor padrão de 1.0;
- *temp*: parâmetro referente a temperatura, tem o valor default de 0;
- *atom-type*: referente ao tipo de aminoácido que o agente representa, também utilizado pelo padrão PDB;
- *atom-parameter*: referente ao tipo de átomo que o agente representa, também utilizado pelo padrão PDB;
- *bond1*, *bond2*, *bond3* e *bond4*: guardam informações sobre as ligações feitas pelo átomo de carbono. O *bond3* guarda sempre o *who* do agente C-Beta em que o C-Alfa está ligado;
- *forbidden*: está relacionado ao tipo de proibição para que o agente possua em relação à movimentação quando o mesmo faz parte de uma estrutura secundária. É preciso

escolher os tipos de proibição a serem levados em conta por cada tipo de movimento (*pivô e crankshaft*);

- *ss*: Aqui estarão guardadas informações sobre que tipo de estrutura secundária o agente tem tendência a formar;
- *amino\_moves\_attempted*: contabiliza as tentativas de movimentação efetuadas pelo agente;
- *energia*: agentes C-Alfa possuem um valor contendo sua energia, porém essa energia é uma energia local e, para calculá-la, os agentes utilizam-se de troca de informações (tipo, distancia, posição) com os agentes que estão a sua volta (agentes C-Alfa e C-Beta) e
- *bond-type*: representação das ligações entre os átomos e, neste primeiro modelo de simulação, essas ligações só tem um parâmetro chamado bond-type e ele pode ser estipulado como duplo ou normal, porém o tipo de ligação não é levado em conta pela função de energia testada.

*Movimentação*: A exploração espacial é feita permitindo a interação entre agentes e a troca de informações. Foram implementadas duas formas para movimentação e serão descritas a seguir porém, antes de descrevê-las, vale destacar que a distância relativa nos eixos x, y e z entre agentes C-Alfas e C-Betas permanece a mesma seja qual for a estratégia escolhida:

Cubo: A estratégia de Cubo é baseada no trabalho de Bortolussi *et al.* onde, para se movimentar o agente escolhe com igual probabilidade um ponto em um cubo centrado em sua posição atual. O comprimento do lado do cubo foi definido experimentalmente em 1.0 Å podendo ser modificado a critério do usuário. Como em nossa simulação todo o espaço de busca é praticamente acessível, a estratégia de movimento é muito simples e garante acatar a ergodicidade<sup>1</sup> e irreducibilidade requerida por métodos de Monte Carlo (requerem uma exploração imparcial sobre o espaço de busca).

---

<sup>1</sup> Em termodinâmica estatística, a hipótese da ergodicidade estabelece que, sobre um período prolongado de tempo, o tempo de permanência em uma dada região do espaço de fase de microestados com a mesma energia é proporcional ao volume da região, ou seja, todos os microestados acessíveis são igualmente prováveis ao longo de um período de tempo prolongado [11]

Distâncias Fixas: Dois aminoácidos consecutivos tendem a permanecer a uma distância de 3.8 Å. Com isso em mente, Bortolussi, Palú, Dovie e Fogolari, em seu trabalho envolvendo Programação Concorrente por Restrições ou *Concurrent Constraint Programming* [14] criou uma estratégia que serviu como base para a implementação de uma estratégia alternativa para movimentar nossos agentes: Movimentando os agentes cegamente no espaço virtual leva muito seguidamente a grandes aumentos na energia do sistema. Nessa estratégia, não implantamos uma restrição rígida em torno dos 3.8 Å, mas sim uma restrição leve, uma forma mais inteligente para o agente se movimentar.

- Se o agente (C-Alfa) a ser movimentado se encontra no meio da cadeia, ou seja, não representa o primeiro ou último aminoácido, ele verifica as distância entre ele e os agentes C-Alfa posterior e anterior, escolhe um ângulo e se movimenta de modo a manter essas distâncias fixas. Em outras palavras, escolhe um novo ponto em uma circunferência (Figura 16). Possui apenas um grau de liberdade.
- Quando o agente representa o primeiro ou último aminoácido da cadeia, ele possui apenas um agente adjacente. Ele verifica a distância que está do agente adjacente e pode se movimentar para qualquer outro ponto que esteja equidistante do mesmo. Pode escolher um ponto na superfície de uma esfera. (Figura 17).

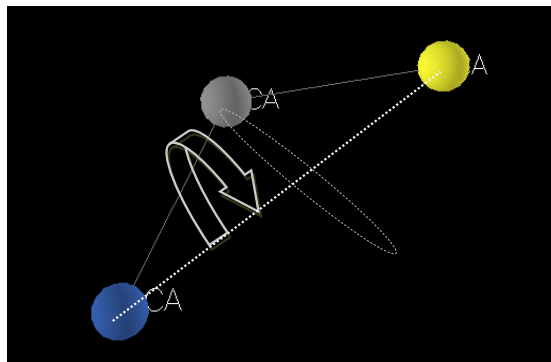


Figura 16 - Estratégia de movimentação - Agentes C-Alfa: Movimentam-se ao longo de uma circunferência que mantém constante a distância entre C-Alfas adjacentes.

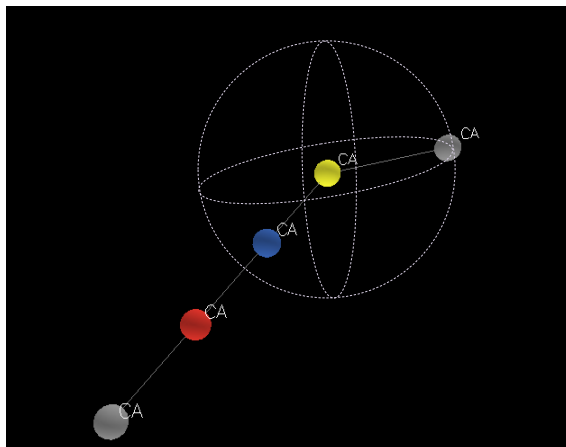


Figura 17 - Estratégia de movimentação - Agentes C-Alfa: Quando estão no início ou fim da sequência de aminoácidos, o agente verifica a distância  $D$  dele para o agente adjacente e é criada uma esfera de raio  $D$  na volta do agente adjacente. Posteriormente o agente escolhe sua nova localização em algum ponto (com igual probabilidade) da superfície da esfera.

*Vizinhança:* Em Bortolussi *et al.* (2007), para evitar sobrecarga na comunicação dos agentes durante a simulação, a lista de vizinhos de cada agente não é atualizada a todo o momento no espaço de tupla e sim a cada  $M$  números de movimentações, onde  $M$  vale 100. Em nossa abordagem, dada as características do software NetLogo, onde a vizinhança é atualizada automaticamente, sempre que um agente se movimenta, a vizinhança é atualizada, portanto não houve preocupações quanto à sobrecarga na comunicação.

A Figura 18 contém o pseudocódigo para os agentes C-Alfa (e também C-Beta).

### 5.2.2 Agentes tipo C-Beta – Nível 1

Os agentes do tipo C-Beta possuem a possibilidade de serem definidos como ligados/desligados, por intermédio de um botão na interface. Um agente C-Beta é implementado como um carbono adicional a um respectivo agente C-Alfa e possui o mesmo *residue-nr* e *atom-type*.

*Movimentação:* É possível escolher a estratégia (Cubo ou Distância Fixa) dos agentes C-Betas independentemente de como os agentes C-Alfa estiverem se movimentando. A hora de se movimentar é o único momento em que os agentes C-Beta trocam informações (relacionadas a posicionamento) com outros agentes. Agentes C-Beta não calculam energias.



---

**Algoritmo 1:** chama\_agentes\_C-Alfa()
 

---

```

1 início
2   energia_atual = computar_energia_local();
3   escolhe_angulo();
4   escolhe_pontos();
5   escolhe_estrategia_movimentacao();
6   movimenta();
7   energia_nova = computar_energia_global();
8   monte_carlo(energia_atual, energia_nova);
   /* Aceita ou não o movimento */
9 fim

```

---

Figura 18 – Pseudocódigo para o funcionamento dos agentes C-Alfa e C-Beta

### 5.2.3 Agente tipo Diretor – Nível 2

Possui conhecimento total sobre a atual conformação da proteína e é responsável por coordenar os agentes aminoácidos e tornar mais eficiente a exploração espacial. O agente Diretor não possui representação no espaço cartesiano e deve ser entendido como um agente que atua na simulação de fora do espaço 3D. O Diretor movimenta os agentes aminoácidos de acordo com estratégias diferenciadas. Foram implementadas duas estratégias, *crankshaft* (Figura 19) e *pivot* (Figura 20). Ambos os movimentos foram implementados via matrizes de rotação e operações de translação e ambos mantêm a distância entre átomos inalterada e possibilitam superar barreiras energéticas introduzidas pela função de energia. A ativação desses movimentos acontece somente quando a temperatura do sistema é alta o suficiente em relação a certo limiar, o que é suficiente para garantir uma fácil superação de barreira energética assim como uma exploração efetiva do espaço de busca conformacional.

Quando ativado, o agente Diretor chama-se recursivamente um número *total\_orch\_moves* de vezes e em cada passo escolhe aleatoriamente qual movimento realizará (*crankshaft* ou *pivot*), os pontos alvo e um ângulo. Após movimentar a cadeia, o agente utiliza-se do método de Monte Carlo para aceitar ou não a nova conformação (com base nas energias potenciais de cada configuração 3D) e então dorme por um número de *ticks* (passos de tempo) *orchestra\_sleep\_time* estipulado pelo usuário (valor padrão igual a 1). Na Figura 21 temos, em pseudocódigo, a representação de como o agente Diretor funciona.

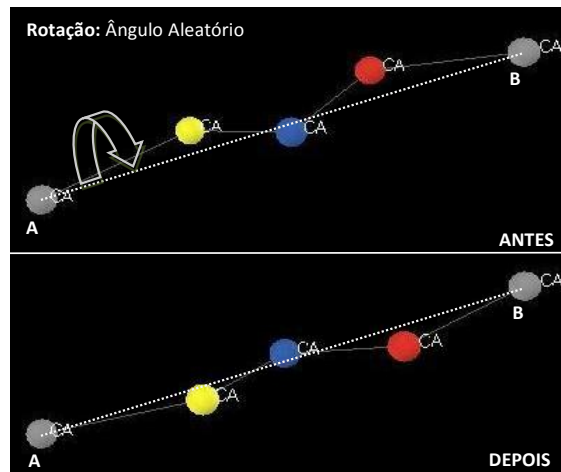


Figura 19 - Movimento *Crankshaft*: Primeiramente escolhe dois aminoácidos A e B da cadeia (a uma distancia de três aminoácidos). Posteriormente aplica rotação aos aminoácidos que separam A e B escolhendo um ângulo aleatório, tendo como eixo a reta AB.

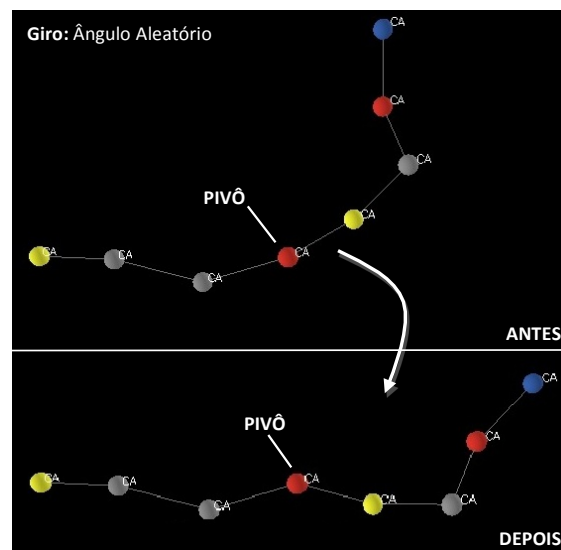


Figura 20 - Movimento *Pivot* : Escolhe um ponto (aminoácido pivô) e gira uma parte da cadeia em torno desse ponto, utilizando um ângulo aleatório.

---

**Algoritmo 2:** chama\_agentes\_Diretores()

---

```

1 início
2   se descansando = FALSO então
3     se temperatura > limiar_diretor então
4       se nr_de_movimentos < max_movimentos então
5         enquanto nr_de_chamadas < total_movimentos_diretor faça
6           energia_atual = computar_energia_global();
7           escolhe_angulo();
8           escolhe_pontos();
9           escolhe_estrategia_movimentacao();
10          movimenta();
11          energia_nova = computar_energia_global();
12          monte_carlo(energia_atual, energia_nova);
13          /* Aceita ou não o movimento */
14          nr_de_chamadas++;
15        fim enquanto
16      fim se
17    dorme();
18  fim se
19 fim

```

---

Figura 21 - Pseudocódigo, funcionamento do agente Diretor.

#### 5.2.4 Agente tipo Ambiente – Nível 3

O agente Ambiente controla todo o sistema, possui variáveis como *old\_average\_attempted\_moves*, *old\_average\_accepted\_moves*, *average\_attempted\_moves* e *average\_accepted\_moves*, que ajudam a controlar quando cada agente deve ser chamado. Além disso, efetua o controle dos passos de temperatura da simulação seguindo as requisições impostas pelo método de arrefecimento simulado (Capítulo 2, Seção 5) e controla os fatores de aceitação de movimentos que aumentam a energia do sistema. Por fim é o responsável pelos critérios de parada da simulação. Os critérios de parada são dois:

- Parada por não melhoria na energia: A cada passo de temperatura o agente Ambiente verifica (de acordo com um limiar mínimo que pode ser atribuído pelo usuário) se a energia total do sistema se alterou ou não significativamente, se a resposta for negativa, o agente incrementa uma variável e, quando essa variável chega a certo valor a simulação é tida como estável de mais para continuar e é finalizada, e

- Parada por alcance da temperatura mínima permitida: O segundo critério de parada é ativado quando o sistema chega até um valor *min\_temperature\_allowed* de temperatura. Quando isso acontece, o agente Ambiente ativa uma nova estratégia de movimentação e aceitação de movimentos (*go-zero-temp*) e, ao término da mesma, encerra a simulação.

---

**Algoritmo 3: chama\_agente\_Ambiente()**


---

```

1 início
  /* Se não há variação considerável na energia por um número máximo
  arbitrário de passos de temperatura */
2 se nao_variacao_energia >= max então
3   temperatura = 0;
4 fim se
  /* Se a temperatura atinge um valor abaixo de certo limiar
  limiar_diretor, os testes de aceitação de movimentos que aumentam a
  energia passam a rejeitar sempre as energias mais altas, dando início a
  um processo de minimização de energia por parte do agente Diretor. */
5 se temperatura < limiar_diretor então
6   fator_temperatura_diretor = 1x10^-20
7 fim se
8 se temperatura < temperatura_minima_permitida então
9   go-zero-temp(); /* Ativa a estratégia para dar início ao processo
  de término da simulação */
10 fim se
11 enquanto numero_de_movimentos < max_movimentos faça
12   chama_agentes_Diretores();
13   chama_agentes_C-Alfa();
14   chama_agentes_C-Beta();
15 fim enqto
16 temperatura = temperatura * 0.98
  /* Decrementa a temperatura de forma logarítmica (Arrefecimento
  Simulado) */
17 se mudanca_consideravel_energia = FALSO então
18   sem_variacao_energia++;
19 fim se
20 chama_agente_Ambiente(); /* Chama a recursão */
21 fim

```

---

Figura 22 - Pseudocódigo, funcionamento do agente Ambiente.

Outra função importante do agente Ambiente é a de gravar informações. A cada passo de temperatura, o agente Ambiente se preocupa em guardar a conformação atual do sistema, juntamente com a energia global atual, o percentual atual de aceitação de movimentos e RMSD com referência à estrutura experimental (se existir). Além disso, ao termino da simulação, o agente disponibiliza um gráfico da energia em função do tempo, o qual pode ser analisado para visualizar o comportamento da proteína ao longo de toda a simulação. O gráfico pode ajudar inclusive a verificar se o mínimo energético encontrado ao final da simulação é realmente o menor valor entre todos outros testados ao longo da simulação. Caso existam outros mínimos, estes podem se tornar potenciais estruturas a serem analisadas. Assim como os agentes C-Alfa, o agente Ambiente também possui uma energia relacionada ao agente Diretor, porém diferentemente do que acontece com os agentes C-Alfa, essa energia é uma energia global e, para calculá-la, os agentes utilizam-se de informações (tipo, distancia, posição) provenientes de todos agentes C-Alfa e C-Beta presentes na simulação. A Figura 22 nos trás o pseudocódigo do agente Ambiente.

### 5.3 Requisitos

No que se dispõe a requisitos do sistema, elencamos:

- É necessário ter o pacote Ruby com a biblioteca Watir instalado;
- É necessário ter Python instalado, juntamente com o pacote NumPy e
- Java(TM) SE Runtime Environment e Java HotSpot(TM) 64-Bit Server VM.

### 5.4 Interface

As Figuras 23, 24 e 25 mostram a interface atual do PROSMART. Embora a Interface do sistema seja unificada (apenas uma janela), para vias de melhor entendimento, aqui dividiremos a interface em três partes: a primeira, que passaremos a chamar de Interface I corresponde aos botões, campos e monitores relacionados com o funcionamento geral da simulação, a segunda parte (Interface II) corresponde a gráficos atualizados em tempo real e, a terceira parte (Interface III) corresponde a parte relacionada às configurações relacionadas especificamente com a estratégia de busca conformacional. Além disso, existe também uma

segunda janela (Figura 24) onde a conformação 3D da proteína e de suas alterações conformacionais podem ser visualizadas. Por meio da Interface I, o usuário pode:

- Escolher o nome do arquivo de entrada e saída de dados;
- Escolher o tamanho dos átomos (para visualização);
- Reiniciar todo o sistema;
- Imprimir (em formato PDB) a conformação atual em que os agentes se encontram;
- Criar interativamente, clicando nos botões de cada aminoácido, os agentes (cadeia) da proteína;
- Requisitar uma predição de estrutura secundária para uma cadeia específica de aminoácidos (valor contido na string “Fasta-Seq”);
- Estipular uma sequência caracterizando as informações a cerca de estruturas secundárias a serem utilizadas pelo sistema;
- Criar uma sequência Fasta diretamente da visualização;
- Requisitar que seja gerado um arquivo PDB contendo não só a conformação atual formada pelos agentes presentes na simulação, mas um arquivo PDB contendo também as cadeias laterais da proteína;
- Carregar uma sequência de aminoácidos (digitada pelo usuário ou fornecida por um arquivo) diretamente para a visualização, por exemplo, utilizando um arquivo contendo a cadeia string “GGG”, o que teria como resultado três agentes representantes de glicina conectados;
- Criar os agentes C-Beta;
- Criar os agentes Diretor e Ambiente;
- Rodar o sistema e
- Fazer o cálculo de RMSD entre duas estruturas.

Além disso, a interface conta com monitores (nomenclatura proveniente do NetLogo) que facilitam o controle do número de átomos/agentes presentes na simulação, o percentual de aceitação dos movimentos, a energia global atual e o passo de temperatura atual.

A Interface II conta com gráficos relacionados à Energia, RMSD, GDT e proporção de aceitação de movimentos. Os gráficos são atualizados em tempo real e tem como finalidade prover ao usuário uma melhor noção do progresso da simulação.

Por meio da Interface III, o usuário pode estipular valores para variáveis referentes ao número de movimentos, ao controle do agente Diretor, dos agentes C-Alfa e C-Beta e agente

Ambiente, além de configurações sob os fatores que regem os métodos de Monte Carlo envolvidos na simulação. Assim sendo, pode-se:

- Escolher uma temperatura a partir da qual abaixo do valor estipulado o Diretor pára de ser executado;
- Escolher o intervalo mínimo e máximo presente no movimento de crankshaft executado pelo Diretor;
- Escolher o intervalo mínimo e máximo presente no movimento de crankshaft executado pelo Diretor;
- Escolher o número mínimo de aminoácidos a serem movidos durante o movimento pivot executado pelo agente Diretor;
- Escolher a quantidade de tentativas de movimento a serem levadas em conta durante a simulação seja quando o agente Diretor está “ligado” ou “desligado” ou quando a simulação encontra-se na menor temperatura permitida (quando acontece uma minimização da energia, atribuindo-se um valor extremamente pequeno para os fatores de Boltzmann presentes nos testes de aceitação dos métodos de Monte Carlo);
- Escolher o ângulo máximo a ser levado em consideração no momento em que os agentes se movimentam. É importante destacar que este parâmetro afeta também as configurações gerais do sistema, ou seja, caso se opte por trocar a função de energia a ser utilizada, este parâmetro permanecerá pois também é levado em conta na movimentação dos agentes C-Alfa e C-Beta;
- Escolher se a função de energia utilizará ou não a predição de estrutura secundária para auxiliar a simulação e, caso positivo, se será permitido ao agente Diretor executar ou não movimentos que envolvam os aminoácidos que possuam estejam marcados como hélices ou fitas;
- Escolher o tipo de movimento padrão a ser executado pelos agentes C-Alfa e C-Beta;
- Especificar quais os valores dos fatores de Boltzmann específicos primeiramente aos agentes C-Alfa e C-Beta e posteriormente ao agente Diretor;
- Especificar a temperatura mínima aceita pelo sistema;.
- O número máximo de não melhorias na energia (relacionado ao término da simulação);
- O percentual utilizado para a diminuição gradativa da temperatura, juntamente com um valor para a variação mínima a ser levada em conta e

- Por fim podemos também escolher qual peso daremos as punições em termos de energia dadas em relação às estruturas secundárias da proteína;

**Setup commands:**

title: 1EDP  
 atom-size: 1.95  
 #NETLOGOPATH: /home/tpaes/netlogo-5.0  
 Reference-Structure: /home/tpaes/netlogo-5.0/inputs/EDP-2.pdb

**Workflow:**

SSP-Seq: CCCCCCCCCCCCCC  
 Fasta-Seq: GCSCSSLMDKECVYFCHLG

**Interactive AA Creator**  
 Non Polar, aliphatic: GLY, VAL, ALA, LEU, PRO, ILE, MET, PHE, TYR, TRP  
 Polar, uncharged: SER, CYS, THR, ASN, GLN  
 Positively charged: LYS, HIS, ARG  
 Negatively charged: GLU, ASP

**RMSD Calculator**  
 PDBReference: ALAS-30-AUGUST-Num1-7.pdb  
 PDBTarget: ALAS-30-AUGUST-Num1-9.pdb  
 RMSD Calc: RMSD

**Nr of Agents**

c-alpha Agents	Environment Agents
19	1
c-beta Agents	Director Agents
19	1

**Second. Structure Predictor**  
 Submit SS Predictor From Fasta-Seq  
 email: thiagopaes@gmail.com

**Energy Calculation**

Tick	energia	RMSD	GDT
229	-18.644	4.4	0

Calculate Energy: temperature 0

Figura 23 - Interface I. Parte da interface na qual o usuário configura a proteína a ser modelada e controla a execução da simulação.



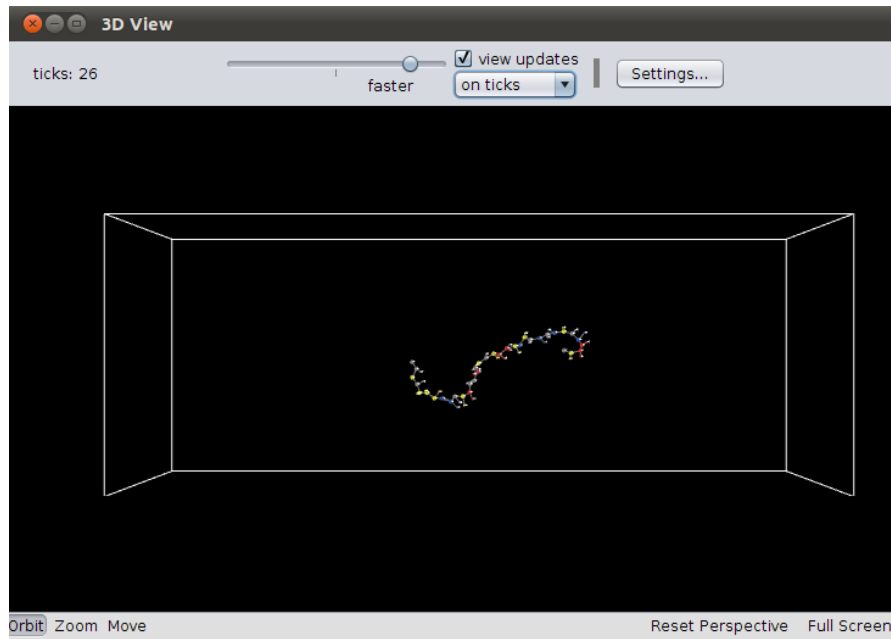


Figura 24 - Visualização 3D. Por intermédio de uma janela separada é possível verificar a conformação atual da proteína que está sendo modelada e, em tempo real, acompanhar as modificações estruturais que estão ocorrendo em virtude das movimentações dos agentes.

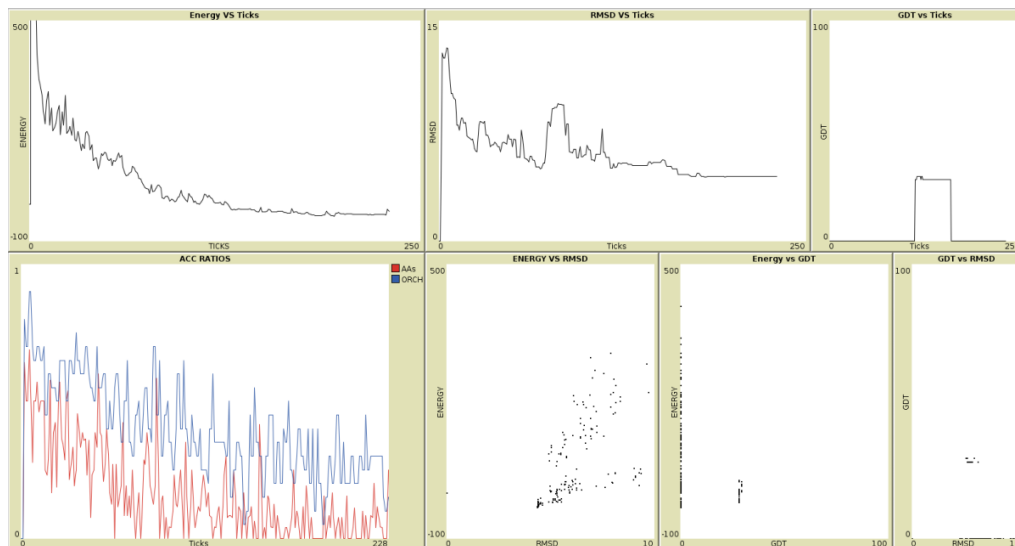


Figura 25 - Interface II. Parte da interface na qual o usuário pode visualizar em tempo real gráficos referentes ao comportamento da simulação.

The screenshot displays the NetLogo interface with several panels of pseudocode and control parameters:

- Attempted Moves Control:** Shows a temperature slider set to 0, a `temp_thr` slider at 0.10, and a `max_moves` slider at 50. It also includes sliders for `attempted_threshold_without_orchestra` (32) and `attempted_threshold_with_orchestra` (15).
- Orchestra Agent Pseudocode:** Contains logic for checking if resting, computing global energy, choosing points and strategies, and moving. It includes a `while` loop for `count_orch` and a `refresh` button.
- Environment Agent Pseudocode:** Contains logic for checking energy variation, stopping if temperature is 0, and updating temperature based on a decrease ratio.
- Control Parameters:** A central panel with sliders for `count_orch` (20), `Delta` (50), `min_crank` (2 AA), `max_crank` (5 AA), `min_pivot_dist` (2 AA), and `max_angle` (90 Degrees). It also has a `Refresh` button.
- C-alfa and C-beta Agents Pseudocode:** Contains logic for asking agents to compute energy and choose points.
- Monte Carlo Factors:** Includes sliders for `fator_temp_buscadores` (1.0000000000000000) and `fator_temp_orch` (1.00), along with initial values for Cas and Cbs.
- Energy Parameters:** Includes sliders for `energy_variation_threshold` (1.0E-4) and `temp_orch_thr` (1.00), and a `SecCoopMult` slider (80).
- Moves:** A panel with buttons for `Moves` (Ras), `Move-C-Alphas` (yes), and `Move-C-Betas` (yes).

Figura 26 - Interface III. Parte da interface na qual o usuário pode configurar a estratégia de busca da conformação de menor energia, assim como o tipo de movimentação que os agentes devem adotar.

Existe também, junto à Interface, uma pequena janela contendo o centro de comandos, por onde o usuário pode interagir por intermédio de comandos de texto na linguagem NetLogo.

## 5. 5 Execução

Com o intuito de elucidar as etapas necessárias para a utilização da ferramenta será apresentado um passo a passo contendo informações sobre como configurar e iniciar a execução da predição de proteínas:

Passo 1 – Preencher o campo “Fasta-Seq” com a sequência de aminoácidos da Proteína alvo. O preenchimento pode ser feito de duas maneiras: caso o usuário já possua a sequência de aminoácidos da proteína, ele pode apenas digitá-la no campo ou, caso não a possua, o usuário pode interativamente criar a proteína clicando em aminoácido por aminoácido. Caso escolha o método interativo, ao finalizar a criação da proteína (que ficará em um formato linear estendido) o usuário deverá clicar no botão “Create Fasta-Seq from View” para que o campo “Fasta-Seq” seja preenchido.

Passo 2 – Cooperação via estruturas secundárias. Caso o usuário opte pela utilização da cooperação via estruturas secundárias, ele deve preencher o campo “SSP-Seq” com as informações relativas à estrutura secundária da proteína em forma de cadeia de caracteres. Caso o usuário não possua essa string de informações, ele tem a opção de, preenchendo o campo “email” com seu e-mail, receber via correio eletrônico a predição de estrutura secundária para a proteína em questão, clicando no botão “Submit SS Predictor From Fasta-Seq”. Além disso, clicando no botão “SSP\_TYPE” o usuário poderá escolher qual abordagem utilizará em termos de peso para a contribuição da cooperação via estruturas secundárias durante a simulação.

Passo 3 – PDB de referência. O usuário deve preencher o caminho contendo a estrutura de referência (cristalográfica) da proteína alvo (caso exista). Essa estrutura será utilizada para os cálculos de RMSD e GDT a serem feitos durante a simulação.

Passo 4 – Pasta de saída. O campo “saída” deverá ser preenchido com o nome da pasta que será criada para guardar os arquivos provenientes da simulação.

Passo 5 – Startup. Clicando no botão “Startup” o usuário reiniciará o sistema redefinindo o tamanho da caixa 3D de simulação (calculado com base no tamanho da proteína). Esta etapa pode ser feita por intermédio de um atalho de teclado, a tecla “Q”.

Passo 6 – Carregando a proteína dentro do ambiente 3D. Uma vez que o ambiente 3D está pronto, o usuário deve clicar no botão “Load Fasta Seq into View” para que a proteína (agentes C-Alfa) sejam carregados dentro do ambiente 3D. Esta etapa pode ser feita por intermédio de um atalho de teclado, a tecla “A”.

Passo 7 – Criando agentes C-Beta. Em seguida o usuário deverá clicar no botão “Create C-Betas” para que os agentes C-Beta sejam criados e carregados dentro do ambiente 3D. Esta etapa pode ser feita por intermédio de um atalho de teclado, a tecla “S”.

Passo 8 – Criando agentes Diretor e Ambiente. Clicando no botão “Create director & environment Agents” o usuário criará os dois últimos agentes que faltavam para que o sistema estivesse completo. Esta etapa pode ser feita por intermédio de um atalho de teclado, a tecla “D”.

Passo 9 – Iniciar simulação. Após o Passo 8 ter sido executado com sucesso, o sistema estará pronto para execução e, para dar início a simulação, ativaremos o agente Ambiente clicando no botão “Environment Agent – Go”. Esta etapa pode ser feita por intermédio de um atalho de teclado, a tecla “F”

Passo 10 – Final da simulação. Quando a simulação chegar ao seu final, será exibida na tela uma mensagem alertando o usuário. Neste momento o usuário poderá verificar a pasta de outputs do PRO-SMART onde irá constar uma nova pasta com o nome escolhido no Passo 4. Nela estarão todos os arquivos de saída da simulação como gráficos contendo o comportamento da energia, GDT e RMSD durante a simulação, os arquivos *.pub.* de cada passo de temperatura ou *tick*, além dos arquivos de saída do método de clusterização incorporado ao PRO-SMART para escolha da melhor estrutura.

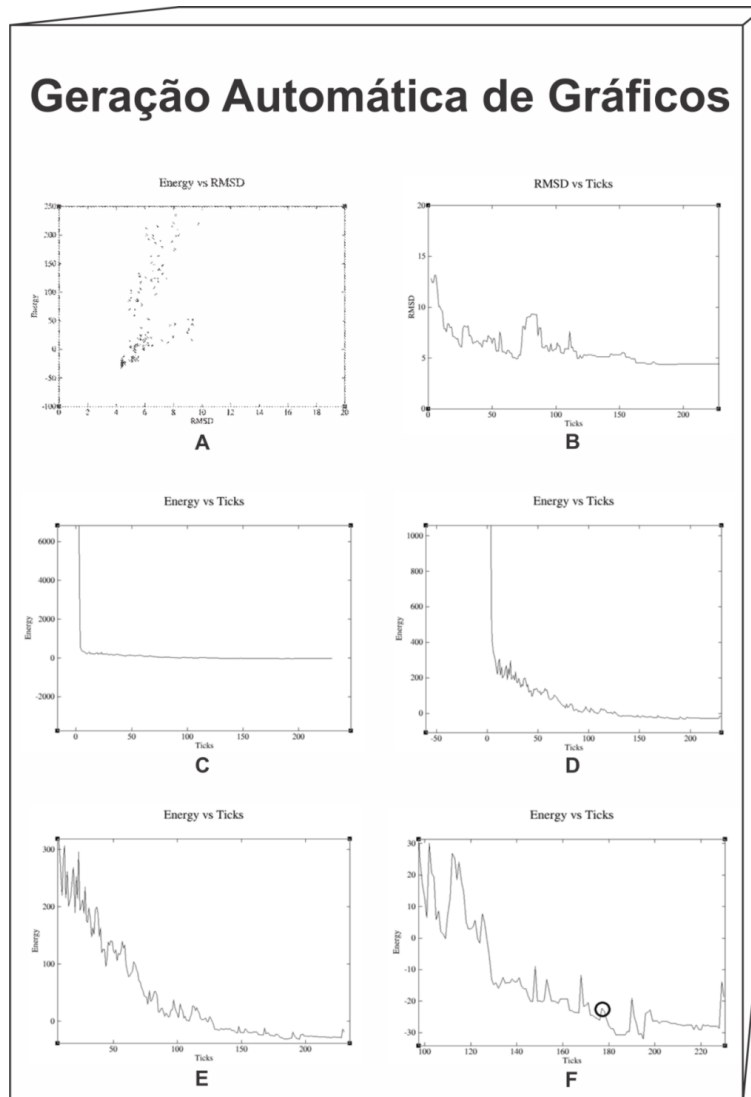


Figura 27 - Exemplos de gráficos gerados automaticamente pelo PRO-SMART. Em (A) podemos verificar a relação de Energia x RMSD ao longo da simulação. Em (B) podemos notar as flutuações de RMSD do início ao fim da simulação e, em (C), (D), (E) e (F) as flutuações da Energia durante a simulação, dando ênfase ao tamanho das mudanças em termos de energia em diferentes fases da simulação.

Na Figura 27 A podemos analisar como foi o comportamento da energia em relação ao RMSD durante a simulação. Na Figura 27 A fica claro que, nesta simulação, os menores valores de RMSD foram obtidos justamente nas menores energias alcançadas, já a Figura 27 B demonstra o comportamento do RMSD ao longo da simulação, onde se pode notar uma gradativa diminuição. Na Figura 27 C podemos verificar que a energia durante a simulação

parece se tornar linear, entretanto precisamos apenas nos aproximar mais para notar as alterações em termos de energia que ocorreram. Modificando-se a escala do eixo y pode-se ter uma melhor noção da simulação. Na Figura 27 D é possível notar que, entre o passo de temperatura 0 e o passo de temperatura 100 o sistema sofreu alterações de energia maiores em comparação às alterações posteriores ao *tick* 100. Confirmando que o sistema está se comportando conforme o esperado, ou seja, superando barreiras energéticas maiores nos primeiros passos de temperatura. Novamente modificando a escala y (energia) é possível notar que existem consideráveis alterações energéticas também em estados referentes a passos de temperaturas mais próximos do final da simulação (Figura 27 E). Por fim, na Figura 27 F estão dispostos apenas valores referentes ao *tick* 100 e superiores, os quais são utilizados pelo algoritmo de clusterização para a escolha do modelo que representará a proteína. O modelo escolhido após o protocolo de clusterização está em destaque. É importante notar que a conformação escolhida não é, necessariamente, a última conformação adotada pela simulação ou aquela de menor energia.

## 6. RESULTADOS

O capítulo a seguir tem o intuito de expor os resultados alcançados pela ferramenta PRO-SMART até então. Será exposto o conjunto de proteínas alvo das simulações, os resultados obtidos pelas abordagens sem e com a utilização de cooperação por estruturas secundárias seguidas de uma comparação do PRO-SMART com o sistema de Bortolussi *et al.* e, por fim, é feita uma análise geral do que se obteve.

### 6.1 Conjunto de Proteínas Alvo

Da análise de trabalhos relacionados um conjunto de polipeptídeos alvo foi criado a fim de tornar possível não só a verificação do potencial de aplicação da ferramenta como também a comparação de resultados do PRO-SMART com os resultados do trabalho base utilizado para a sua criação. A seguir na Tabela 1 e Figura 28 estão dispostas as proteínas escolhidas, com seus identificadores no PDB e a quantidade de aminoácidos presente em cada uma, além de informações sobre o tipo de estrutura secundária que possuem. Proteínas que possuem somente hélices são da classe  $\alpha$ , proteínas que possuem somente folhas são da classe  $\beta$  e proteínas que possuem hélices e folhas, da classe  $\alpha\beta$ .

É importante atentar que, com o intuito de obter mais precisão no cálculo de energia das proteínas em questão, optou-se pela adição de aminoácidos do tipo glicina (sem cadeia lateral) tanto no começo das proteínas quanto no final, assim aumentando em dois As o número de aminoácidos usados nas simulações (se comparado aos números contidos na tabela abaixo). O número do resíduo ou *residue nr* é um campo dos arquivos no formato PDB o qual é utilizado na geração dos diagramas de Ramachandran e, por isso, em nossas análises, quando fizermos referências ao início e fim de hélices, por exemplo, e seus respectivo *residue nrs*, haverá uma pequena diferença em relação às informações contidas no PDB. Outra informação importante de ser salientada diz respeito às análises feitas, as quais utilizaram como fonte de informações sobre estrutura secundárias o DSSP [9].

Tabela 1 - Conjunto de Proteínas alvo de simulações, junto de seu PDB ID, do artigo publicado que a descreve, número de aminoácidos que possui e classe.

PDB ID	Artigo de referência	Nº de AAs	Classe
1LE0	[19]	12	$\beta$
1KVG	[67]	12	$\beta$
1LE3	[19]	17	$\beta$
1EDP	[4]	17	$\alpha$
1PG1	[28]	18	$\beta$
1ZDD	[69]	34	$\alpha$
1VII	[55]	36	$\alpha$
2GP8	[70]	40	$\alpha$
1ED0	[64]	46	$\alpha\beta$

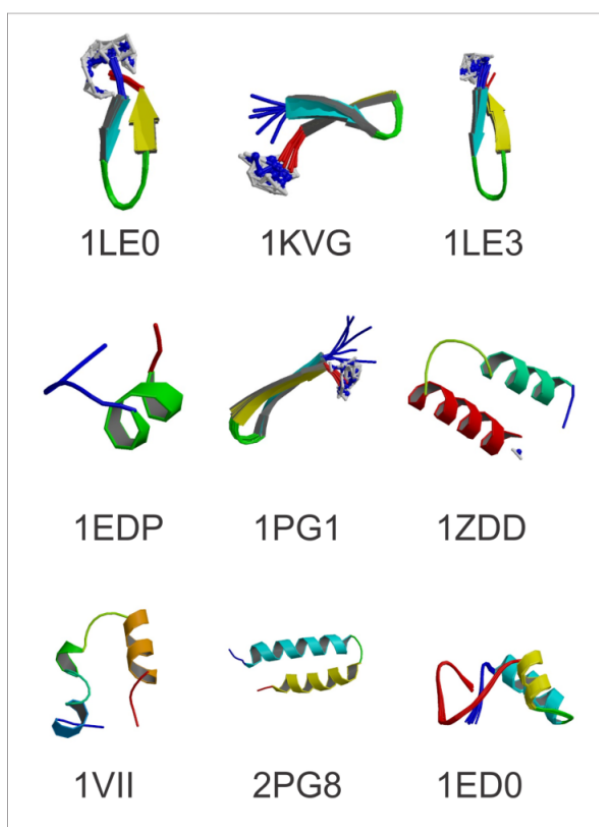


Figura 28 - Estruturas 3D das proteínas alvo de simulação.



## 6.2 Simulações Sem Cooperação via Estruturas Secundárias.

### 6.2.1 Configuração

A Tabela 2 trás dados referentes à configuração do PRO-SMART para as simulações sem cooperação via estruturas secundárias. Testes iniciais feitos durante a dissertação (utilizados como exemplos nos resumos publicados em anais de congresso [50]) contribuíram para o ajuste do valor de variáveis do sistema. Uma informação a destacar é a de que a estratégia de movimentação com distâncias fixas se mostrou pouco efetiva, levando a mínimos locais e, isto posto, os testes (tanto com a cooperação por estruturas secundárias ativada quanto desativada) utilizaram a estratégia de movimentação por cubo.

Tabela 2 - Conjunto de variáveis e funcionalidades configuradas nas simulações com cooperação via estruturas secundária desativada.

Variável/Funcionalidade	Valor
temp_thr	0,10
attempted_threshold_without_orchestra	32
attempted_threshold_with_orchestra	13
temp_zeromoves	50
total_orch_moves	20
orchestra_sleep_time	1
min_crank	2 AAs
max_crank	5 AAs
min_pivot_dist	2 AAs
max_angle	90°
moves	Cubo
move-C-Alphas	Yes
move-C-Betas	Yes
fator_temp_buscaadores	1
fator_temp_orch	1
max_number_of_no_improvement_in_energy	6 Ticks
min_temperature_allowed	0,099
temp_decrease_ratio	0,98
energy_variation_threshold	$1 \times 10^{-4}$
temp_orch_thr	1
initial_temperature	10
SSP	OFF

### 6.2.2 Resultados Obtidos

As tabelas a seguir demonstram o desempenho do PRO-SMART levando-se em conta diferentes formas de avaliação de similaridade de proteínas. Com base na tabela de cada

proteína foram analisados os resultados de cada simulação e, deixando o RMSD de lado, buscou-se encontrar as estruturas que, segundo os outros critérios de avaliação, alcançaram resultados acima da média. As simulações que teoricamente obtiveram maior êxito no que se trata da avaliação geométrica da proteína estão em destaque. Uma vez feita a análise da geometria adotada pelas conformações, parte-se para a avaliação estereoquímica. Em um primeiro momento, optou-se por gerar os diagramas de Ramachandran e analisar todas as simulações, entretanto essa abordagem se mostrou ineficaz devido à grande diferença nos resultados aliada a excessiva quantidade de dados disposta no diagrama, o que notavelmente prejudica a assimilação da informação. O mapa de Ramachandran citado acima pode ser encontrado no Apêndice B.

Optou-se então pela análise estereoquímica apenas das estruturas de maior êxito em se tratando de geometria.

Tabela 3 - Resultados obtidos para as diferentes medidas de similaridade em relação à proteína de PDB ID: 1EDP, sem cooperação via estrutura secundária.

Simulação	1	2	3	4	5	6	7	8	9	10	Média ( $\bar{O}$ )	Desvio ( $\sigma$ )
RMSD (Å)	7,80	6,20	6,10	4,40	5,80	8,10	4,70	4,80	4,40	7,20	5,95	1,39
TM-Score	0,09	0,08	0,15	0,13	0,12	0,15	0,14	0,19	0,21	0,13	0,14	0,04
MaxSub	0,31	0,34	0,28	0,40	0,35	0,36	0,39	0,43	0,41	0,40	0,37	0,05
GDT_TS	0,41	0,51	0,49	0,56	0,51	0,46	0,57	0,56	0,54	0,56	0,52	0,05
GDT_HA	0,26	0,31	0,31	0,35	0,32	0,32	0,35	0,38	0,37	0,35	0,33	0,03
Energia	-37,21	-15,12	-16,75	-21,82	-24,60	-28,99	-46,50	-25,92	-25,67	-45,85	-28,84	11,01
Tick	181	223	188	173	195	125	200	157	177	201	182	26

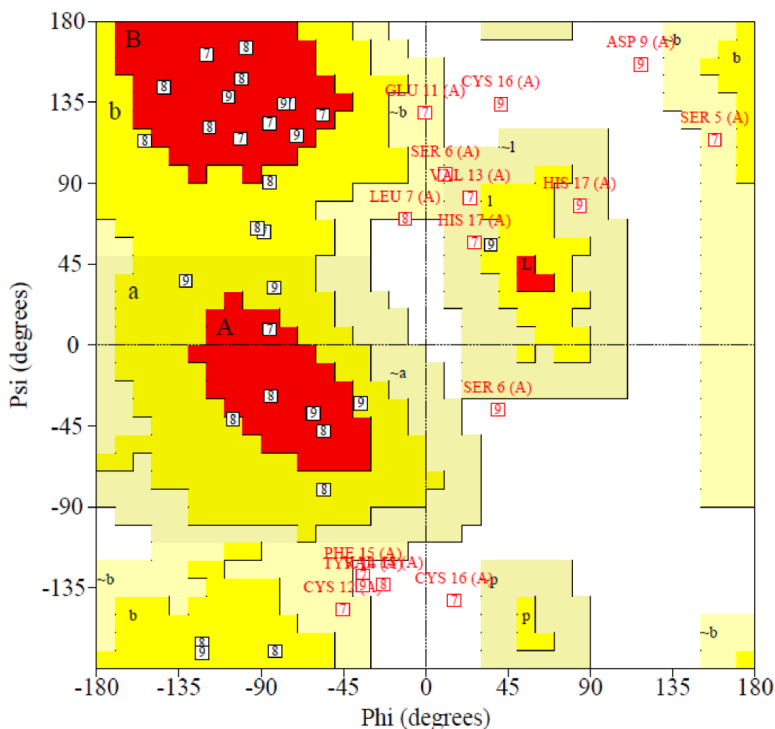


Figura 29 - Diagrama de Ramachandran contendo a análise estereoquímica das estruturas escolhidas como representantes das simulações 7, 8 e 9. Proteína de PDB ID: 1EDP.

A proteína 1EDP, composta por 17 resíduos de aminoácidos, possui uma hélice  $\alpha$  de sete resíduos (LYS10 a CYS16), flanqueada por regiões sem estrutura secundária regular definida: N-terminal (CYS1 a ASP9) e C-terminal (HIS16 a LEU17). O enovelamento desta proteína se dá pela formação de duas ligações dissulfídicas entre a hélice e a alça N-terminal (CYS1 a MET8) [5].

Análise dos modelos 200, 157, e 177, provenientes das simulações 7, 8 e 9, respectivamente, mostra que eles possuem o dobramento correto, entretanto no modelo proveniente da simulação 7, os resíduos que deveriam compor a hélice  $\alpha$  não se situam nas posições alfa do mapa de Ramachandran, diferentemente dos modelos provenientes das simulações 8 e 9. Esses resultados são corroborados tanto pelo mapa de Ramachandran quanto pelos valores de TM-Score. Comparando as simulações 8 e 9 e contabilizando os percentuais de resíduos dispostos em regiões não permitidas verifica-se que, no modelo da simulação 9 existem quatro resíduos em regiões não permitidas contra apenas um do modelo encontrado na simulação 8, o que nos poderia levar a dizer ser o modelo 8 o mais bem

adaptado modelo para servir como resultado em busca da estrutura nativa, entretanto, é preciso que se faça a análise de quais são estes resíduos que estão em locais não permitidos. Ao verificarmos quais são os resíduos fora de regiões permitidas encontramos: ASP9, TYR14, CYS16 e HIS17, onde ASP9 é um resíduo posicionado entre uma volta e uma hélice e CYS16 e HIS17 são, respectivamente, o último resíduo que compõe a hélice e seu conseqüente, o que caracteriza os três resíduos como passíveis de certa instabilidade. Já a TYR14 é um resíduo situado dentro da hélice o que de fato contribui para a diminuição da qualidade da predição, entretanto o mesmo acontece com o modelo proveniente da simulação 8, onde a VAL13 também se posiciona no interior da hélice e também, segundo o diagrama de Ramachandran, se situa em região não permitida. Ao avaliarmos resíduo por resíduo o mapa de Ramachandran da Figura 29 com comparação com a estrutura experimental, verificamos que em ambos os modelos a maioria dos resíduos que compõem a hélice da proteína posicionam-se nas devidas posições do mapa de Ramachandran. A seguir, na Figura 30, apresentamos a sobreposição da estrutura experimental e da estrutura predita pela simulação de número 9.

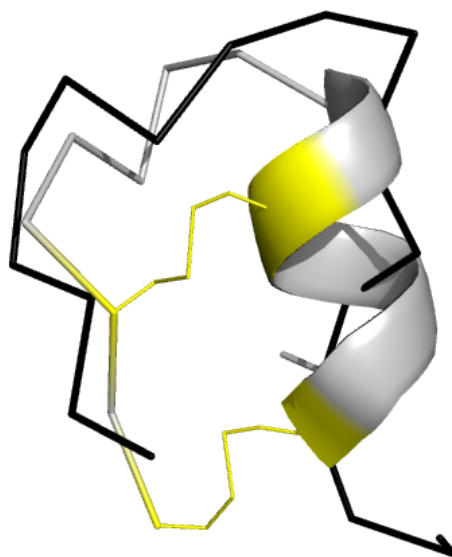


Figura 30 - Estrutura experimental (em cinza) e estrutura predita (em preto) pela simulação 9, proteína de PDB ID: 1EDP.

Tabela 4 - Resultados obtidos para as diferentes medidas de similaridade em relação à proteína de PDB ID: 1PG1, sem cooperação via estrutura secundária.

Simulação	1	2	3	4	5	6	7	8	9	10	Média ( $\bar{O}$ )	Desvio ( $\sigma$ )
RMSD (Å)	5,70	8,6	5,50	5,50	7,9	7,00	9,5	6,00	8,40	5,80	6,99	1,50
TM-Score	0,15	0,16	0,12	0,20	0,12	0,14	0,11	0,16	0,16	0,17	0,15	0,03
MaxSub	0,41	0,32	0,42	0,34	0,35	0,40	0,47	0,36	0,34	0,44	0,39	0,05
GDT_TS	0,50	0,40	0,53	0,54	0,44	0,50	0,49	0,53	0,46	0,54	0,49	0,05
GDT_HA	0,33	0,29	0,35	0,38	0,31	0,33	0,36	0,33	0,33	0,38	0,34	0,03
Energia	-46,55	-18,51	-51,79	-20,48	-42,40	-29,61	-36,83	-50,48	-33,69	-25,63	-35,60	12,07
Tick	208	179	191	194	211	212	211	201	190	170	196	14

A proteína 1PG1, composta por 18 resíduos de aminoácidos, possui duas fitas  $\beta$  de cinco resíduos (ARG5 à CYS9 e PHE13 à GLY18) e uma volta (ARG11 e ARG12). É rica em argininas e cisteínas, contendo seis argininas e quatro cisteínas que formam duas ligações dissulfídicas [28].

Quanto à análise estereoquímica (feita com base na Figura 32), esta demonstra apenas um resíduo situado em região não permitida, a CYS9. Verificando a estrutura 3D obtida (Figura 31) é possível notar que a CIS9, embora disposta em posição não permitida, por se situar no final da primeira fita, não influencia no dobramento da proteína. A estrutura 3D retrata bem a primeira fita, entretanto a partir da estrutura de volta a conformação levemente demonstra padrões diferentes da ER esperada.

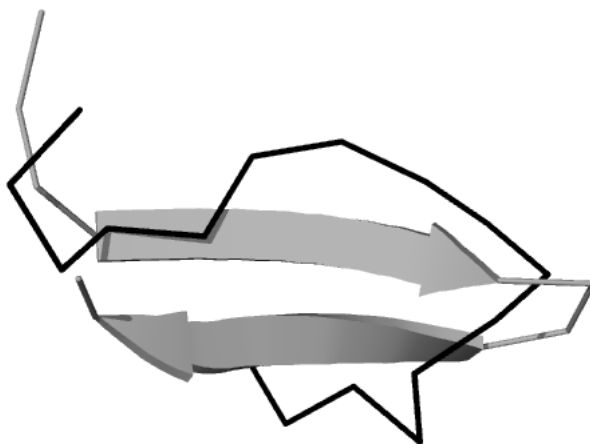


Figura 31 - Estrutura experimental (em cinza) e estrutura predita (em preto) pela simulação 10, proteína de PDB ID: 1PG1. Na sobreposição apenas dos resíduos que compõem as fitas temos: 1,29 Å para a primeira fita e 0,92 Å para a segunda.

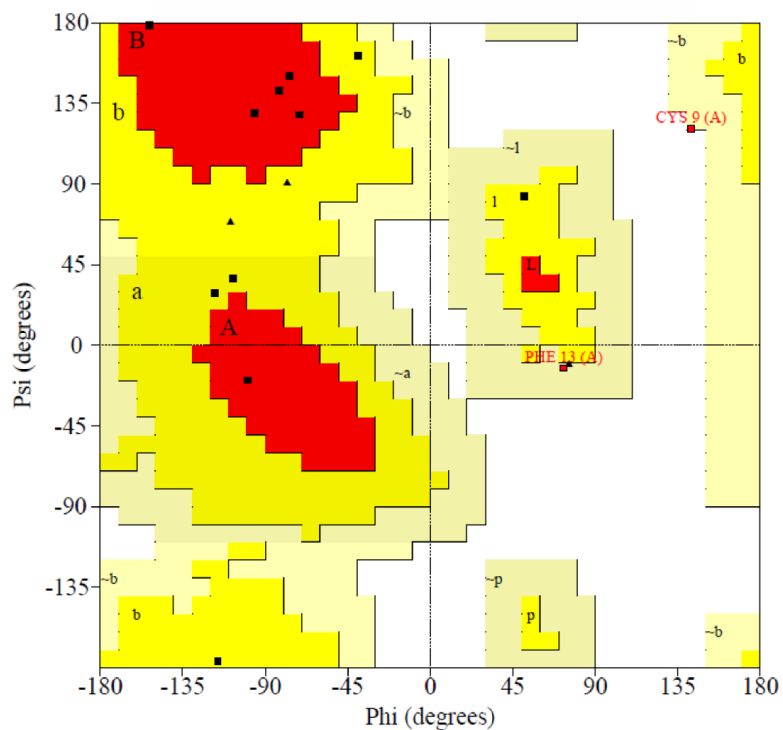


Figura 32 - Diagrama de Ramachandran do modelo obtidos na simulação 10. Proteína PDB ID: 1PG1. Cooperação via estrutura secundária desativada. Triângulos simbolizam glicinas.

Tabela 5 - Resultados obtidos para as diferentes medidas de similaridade em relação à proteína de PDB ID: 1LE0, sem cooperação via estrutura secundária.

Simulação	1	2	3	4	5	6	7	8	9	10	Média ( $\bar{O}$ )	Desvio ( $\sigma$ )
RMSD (Å)	6,60	6,10	6,90	3,80	8,40	6,40	5,60	7,40	8,00	4,50	6,37	1,45
TM-Score	0,25	0,18	0,16	0,14	0,25	0,23	0,23	0,21	0,21	0,24	0,21	0,04
MaxSub	0,46	0,48	0,39	0,46	0,46	0,55	0,46	0,45	0,46	0,54	0,47	0,05
GDT_TS	0,56	0,56	0,50	0,63	0,54	0,63	0,60	0,52	0,50	0,67	0,57	0,06
GDT_HA	0,46	0,38	0,35	0,42	0,44	0,50	0,46	0,44	0,40	0,48	0,43	0,05
Energia	-7,22	-12,42	-18,77	-14,18	-6,31	-18,96	-7,79	-16,00	-19,04	-16,96	-13,76	5,07
Tick	213	122	212	190	151	171	117	150	179	146	165	33

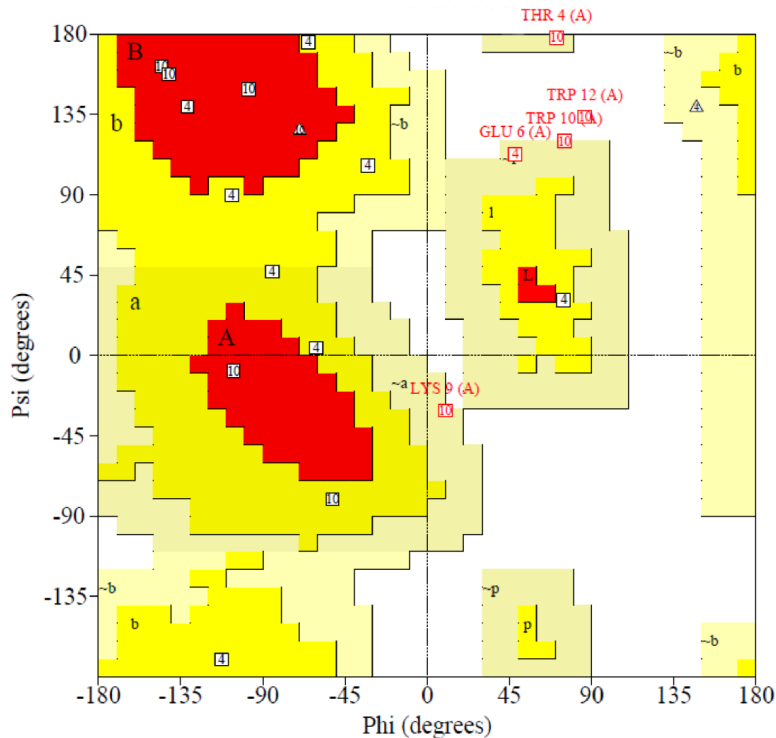


Figura 33 - Diagrama de Ramachandran dos modelos obtidos nas simulações 4 e 10. Proteína PDB ID: 1LE0. Cooperação via estrutura secundária desativada.

A proteína 1LE0 é formada um motivo estrutural chamado zíper de triptofano o qual estabiliza uma conformação de *beta hairpin* com duas fitas compostas pelos resíduos TRP-THR-TRP sendo a primeira fita formada pelos resíduos TRP3, THR4 e TRP5 e a segunda fita formada pelos resíduos de número TRP11, THR12 e TRP13 [19].

A análise do diagrama de Ramachandran do modelo 190 da simulação 4 (Figura 33) mostra que tal conformação não possui resíduos em regiões não permitidas, porém alguns resíduos e as ER indicadas pelo mapa possuem certas discrepâncias em relação à estrutura experimental, onde resíduos que deveriam ser classificados como fitas são classificados como hélices. A baixa resolução da função de energia utilizada nos dá margem a possibilidade da causa para tal acontecimento ter origem nas características peculiares da proteína em questão, a qual foi criada sinteticamente e possui como característica conformacional o já citado “zíper de triptofanos”. O mapa de Ramachandran para o modelo 146 proveniente da simulação 10 (também presente na Figura 33) mostra dois resíduos em regiões não permitidas e, a análise desses resíduos revela que são justamente resíduos relacionados à segunda fita, entretanto, os demais resíduos se dispõem em conformidade com a estrutura experimental.

Deste modo os dois modelos se mostraram capazes de adotar a mesma topologia da estrutura experimental (o que já era esperado tendo em vista os valores obtidos principalmente em termos de MaxSub e GDT), embora não tenham sido capazes de prever ambas as fitas de folha com propriedade. A Figura 34 representa a estrutura 3D do modelo 10 sobreposta à estrutura experimental levando-se em conta todos os átomos da proteína. Entretanto, ao verificarmos a sobreposição apenas dos resíduos que formam duas fitas que compõem a folha da proteína, alcançamos 0,03 Å e 0,4 Å para a primeira e segunda fita, respectivamente.

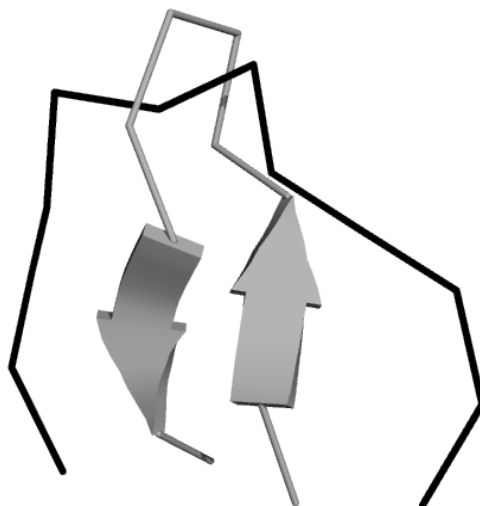


Figura 34 - Estrutura experimental (em cinza) e estrutura predita (em preto) pela simulação 10, proteína de PDB ID: 1LE0.

Tabela 6 - Resultados obtidos para as diferentes medidas de similaridade em relação à proteína de PDB ID: 1ZDD, sem cooperação via estrutura secundária.

Simulação	1	2	3	4	5	6	7	8	9	10	Média ( $\bar{O}$ )	Desvio ( $\sigma$ )
RMSD (Å)	9,30	8,63	8,04	8,10	8,70	8,50	7,40	9,20	8,70	7,87	8,44	0,59
TM-Score	0,16	0,16	0,14	0,16	0,14	0,13	0,15	0,13	0,14	0,15	0,15	0,01
MaxSub	0,18	0,19	0,17	0,22	0,18	0,22	0,22	0,17	0,18	0,19	0,19	0,02
GDT_TS	0,30	0,34	0,30	0,36	0,33	0,35	0,36	0,29	0,33	0,32	0,33	0,03
GDT_HA	0,19	0,18	0,17	0,20	0,18	0,21	0,21	0,17	0,18	0,18	0,19	0,02
Energia	-34,90	-29,93	-18,29	-61,64	-55,00	-25,00	-52,00	-57,00	-23,86	-71,92	-42,95	18,68
Tick	187	132	183	203	208	180	230	203	176	163	192	20



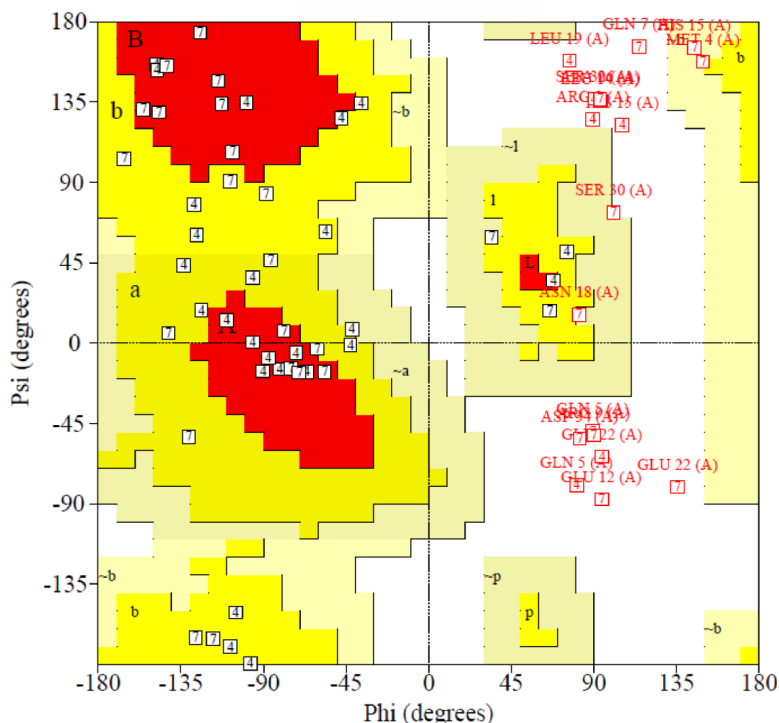


Figura 35 - Diagrama de Ramachandran dos modelos obtidos nas simulações 4 e 7. Proteína PDB ID: 1ZDD. Cooperação via estrutura secundária desativada.

A proteína 1ZDD, composta por 35 resíduos, possui duas hélices  $\alpha$ , a primeira composta por 11 resíduos (MET4 à LEU14) e a segunda composta por 14 resíduos (GLU21 à ASP34). A proteína é estabilizada por duas ligações dissulfídicas [69].

A análise da estereoquímica dos modelos gerados pelas simulações 4 e 7 (Figura 35) revela que o método não se comportou de maneira adequada e que a maioria dos resíduos, na comparação entre os ângulos adotados pela estrutura experimental, ou se posicionam em regiões não permitida ou em regiões diferentes das quais deveria, o que pode ser confirmado na Figura 36.



Figura 36 - Estrutura experimental (em cinza) e estrutura predita (em preto) pela simulação 7, proteína de PDB ID: 1ZDD.

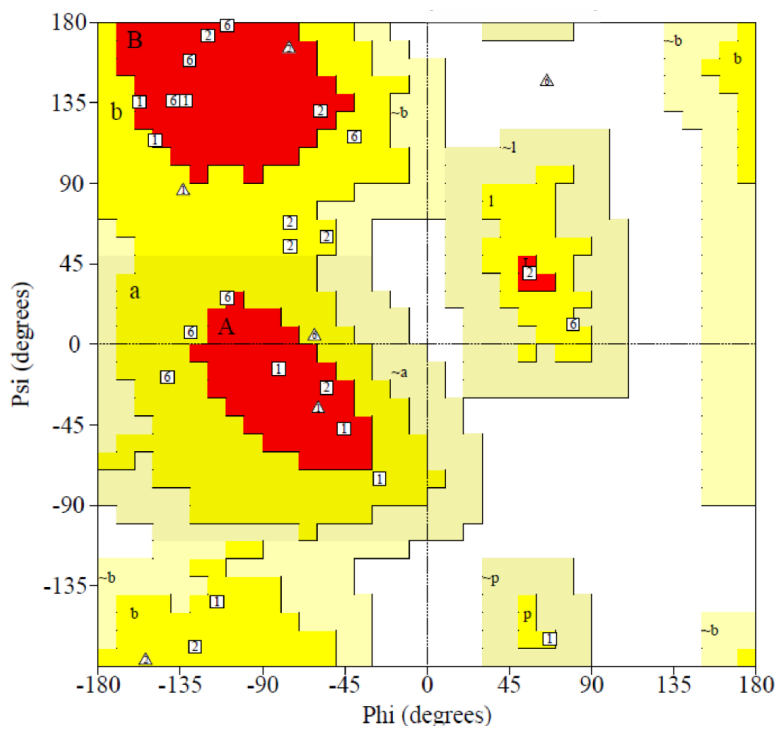


Figura 37 - Diagrama de Ramachandran dos modelos obtidos nas simulações 1, 2 e 6. Proteína PDB ID: 1KVG. Cooperação via estrutura secundária desativada.

Tabela 7 - Resultados obtidos para as diferentes medidas de similaridade em relação à proteína de PDB ID: 1KVG, sem cooperação via estrutura secundária.

Simulação	1	2	3	4	5	6	7	8	9	10	Média ( $\bar{O}$ )	Desvio ( $\sigma$ )
RMSD (Å)	7,50	7,00	6,20	7,50	7,0	3,80	6,50	5,20	6,90	6,4	6,40	1,14
TM-Score	0,19	0,18	0,32	0,15	0,20	0,19	0,14	0,11	0,28	0,19	0,20	0,06
MaxSub	0,45	0,50	0,47	0,40	0,43	0,57	0,48	0,46	0,44	0,41	0,46	0,05
GDT_TS	0,52	0,56	0,56	0,52	0,48	0,65	0,54	0,52	0,52	0,52	0,54	0,04
GDT_HA	0,42	0,44	0,44	0,35	0,40	0,44	0,38	0,31	0,44	0,40	0,40	0,04
Energia	-31,01	-26	-30,25	-30	-27,8	-23,68	-30	-19,24	-27,99	-20,5	-26,65	4,21
Tick	193	182	224	212	176	144	211	171	156	196	186	25

A proteína 1KVG é formada por 13 aminoácidos e adota uma conformação de beta-hairpin quando cristalizada em conjunto ao EPO-R (porção extracelular do hormônio receptor eritropoietina a qual o polipeptídico se liga). A 1KVG possui duas fitas (CIS3 a GLY6 e GLY9 a CIS12) [67].

Por meio da Figura 37, a qual contém o mapa de Ramachandran criado a partir dos modelos 182, 224 e 144 provenientes das simulações 2, 3 e 6, respectivamente, é possível notar que nenhum dos modelos possui resíduos em regiões não permitidas. Partindo-se então para a verificação das conformações adotadas pelos resíduos em relação à estrutura experimental se pôde notar que, enquanto os modelos 182 e 144 possuem resíduos classificados em regiões  $\alpha$  quando deveriam ter sido classificados como  $\beta$ , o mesmo não acontece no modelo 224, o que o caracteriza como o mais adaptado para representar a estrutura nativa da proteína. Além disso, em se tratando das medidas de similaridade, podemos notar pela comparação dos TM-Scores que o modelo 224 proveniente da simulação 3 se mostra diferenciado mesmo antes da análise estereoquímica (embora possua um RMSD maior que o modelo proveniente da simulação 6).

Analisando a conformação 3D da proteína sobreposta da conformação experimental (Figura 38), no entanto, nota-se que a estrutura irregular de volta composta pelos resíduos PRO7 e LEU8 não foi predita, ocasionando uma conformação de baixa qualidade. Com o intuito de se verificar o quão próximo das fitas o método chegou com sua estrutura foram calculados os valores de RMSD somente para os resíduos das fitas, obtendo como resposta os valores de 1,22 Å e 1,82 Å para a primeira e segunda fita, respectivamente.

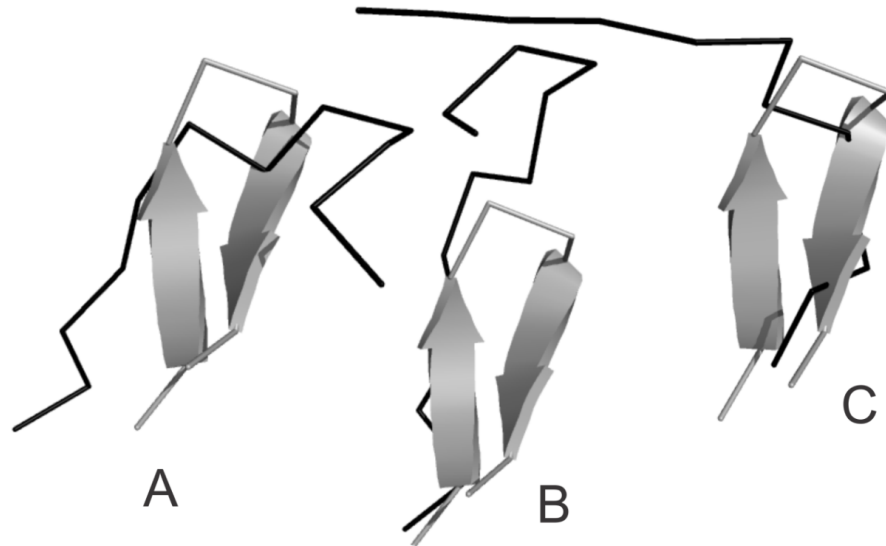


Figura 38 - Estrutura experimental (em cinza) e estrutura predita (em preto) pela simulação 3, proteína de PDB ID: 1KVG.

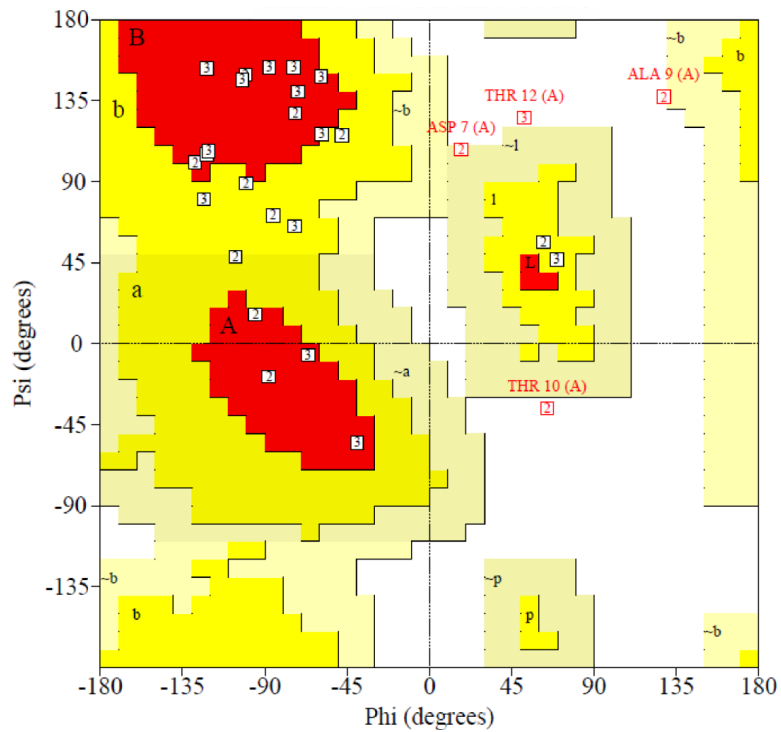


Figura 39 - Diagrama de Ramachandran dos modelos obtidos nas simulações 2 e 3. Proteína PDB ID: 1LE3. Cooperação via estrutura secundária desativada.

Tabela 8 - Resultados obtidos para as diferentes medidas de similaridade em relação à proteína de PDB ID: 1LE3, sem cooperação via estrutura secundária.

Simulação	1	2	3	4	5	6	7	8	9	10	Média ( $\bar{O}$ )	Desvio ( $\sigma$ )
RMSD (Å)	7,20	6,30	4,85	9,90	8,40	5,30	7,70	8,90	6,16	6,40	7,11	1,61
TM-Score	0,18	0,21	0,21	0,20	0,19	0,20	0,18	0,26	0,21	0,14	0,20	0,03
MaxSub	0,43	0,46	0,46	0,36	0,40	0,39	0,44	0,40	0,37	0,34	0,40	0,04
GDT_TS	0,52	0,53	0,61	0,45	0,45	0,53	0,50	0,48	0,52	0,47	0,51	0,05
GDT_HA	0,41	0,39	0,44	0,36	0,36	0,34	0,39	0,38	0,38	0,28	0,37	0,04
Energia	-17,50	-37,08	-23,66	-25,11	-28,22	-2,81	-17,48	-19,65	-22,53	-23,03	-21,71	8,78
Tick	160	174	154	189	157	227	188	157	119	165	169	28

A proteína 1LE3 possui 17 resíduos de aminoácidos e é similar à proteína 1LE0 anteriormente testada. A 1LE3, assim como a 1LE0, também possui um motivo estrutural chamado “zíper de triptofano” o qual estabiliza sua conformação de *beta-hairpin* com duas fitas (GLU3 à ASP7 e THR12 à THR16). Existe ainda uma volta formada por quatro resíduos, indo de ASP8 até a LIS11 [19].

Em um primeiro momento a análise do diagrama de Ramachandran (Figura 39) para as estruturas provenientes das simulações 2 e 3 revela que existem dois resíduos em regiões não permitidas no modelo 174 advento da simulação 2 e apenas uma no modelo 154 advento da simulação 3. Além disso, em se tratando de resíduos nas regiões mais favoráveis, a estrutura proveniente da simulação 2 possui seis resíduos nas regiões principais ou *core* do mapa (regiões em mais escuras), contra nove resíduos em se tratando do modelo 3. Analisando com mais propriedade o mapa de Ramachandran, resíduo a resíduo em comparação ao mapa de Ramachandran da estrutura experimental, nota-se que na estrutura experimental os únicos resíduos não classificados como  $\beta$  são exatamente os quatro resíduos que compõem a estrutura irregular do tipo volta presente na conformação da proteína (estes são classificados como  $\alpha$ ). No modelo encontrado na simulação 2, os dois resíduos dispostos em regiões não permitidas foram a THR10 e a ALA9, ambos os presentes na estrutura irregular de volta da proteína. Quando tratando da simulação 3, o único resíduo disposto em região não permitida no modelo é a THR12, a qual esta presente no início da segunda fita, local também de certa instabilidade. Como não foi possível chegar a uma decisão sobre qual modelo utilizar para representar a potencial estrutura nativa da proteína, passamos à análise dos resíduos que estão

dispostos em regiões marcadas como  $\alpha$ . No modelo 174 gerado na simulação de número 2 os resíduos marcados como hélices são a THR6, a qual se dispõe logo antes do início da estrutura irregular de volta (o que lhe fornece fundamento para ser instável) e a ASP8, a qual faz parte da volta. No modelo 154 gerado na terceira simulação os resíduos marcados como hélices são ASP7 e THR10, ambos parte da estrutura irregular de volta. Assim sendo fica evidente a total conformidade dos resultados obtidos em termos estereoquímicos com a experimental, tanto para a estrutura obtida da simulação 2 quanto para estrutura obtida da simulação 3.

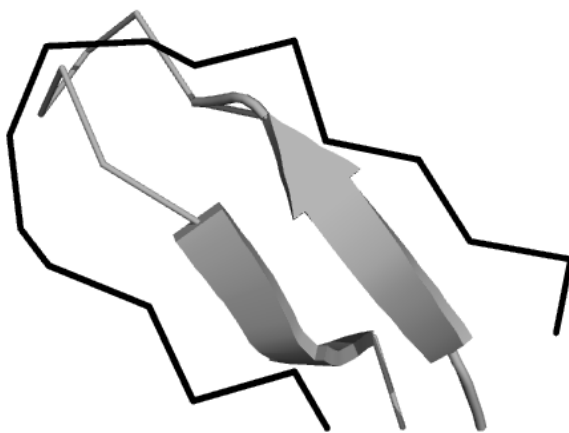


Figura 40 - Estrutura experimental (em cinza) e estrutura predita (em preto) pela simulação 3, proteína de PDB ID: 1LE3. Cooperação via estrutura secundária desativada.

Tabela 9 - Resultados obtidos para as diferentes medidas de similaridade em relação à proteína de PDB ID: 1VII, sem cooperação via estrutura secundária.

Simulação	1	2	3	4	5	6	7	8	9	10	Média ( $\bar{O}$ )	Desvio ( $\sigma$ )
RMSD (Å)	9,3	9	10,5	10,7	6,5	10,4	8,9	9,6	9,5	8,9	9,33	1,20
TM-Score	0,15	0,18	0,15	0,12	0,20	0,14	0,14	0,16	0,14	0,14	0,15	0,02
MaxSub	0,18	0,24	0,18	0,17	0,29	0,18	0,17	0,20	0,17	0,19	0,20	0,04
GDT_TS	0,31	0,32	0,28	0,26	0,40	0,29	0,31	0,31	0,26	0,30	0,30	0,04
GDT_HA	0,16	0,19	0,17	0,15	0,24	0,17	0,19	0,19	0,15	0,18	0,18	0,02
Energia	-73,33	-50,19	-79,51	-40	-55,97	-69,06	-59,32	-48,83	-31,5	-42	-54,97	15,50
Tick	192	212	177	180	169	219	205	197	160	176	188	19

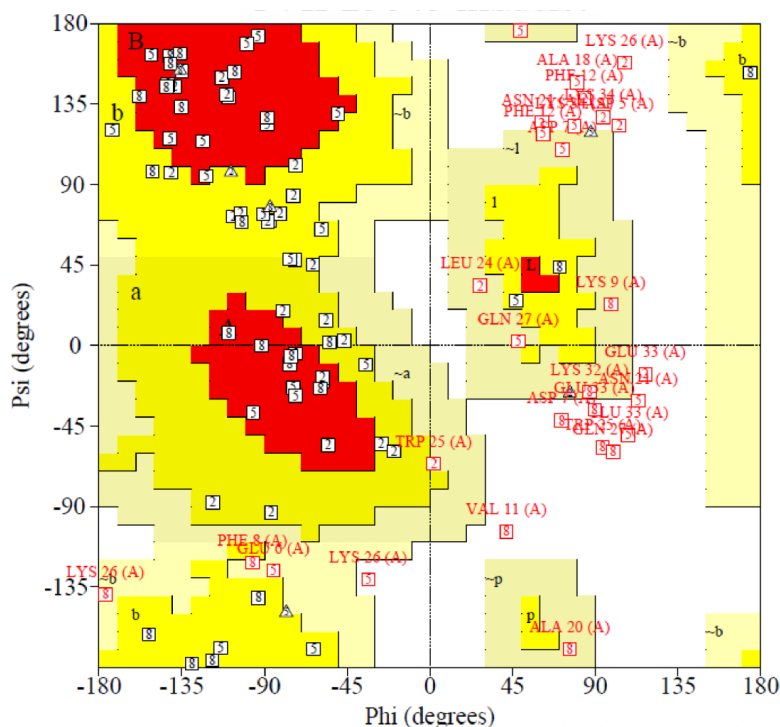


Figura 41 – Diagrama de Ramachandran dos modelos obtidos nas simulações 2, 5 e 8. Proteína PDB ID: 1VII. Cooperação via estrutura secundária desativada.

A proteína 1VII possui 36 resíduos de aminoácidos e se constitui de um feixe de três hélices, a primeira formada de ASP5 até LYS9, a segunda composta da ARG16 até a PHE19 e a terceira composta da LEU24 até GLU33 [55]. Por meio do mapa de Ramachandran disposto na Figura 41 é possível notar que os três modelos escolhidos para análise estereoquímica apresentam entre cinco e seis resíduos em regiões não permitidas. Verificando quais resíduos são esses, percebe-se que alguns deles se repetem (ASN21, LYS26, GLU33 e LYS34) e que a estrutura que possui mais resíduos pertencentes à estruturas secundárias em tais regiões é o modelo proveniente da simulação 8.

Análise resíduo a resíduo dos ângulos phi e psi dispostos no mapa de Ramachandran gerado nas simulações em comparação ao valor dos ângulos adotado pela conformação experimental revela que nenhuma das três conformações tem os resíduos pertencentes à primeira hélice em posições adequadas. Entretanto, tanto o modelo 169 quanto o modelo 197 demonstraram adotar posições no mapa em conformidade com a estrutura experimental quando analisada a terceira hélice. Dentre os modelos 169 e 197 o modelo mais próximo das

hélices em se tratando de estrutura 3D foi o modelo 169, originário da simulação 5 (Figura 42).

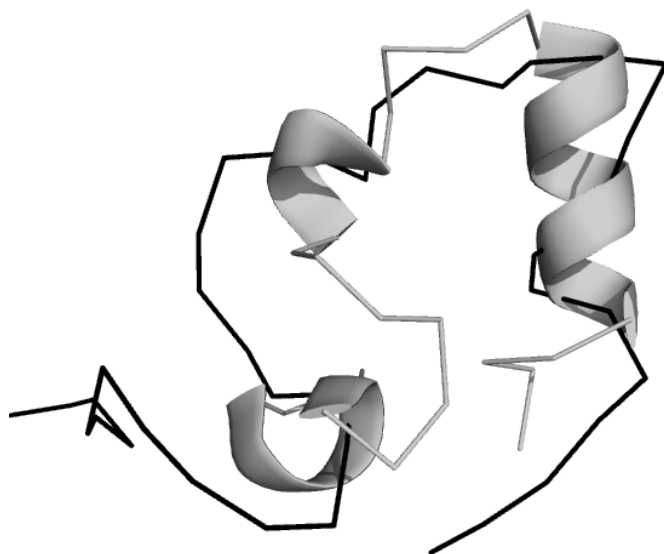


Figura 42 - Estrutura experimental (em cinza) e estrutura predita (em preto) pela simulação 5, proteína de PDB ID: 1VII. Cooperação via estrutura secundária desativada. O RMSD levando-se em conta uma sobreposição somente os 10 resíduos que formam a terceira hélice tem o valor de 3,70 Å.

Tabela 10 - Resultados obtidos para as diferentes medidas de similaridade em relação à proteína de PDB ID: 2GP8, sem cooperação via estrutura secundária.

Simulação	1	2	3	4	5	6	7	8	9	10	Média ( $\bar{O}$ )	Desvio ( $\sigma$ )
RMSD (Å)	9,10	12,09	12,83	11,91	12,34	13,40	12,73	11,76	9,52	13,26	13,12	1,44
TM-Score	0,17	0,16	0,15	0,14	0,13	0,14	0,14	0,13	0,12	0,15	0,14	0,01
MaxSub	0,19	0,18	0,18	0,16	0,14	0,15	0,15	0,15	0,13	0,19	0,15	0,02
GDT_TS	0,30	0,30	0,25	0,27	0,25	0,25	0,31	0,21	0,24	0,28	0,25	0,03
GDT_HA	0,18	0,18	0,15	0,16	0,15	0,14	0,17	0,14	0,13	0,16	0,16	0,02
Energia	-58,73	-73,90	-67,02	-56,08	-31,36	-58,3	-44,58	-32,92	-56,70	-65,53	-54,05	14,89
Tick	205	211	208	211	208	207	202	178	205	217	205	10





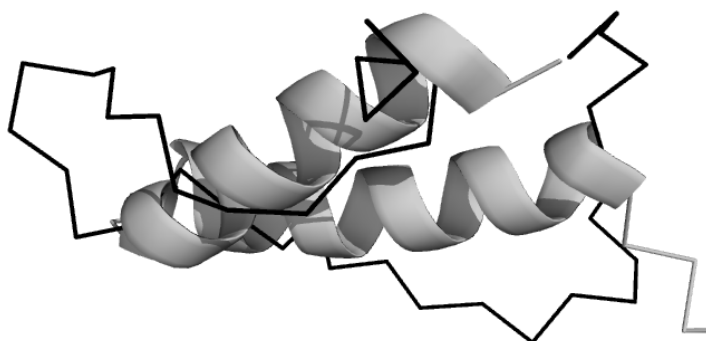


Figura 44 - Estrutura experimental (em cinza) e estrutura predita (em preto) pela simulação 1, proteína de PDB ID: 2GP8. Cooperação via estrutura secundária desativada.

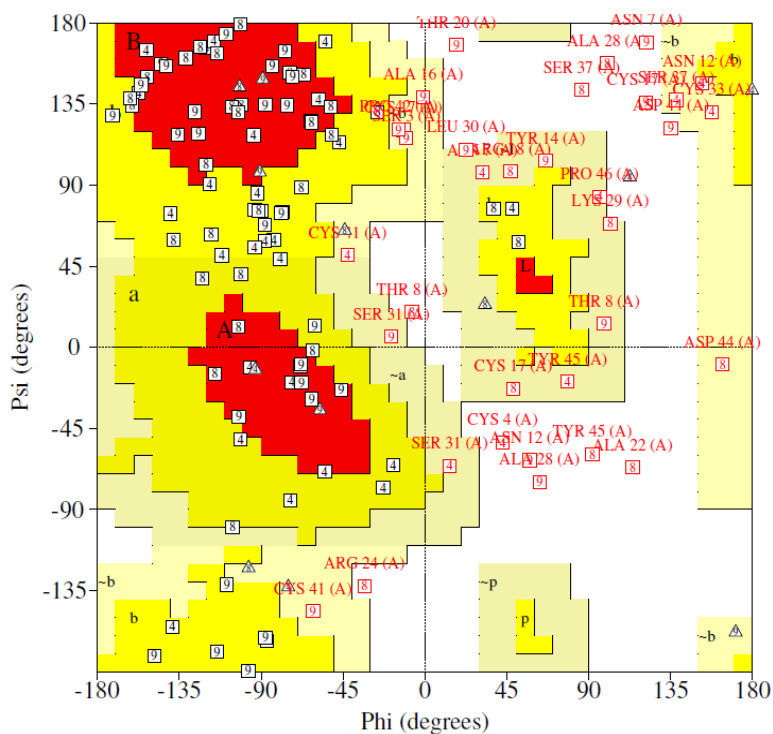


Figura 45 - Diagrama de Ramachandran dos modelos obtidos nas simulações 4, 8 e 9. Proteína PDB ID: 1ED0. Cooperação via estrutura secundária desativada. Triângulos simbolizam glicinas.

Tabela 11 - Resultados obtidos para as diferentes medidas de similaridade em relação à proteína de PDB ID: 1ED0, sem cooperação via estrutura secundária.

Simulação	1	2	3	4	5	6	7	8	9	10	Média ( $\bar{O}$ )	Desvio ( $\sigma$ )
RMSD (Å)	11,50	13,2	11,9	13,7	15,6	12,2	12,5	10,5	11,8	11,7	12,57	1,38
TM-Score	0,14	0,15	0,16	0,14	0,12	0,12	0,15	0,17	0,14	0,13	0,14	0,02
MaxSub	0,16	0,15	0,18	0,15	0,13	0,14	0,16	0,20	0,16	0,16	0,16	0,02
GDT_TS	0,23	0,23	0,28	0,24	0,22	0,21	0,21	0,27	0,24	0,23	0,24	0,02
GDT_HA	0,14	0,15	0,16	0,14	0,12	0,13	0,15	0,15	0,16	0,14	0,14	0,01
Energia	14,81	-44,7	-24,73	-38,24	-57,24	-89,31	-40,2	-18,3	-43,58	4,91	-33,66	28,44
Tick	122	228	181	150	185	199	192	166	206	127	176	33

A proteína 1ED0 possui 46 resíduos e uma topologia beta-alfa-alfa-beta, contando com uma folha  $\beta$  composta pelas fitas (LIS2 à CYS4 e LYS34 à ILE36) e duas hélices  $\alpha$  (THR8 à LEU19 e ARG24 à SER31) [64].

Por meio das 10 execuções feitas simulando a proteína os modelos 150, 166 e 206 provenientes das simulações 4, 8 e 9, respectivamente, foram os que obtiveram os melhores resultados no que se trata de avaliação de similaridade.

Análise do mapa de Ramachandran (Figura 45) para tais modelos revela que os modelos 166 e 206 possuem um considerável número de resíduos em regiões não permitidas (cerca de 17,10%), enquanto o modelo 150 proveniente da simulação de número 4 possui apenas um resíduo (2,90%) em tal região. Ao analisarmos resíduo a resíduo confirmamos que o modelo 150, apesar de possuir o pior valor em termos de RMSD, é aquele que mais acerta no que se trata de estereoquímica em comparação à estrutura experimental. Ao verificarmos a sobreposição com a estrutura experimental (Figura 46) nos deparamos com um resultado não promissor, entretanto, ao alinharmos apenas as estruturas de hélice (relacionadas a 20 resíduos ou 46% da proteína), verificamos um RMSD relativamente mais baixo, de 4,47 e 3,39 para a primeira e segunda hélices, respectivamente.

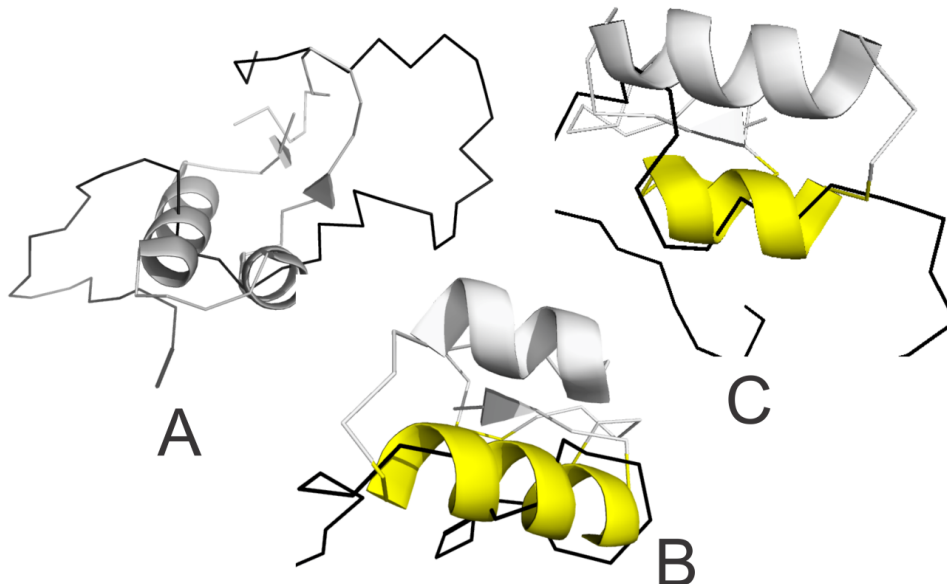


Figura 46 - Estruturas experimental (em cinza e amarelo) e estruturas previstas (em preto) pela simulação 4, proteína de PDB ID: 1ED0. Cooperação via estrutura secundária desativada. Em A temos a melhor sobreposição entre a estrutura experimental e a prevista. Em B a sobreposição apenas dos resíduos que compõem a primeira hélice e, em C, a sobreposição apenas dos resíduos que compõem a segunda hélice.

### 6.2.3 Comparação

Bortolussi e colaboradores utilizaram apenas duas métricas para avaliar a qualidade de seus modelos previstos: RMSD e energia. Embora tenhamos em mente não ser essa a abordagem ideal (pelas deficiências que o RMSD possui quando utilizado na avaliação da similaridade de proteínas) se optou pela comparação dos resultados obtidos pelos dois métodos em termos de RMSD e energia, buscando evidenciar os casos em que cada abordagem mostrou-se melhor ou pior adaptados. Essa comparação deu origem a Tabela 12 e Tabela 13 onde resultados tidos como “melhores” segundo a teoria de Anfinsen estão em negrito.

Por meio da Tabela 12 é possível notar que o PROSMART obteve melhores resultados em se tratando das proteínas 1LE3, 1EDP, 1PG1 e 1ZDD, entretanto alcançando resultados piores para as proteínas 1LE0, 1KVG, 1VII, 2GP8 e 1ED0. Quanto à comparação das energias encontradas e presentes na Tabela 13, fica evidente que a estratégia de busca implementada nesta dissertação se mostrou mais efetiva que a de Bortolussi *et al*, alcançando melhores resultados em todas as simulações. Notou-se também que os desvios emergentes das simulações no PRO-SMART foram significativamente maiores, o que pode ser endereçado

ao método de clusterização utilizado para a escolha dos modelos resposta de cada simulação. Além disso, tendo em mente que uma função de energia possui como característica estar imersa em um vasto cenário de conformações, é factível que, partindo-se de posições diferentes do cenário (o que acontece em nossa simulação devido à semente aleatória gerada antes de cada simulação), a função de energia percorra diferentes caminhos.

Tabela 12 - Valores (em Å) Comparação em termos de RMSD entre o PROSMART e o *framework* de Bortolussi *et al.* Cooperação via estrutura secundária desativada.

PDB ID	PRO-SMART	Bortolussi <i>et al.</i>
1LE0	6,37 ±1,45)	<b>5.41 ±0,70</b>
1KVG	6.40 ±1,14)	<b>5.06 ±0,93</b>
1LE3	<b>7.11 ±1,61</b>	7,21 ±0,92
1EDP	<b>5,95 ±1,39</b>	6,17 ±0,79
1PG1	<b>6,99 ±1,50</b>	9,12 ±1,33
1ZDD	<b>8,44 ±0,59</b>	8,51 ±1,29
1VII	9,33 ±1,20	<b>8,40 ±1,08</b>
2GP8	13,12 ±1,44	<b>7,96 ±1,50</b>
1ED0	12,57 ±1,38	<b>10,80 ±0,87</b>

Tabela 13 - Comparação em termos de energia entre o PROSMART e o *framework* de Bortolussi *et al.* Cooperação via estrutura secundária desativada.

PDB ID	PRO-SMART	Bortolussi <i>et al.</i>
1LE0	<b>-13,76 ±5,07</b>	2,88 ±1,43
1KVG	<b>-26,65 ±4,21</b>	-1,33 ±1,18
1LE3	<b>-21,71 ±8,78</b>	1,82 ±1,77
1EDP	<b>-28,84 ±11,01</b>	-3,55 ±0,77
1PG1	<b>-35,60 ±12,07</b>	23,79 ±1,37
1ZDD	<b>-42,95 ±18,68</b>	14,20 ±3,12
1VII	<b>-54,97 ±15,50</b>	11,33 ±3,88
2GP8	<b>-54,05 ±14,89</b>	48,63 ±4,52
1ED0	<b>-33,66 ±28,44</b>	16,72 ±2,19

### 6.3 Simulações com Cooperação via Estruturas Secundária

#### 6.3.1 Configuração

A configuração do PRO-SMART utilizada nas simulações com cooperação via estrutura secundária difere, em relação às simulações sem cooperação, em apenas três variáveis: (i) a variável chamada “SSP”, a qual é tem o valor “ON” atribuído, (ii) a variável “SSP\_TYPE” a qual não tinha valor atribuído e se torna responsável por escolher o peso

estipulado para a cooperação via estrutura secundária e (iii) o campo “SSP-SEQ”, o qual deve ser preenchido com a predição de estrutura secundária proveniente de um software auxiliar.

Tabela 14 - Conjunto de atributos e funcionalidades configuradas nas simulações com cooperação via estruturas secundárias ativada com peso=1.

Variável/Funcionalidade	Valor
temp_thr	0,10
attempted_threshold_without_orchestra	32
attempted_threshold_with_orchestra	13
temp_zeromoves	50
total_orch_moves	20
orchestra_sleep_time	1
min_crank	2 AAs
max_crank	5 AAs
min_pivot_dist	2 AAs
max_angle	90°
moves	Cubo
move-C-Alphas	Yes
move-C-Betas	Yes
fator_temp_buscadores	1
fator_temp_orch	1
max_number_of_no_improvement_in_energy	6 Ticks
min_temperature_allowed	0,099
temp_decrease_ratio	0,98
energy_variation_threshold	$1 \times 10^{-4}$
temp_orch_thr	1
initial_temperature	10
SSP	ON
SSP_TYPE	1 ou 1/10

A Tabela 15 conta com os valores do campo SSP-SEQ utilizado para cada proteína. Aqui é importante lembrar a importante diferença entre a abordagem de Bortolussi *et al.* quanto a estruturas secundárias e a utilizada pelo PRO-SMART. Enquanto Bortolussi *et al.* utiliza informações sobre a conformação exata (retirada do PDB) dos resíduos que formam estruturas secundárias, o PRO-SMART se utiliza apenas de uma informação, a sequencia “SSP-Seq”. Além disso, a versão do PRO-SMART aqui disposta possui ainda uma considerável limitação, a de possuir heurísticas para tratar de hélices. Pesquisas relacionadas ao padrão conformacional de fitas que formam folhas  $\beta$  vem sendo objeto de estudo nos últimos anos e pretende-se que heurísticas para tal finalidade sejam incorporadas ao sistema em breve. Assim sendo, optou-se por simular com a cooperação via estrutura secundária

ativada, apenas proteínas que não possuísem fitas, diminuindo o número de proteínas simuladas em relação aos testes sem cooperação.

Tabela 15 - Sequências referentes à estrutura secundária das proteínas simuladas. Única informação utilizada pelo PRO-SMART. A letra C simboliza resíduos que, segundo o preditor, não formam estruturas secundárias (estes resíduos são chamados *coils*), a letra H simboliza resíduos que, segundo o preditor, formam hélices.

PDB ID	SSP-Seq
1EDP	"CCCCCCCCCCHHHHCCCC"
1ZDD	"CCCHHHHHHHHHHHHHCCCCCHHHHHHHHHHHHHHHHCC"
1VII	"CCCCHHHHHHHHHCCCHHHHHCCCHHHHHHHHHHHHHCCCC"
2GP8	"CCCCCCHHHHHHHHHHHHHHHHHHHCCCCCHHHHHHHHHHHHHCCCC"

### 6.3.2 Resultados Obtidos

Nas Tabelas 16 a 23 constam os resultados obtidos nas simulações utilizando-se o mesmo peso=1 utilizado por Bortolussi *et al.* em seu trabalho e, nas Tabelas 24 a 27 constam os resultados obtidos nas simulações utilizando-se uma nova heurística, onde o peso atribuído para a cooperação via estrutura secundária ou em outras palavras, a punição em termos de energia é baseada no potencial atual do sistema, tendo o valor de 1 décimo da energia.

Novamente tendo como alvo de simulação a proteína 1EDP de estrutura topológica já descrita na seção anterior, com o objetivo de melhorar a predição obtida por meio da adição de cooperação via estrutura secundárias, foram executadas 10 simulações com cooperação ativada e peso=1, onde os melhores resultados em termos de geometria e medidas de similaridade foram obtidos nas simulações 2, 6, 9 e 10. O diagrama de Ramachandran (Figura 47) para as quatro simulações em destaque revela que as simulações 6 e 9 possuem resíduos adotando regiões não permitidas tanto no interior quanto no exterior da hélice da proteína, o que prejudica a obtenção de uma maior similaridade com a estrutura experimental, mesmo sendo o modelo proveniente da simulação 6 o de menor valor de RMSD entre os destacados. Assim sendo fica evidente o motivo do modelo 6, apesar de possuir um RMSD melhor, não possuir melhores valores em termos de MaxSub, TM-Score e GDT.

Tabela 16 - Simulações com cooperação via estrutura secundária. Proteína de PDB ID: 1EDP. Peso = 1.

Simulação	1	2	3	4	5	6	7	8	9	10	Média ( $\bar{O}$ )	Desvio ( $\sigma$ )
RMSD (Å)	5,60	5,70	5,70	6,00	6,90	4,80	5,20	6,80	7,10	6,10	5,99	0,75
TM-Score	0,13	0,15	0,12	0,16	0,18	0,16	0,16	0,09	0,16	0,15	0,15	0,03
MaxSub	0,31	0,40	0,35	0,34	0,31	0,40	0,35	0,34	0,41	0,40	0,36	0,04
GDT_TS	0,51	0,54	0,50	0,49	0,43	0,56	0,51	0,47	0,50	0,54	0,51	0,04
GDT_HA	0,31	0,40	0,31	0,31	0,31	0,37	0,31	0,28	0,34	0,38	0,33	0,04
Energia	-12,01	-17,73	-22,04	-19,00	-20,05	-1,11	-29,60	-19,79	-11,00	-12,00	-16,43	7,75
Tick	216	189	197	181	201	185	198	116	191	223	189	29

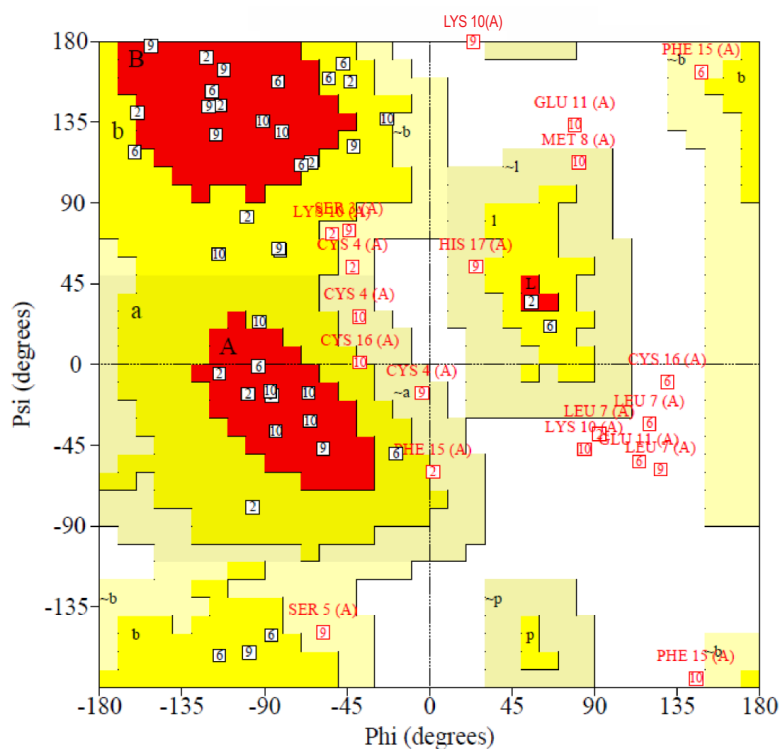


Figura 47 - Diagrama de Ramachandran dos modelos obtidos nas simulações 2, 6, 9 e 10. Proteína PDB ID: 1EDP. Cooperação via estruturas secundárias ativada, com peso = 1.

Ao analisarmos os modelos originários das simulações 2 e 10 é possível verificar que esses possuem apenas um resíduo em região não permitida e em ambos os casos o resíduo é a LEU7, a qual faz parte da estrutura irregular do tipo volta existente na conformação da proteína. A visualização da estrutura 3D (Figura 48) revela conformações com dobramento



mais próximo ao da experimental em comparação aos modelos 6 e 9, porém possuindo desconformidades justamente em posições posteriores às posições de suas LEU7. A análise do mapa de Ramachandran (Figura 47) para cada resíduo de aminoácido dando ênfase as regiões do mapa em que se enquadram os resíduos que devem formar a ER  $\alpha$  da proteína (LYS10 até CYS16) revela que o único modelo que possui um número relevante de resíduos em tal região é o modelo 185 proveniente da simulação 6, onde 5 resíduos pertencem à região  $\alpha$  do mapa e apenas 2 dos resíduos que deveriam adotar tal região (LYS10 e GLU11) não o fazem (e estão no início da hélice). Segundo nossa análise estereoquímica o modelo melhor adaptado para representar a proteína, dentre os encontrados nas simulações, é o modelo proveniente da simulação 6, o que condiz com os valores encontrados em nossa análise geométrica na qual o modelo obteve na, grande maioria dos casos, valores mais atraentes.

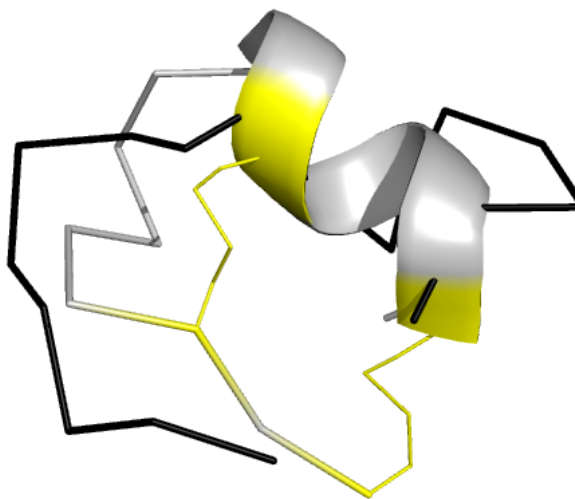


Figura 48 - Estrutura experimental (em cinza) e estrutura predita (em preto) pela simulação 6, proteína de PDB ID: 1EDP. Cooperação via estrutura secundária ativada com peso = 1. As ligações dissulfídicas estão destacadas em amarelo.

Tabela 17 - Simulações com cooperação via estrutura secundária. Proteína de PDB ID: 1VII. Peso = 1.

Simulação	1	2	3	4	5	6	7	8	9	10	Média ( $\bar{\theta}$ )	Desvio ( $\sigma$ )
RMSD (Å)	18,80	12,00	12,80	10,40	10,60	12,50	9,50	13,60	8,60	15,80	12,46	3,06
TM-Score	0,14	0,16	0,14	0,17	0,14	0,13	0,14	0,15	0,14	0,14	0,14	0,01
MaxSub	0,16	0,18	0,22	0,24	0,19	0,18	0,22	0,20	0,18	0,19	0,20	0,02
GDT_TS	0,26	0,26	0,28	0,31	0,28	0,26	0,34	0,27	0,28	0,27	0,28	0,03
GDT_HA	0,17	0,17	0,17	0,20	0,17	0,15	0,20	0,17	0,17	0,17	0,18	0,02
Energia	-54,59	-11,87	-49,56	-52,64	-53,69	-41,44	-72,55	-51,5	-14,41	-36,28	-43,85	18,73
Tick	209	182	202	191	215	201	188	188	210	202	198	11

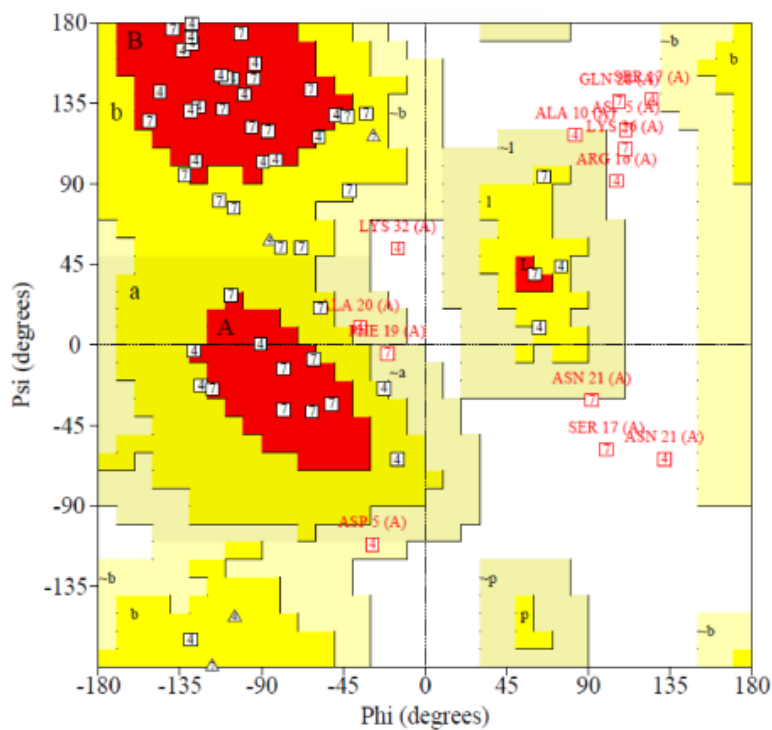


Figura 49 - Diagrama de Ramachandran dos modelos obtidos nas simulações 4 e 10. Proteína PDB ID: 1VII. Cooperação via estruturas secundárias ativada, com peso = 1.

Segundo o mapa de Ramachandran (Figura 49) o modelo 191, proveniente da simulação 4 possui quatro resíduos em região não permitida enquanto o modelo 188, proveniente da simulação 7, possui cinco resíduos em regiões não permitidas. Verificando resíduo a resíduo no mapa verifica-se que nenhum dos modelos possui grande numero de resíduos a adotar as regiões esperadas no mapa levando-se em conta a estrutura experimental, o que nos dá como resultado uma estrutura 3D de baixa qualidade.

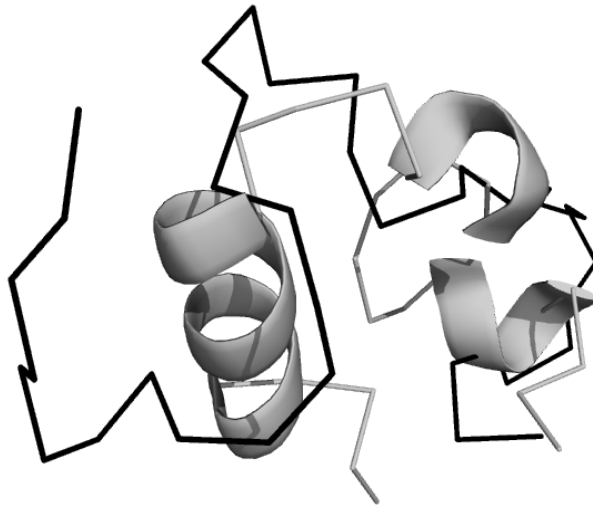


Figura 50 - Estrutura experimental (em cinza) e estrutura predita (em preto) pela simulação 6, proteína de PDB ID: 1VII. Cooperação via estrutura secundária ativada com peso = 1.

Tabela 18 - Simulações com cooperação via estrutura secundária. Proteína de PDB ID: 1ZDD. Peso = 1.

Simulação	1	2	3	4	5	6	7	8	9	10	Média ( $\bar{\theta}$ )	Desvio ( $\sigma$ )
RMSD (Å)	6,40	6,40	8,00	8,70	11,40	10,07	13,70	13,10	6,30	10,87	9,49	2,76
TM-Score	0,16	0,13	0,12	0,14	0,15	0,17	0,12	0,13	0,16	0,14	0,14	0,02
MaxSub	0,26	0,19	0,20	0,21	0,21	0,23	0,18	0,21	0,20	0,22	0,21	0,02
GDT_TS	0,38	0,36	0,29	0,32	0,29	0,32	0,27	0,32	0,38	0,32	0,33	0,04
GDT_HA	0,19	0,17	0,18	0,18	0,18	0,20	0,16	0,19	0,20	0,18	0,18	0,01
Energia	-18,00	-30,49	-12,00	-19,06	-37,87	-12,60	21,42	-32,75	-26,85	-29,36	-19,76	16,88
Tick	162	186	183	210	165	198	188	211	159	203	186	19

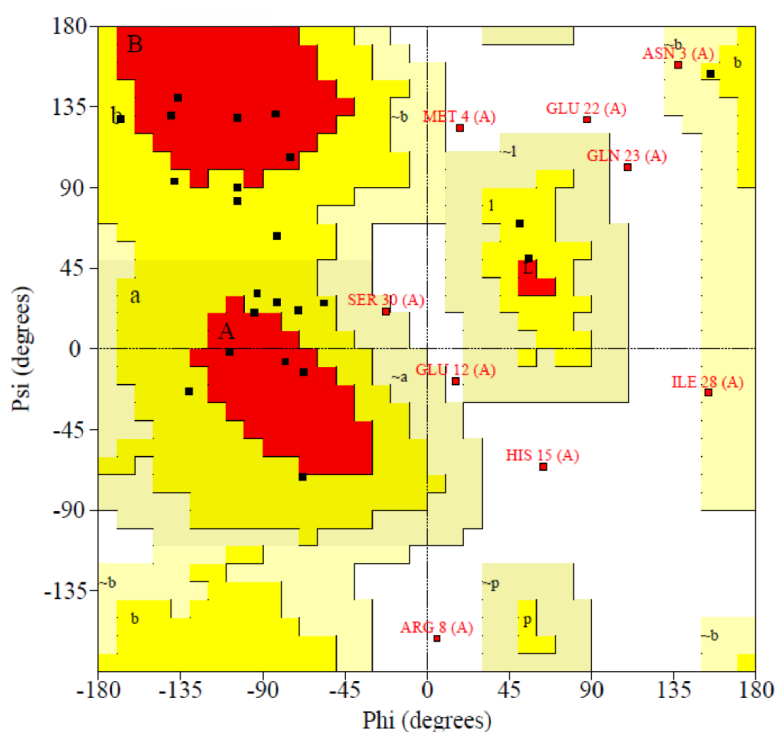


Figura 51 - Diagrama de Ramachandran dos modelos obtidos na simulação 1. Proteína PDB ID: 1ZDD. Cooperação via estruturas secundárias ativada, com peso = 1.

Novamente tendo como alvo de 10 simulações a proteína 1ZDD (de estrutura topológica já descrita na seção anterior), porém desta vez utilizando cooperação via estruturas secundárias com peso=1 seguido da análise geométrica dos resultados encontrados chegou-se ao mapa de Ramachandran disposto na Figura 51, referente ao modelo 162 proveniente da simulação 1. A análise do mapa de Ramachandran revela que os resíduos parte da primeira hélice, em sua maioria, se colocaram nas devidas posições do mapa em relação a estrutura experimental. Entretanto, a influência dos métodos utilizados para completar o esqueleto e a cadeia lateral dos aminoácidos que compõem a estrutura predita se mostrou não efetiva pelo fato de ter gerado conformações com tamanhos de ligações impróprios. Além disso, analisando os ângulos phi e psi adotados pelos resíduos que compõem a segunda hélice verificamos que estes em sua maioria não adotaram as posições desejadas.



Figura 52 - Estrutura experimental (em cinza) e estrutura predita (em preto) pela simulação 1, proteína de PDB ID: 1ZDD. Cooperação via estrutura secundária ativada com peso = 1.

Tabela 19 - Simulações com cooperação via estrutura secundária. Proteína de PDB ID: 2GP8. Peso = 1.

Execução	1	2	3	4	5	6	7	8	9	10	Média ( $\bar{O}$ )	Desvio ( $\sigma$ )
RMSD (Å)	9,10	12,93	12,50	18,70	11,76	9,40	19,36	9,70	16,10	9,00	12,86	3,93
TM-Score	0,17	0,14	0,13	0,12	0,13	0,15	0,13	0,15	0,14	0,17	0,14	0,02
MaxSub	0,19	0,18	0,16	0,15	0,17	0,18	0,17	0,18	0,17	0,21	0,18	0,02
GDT_TS	0,29	0,27	0,25	0,25	0,26	0,26	0,25	0,27	0,28	0,30	0,27	0,02
GDT_HA	0,18	0,16	0,15	0,15	0,16	0,16	0,16	0,16	0,18	0,18	0,16	0,01
Energia	-28,75	-26,16	-34,00	-20,00	-42,80	-26,24	-31,58	1,70	-18,10	-52,23	-27,82	14,55
Tick	186	217	173	212	174	193	212	150	160	214	189	24

Novamente tendo como alvo a proteína 2GP8 de estrutura já descrita anteriormente, com a adição da cooperação via estruturas secundárias com peso equivalente a 1, os modelos originários das simulações 1, 8 e 10 não demonstraram significativa melhora, alcançando resultados do mesmo calão em termos de RMSD, entretanto, no que se trata de estereoquímica, foi possível verificar melhorias tanto no que se trata do número de resíduos em regiões não permitidas quanto em termos da disposição dos resíduos pertencentes às estruturas secundárias. A seguir está disposta a estrutura 3D do modelo 186, originário da simulação 1, o qual possui mais resíduos dispostos na região de hélices  $\alpha$ , segundo Figura 53.

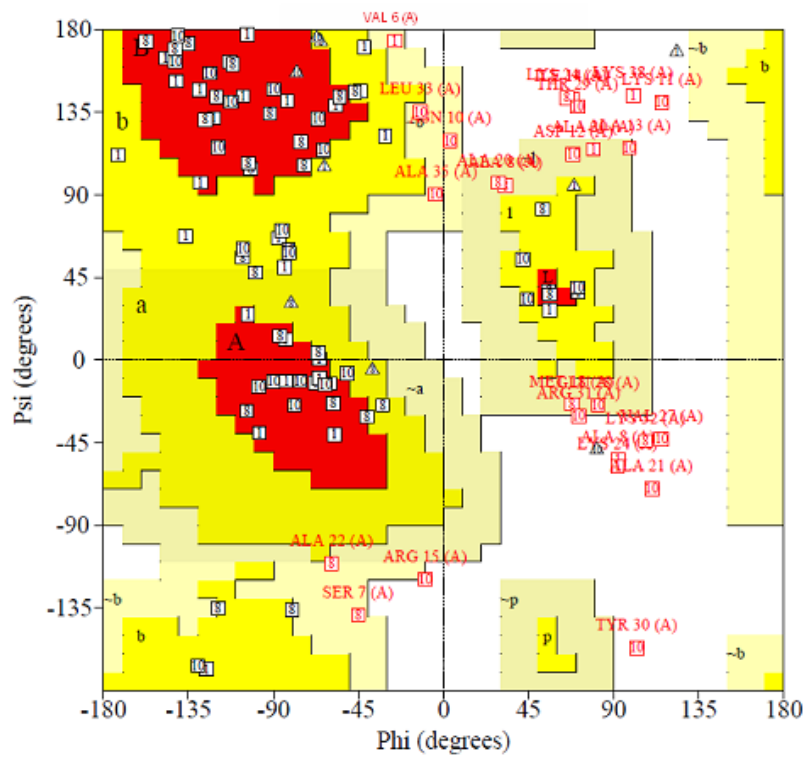


Figura 53 - Diagrama de Ramachandran dos modelos obtidos nas simulações 1, 8 e 10. Proteína PDB ID: 2GP8. Cooperação via estruturas secundárias ativada, com peso = 1.

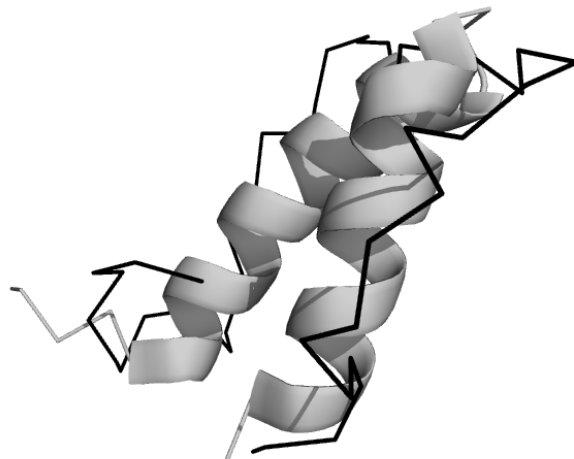


Figura 54 - Estrutura experimental (em cinza) e estrutura predita (em preto) pela simulação 1, proteína de PDB ID: 2GP8. Cooperação via estrutura secundária ativada com peso = 1.

Tabela 20 - Simulações com cooperação via estrutura secundária. Proteína de PDB ID: 1EDP. Peso = energia/10.

Simulação	1	2	3	4	5	6	7	8	9	10	Média ( $\bar{\theta}$ )	Desvio ( $\sigma$ )
RMSD (Å)	5,20	5,80	4,70	6,70	6,70	7,80	4,00	6,90	4,90	5,00	5,77	1,26
TM-Score	0,13	0,13	0,21	0,14	0,11	0,15	0,09	0,08	0,21	0,22	0,15	0,05
MaxSub	0,37	0,51	0,37	0,36	0,34	0,31	0,46	0,37	0,36	0,41	0,39	0,06
GDT_TS	0,50	0,57	0,53	0,47	0,50	0,43	0,54	0,49	0,51	0,53	0,51	0,04
GDT_HA	0,32	0,38	0,35	0,29	0,29	0,32	0,32	0,29	0,32	0,35	0,33	0,03
Energia	-11,43	-5,66	-19,17	-21,51	-12,22	-8,16	0,00	-0,25	6,65	-22,40	-9,42	9,82
Tick	197	161	191	175	143	186	155	158	153	201	172	20

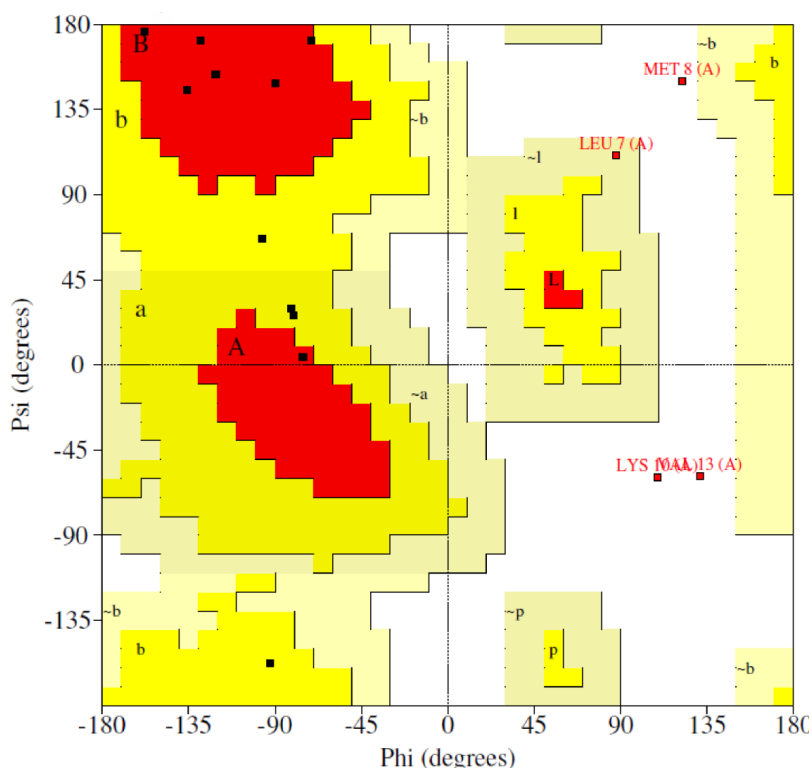


Figura 55 - Diagrama de Ramachandran do modelo obtido na simulação 10. Proteína PDB ID: 1EDP. Cooperação via estruturas secundárias ativada, com peso = energia/10.



Novamente tratando de simulações envolvendo a proteína 1EDP, agora com a cooperação via estruturas secundária ativa, com peso equivalente à um décimo da energia da proteína. A estrutura escolhida para representar a estrutura nativa da proteína neste caso foi o modelo proveniente da simulação 10 e, como pode ser notado, este possui um valor inferior em termos de RMSD em relação à estrutura gerada na simulação 7. Ao analisarmos os outros parâmetros de similaridade podemos notar que a estrutura proveniente da simulação 7, embora tenha RMSD menor (melhor), possui um valor de TM-Score muito abaixo da média, o que nos leva a crer que, teoricamente, seu enovelamento não deve ser bom. Isso foi confirmado através da visualização 3D da estrutura que confirma o fato de má formação conformacional tanto na parte da proteína onde não esta presente a ER como também na parte onde existe ER.

No modelo gerado a partir da simulação 10, segundo o mapa de Ramachandran, existem três resíduos em regiões não permitidas, sendo que dois deles fazem parte da parte da estrutura onde uma estrutura regular do tipo hélice é formada, assim dando indícios de uma conformação que embora tenha um RMSD aceitável, não deve condizer com a estrutura experimental.

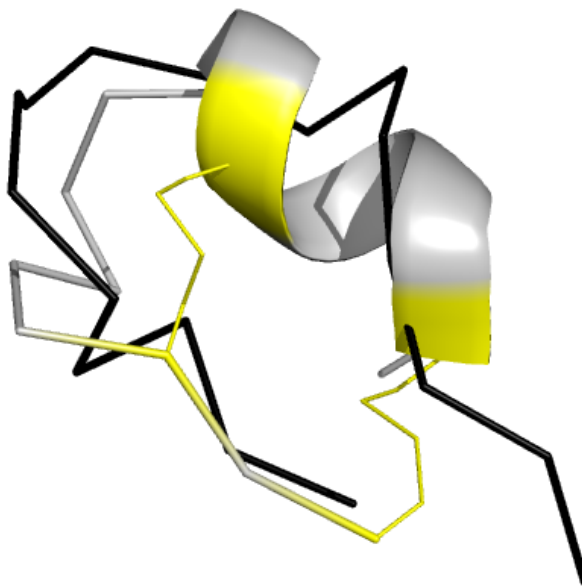


Figura 56 - Estrutura experimental (em cinza) e estrutura predita (em preto) pela simulação 1, proteína de PDB ID: 1EDP. Cooperação via estrutura secundária ativada com peso = energia/10.. Pontes bissulfídricas em amarelo.

Tabela 21 - Simulações com cooperação via estrutura secundária. Proteína de PDB ID: 1ZDD. Peso = energia/10.

Simulação	1	2	3	4	5	6	7	8	9	10	Média ( $\bar{\theta}$ )	Desvio ( $\sigma$ )
RMSD (Å)	6,50	8,20	6,30	10,00	9,10	9,90	9,90	8,30	8,00	11,50	8,77	1,63
TM-Score	0,14	0,13	0,18	0,13	0,15	0,22	0,16	0,14	0,14	0,14	0,15	0,03
MaxSub	0,20	0,20	0,28	0,15	0,20	0,32	0,22	0,20	0,22	0,19	0,22	0,05
GDT_TS	0,36	0,30	0,38	0,25	0,28	0,38	0,32	0,30	0,36	0,34	0,33	0,04
GDT_HA	0,18	0,18	0,21	0,15	0,17	0,24	0,20	0,18	0,19	0,18	0,19	0,02
Energia	19,10	0,00	17,20	0,04	16,34	11,52	-0,39	0,02	39,69	0,00	10,32	13,16
Tick	226	171	207	168	114	114	166	206	128	170	167	39

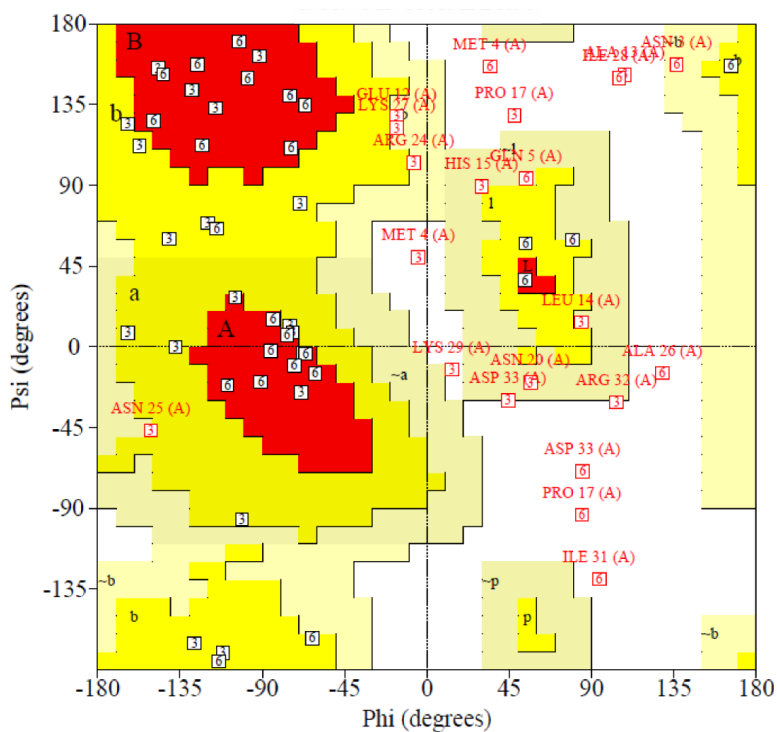


Figura 57 - Diagrama de Ramachandran dos modelos obtidos nas simulações 3 e 6. Proteína PDB ID: 1ZDD. Cooperação via estruturas secundárias ativada, com peso = energia/10.

Nas simulações com a cooperação via estruturas secundárias com peso estipulado em um décimo da energia os modelos tidos como mais adaptados para serem representantes da estrutura nativa da 1ZDD são os modelos 207 e 114 provenientes das simulações 3 e 6, respectivamente. Pela análise do mapa de Ramachandran é possível verificar que em ambos os casos cinco resíduos adotam posições não permitidas do mapa de Ramachandran. Quando tratamos da verificação das posições de todos os resíduos em comparação ao mapa de Ramachandran proveniente da estrutura experimental nota-se que o modelo 114 é aquele que possui maior conformidade principalmente no que diz respeito às hélices que compõem a proteína, o que explica os melhores valores em termos de TM-Score, MaxSub e GDT obtidos.

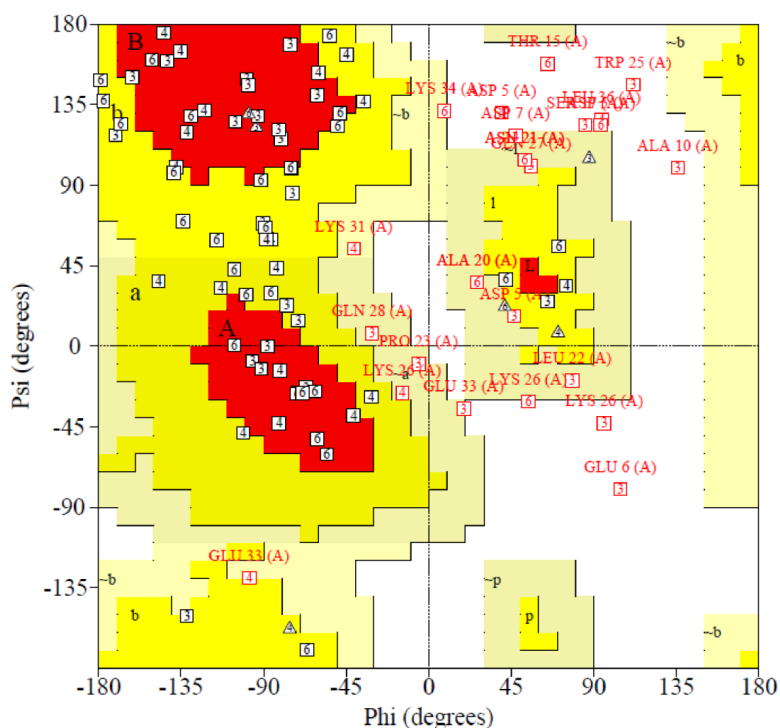


Figura 58 - Diagrama de Ramachandran dos modelos obtidos nas simulações 3, 4 e 6. Proteína PDB ID: 1VII. Cooperação via estruturas secundárias ativada, com peso = energia/10.

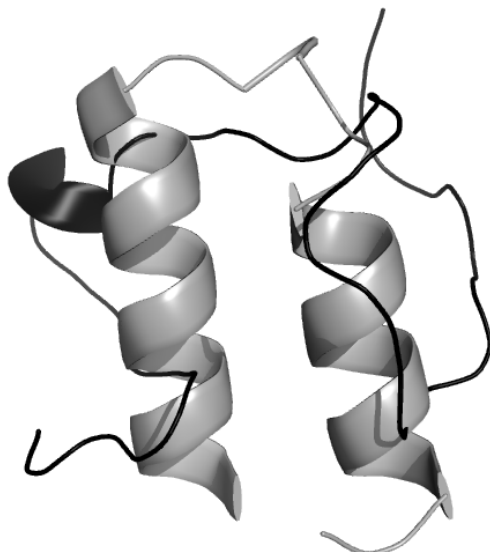


Figura 59 - Estrutura experimental (em cinza) e estrutura predita (em preto) pela simulação 6, proteína de PDB ID: 1ZDD. Cooperação via estrutura secundária ativada com peso = energia/10.

Tabela 22 - Simulações com cooperação via estrutura secundária. Proteína de PDB ID: 1VII. Peso = energia/10.

Simulação	1	2	3	4	5	6	7	8	9	10	Média ( $\bar{O}$ )	Desvio ( $\sigma$ )
RMSD (Å)	8,50	7,50	8,60	5,90	8,20	7,00	7,30	6,90	8,70	9,90	7,85	1,15
TM-Score	0,15	0,15	0,17	0,16	0,15	0,16	0,15	0,13	0,14	0,14	0,15	0,01
MaxSub	0,19	0,18	0,29	0,23	0,20	0,21	0,19	0,17	0,17	0,19	0,20	0,03
GDT_TS	0,32	0,31	0,33	0,40	0,30	0,37	0,33	0,35	0,29	0,31	0,33	0,03
GDT_HA	0,19	0,16	0,22	0,20	0,17	0,21	0,18	0,17	0,17	0,17	0,18	0,02
Energia	0,00	0,00	-0,02	18,70	0,00	19,00	-40,00	0,00	33,00	-15,00	1,57	20,00
Tick	115	149	132	229	204	173	188	145	105	128	156	40

Com a cooperação via estruturas secundárias com peso estipulado em um décimo da energia para a proteína 1VII os resultados demonstraram uma considerável melhora em termos de RMSD em relação à estrutura experimental, entretanto a análise estereoquímica revela que os resíduos ainda assim não foram capazes de adotar as devidas regiões.

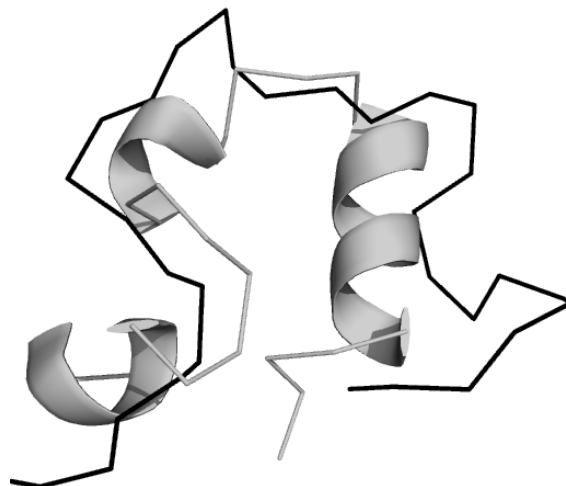


Figura 60 - Estrutura experimental (em cinza) e estrutura predita (em preto) pela simulação 6, proteína de PDB ID: 1VII. Cooperação via estrutura secundária ativada com peso = energia/10.

Tabela 23 - Simulações com cooperação via estrutura secundária. Proteína de PDB ID: 2GP8. Peso = energia/10.

Simulação	1	2	3	4	5	6	7	8	9	10	Média ( $\bar{O}$ )	Desvio ( $\sigma$ )
RMSD (Å)	7,00	7,90	9,50	8,90	8,10	9,40	9,40	9,10	7,30	7,30	8,39	0,98
TM-Score	0,19	0,15	0,15	0,13	0,15	0,15	0,17	0,14	0,16	0,15	0,15	0,02
MaxSub	0,26	0,17	0,18	0,15	0,17	0,19	0,23	0,18	0,18	0,16	0,19	0,03
GDT_TS	0,36	0,29	0,29	0,25	0,28	0,31	0,30	0,31	0,35	0,30	0,30	0,03
GDT_HA	0,20	0,16	0,17	0,13	0,16	0,18	0,19	0,16	0,18	0,14	0,17	0,02
Energia	-0,90	12,98	55,94	0,10	0,04	43,00	-0,03	0,07	0,01	-0,30	11,09	20,86
Tick	228	225	117	159	173	120	167	220	173	192	177	39

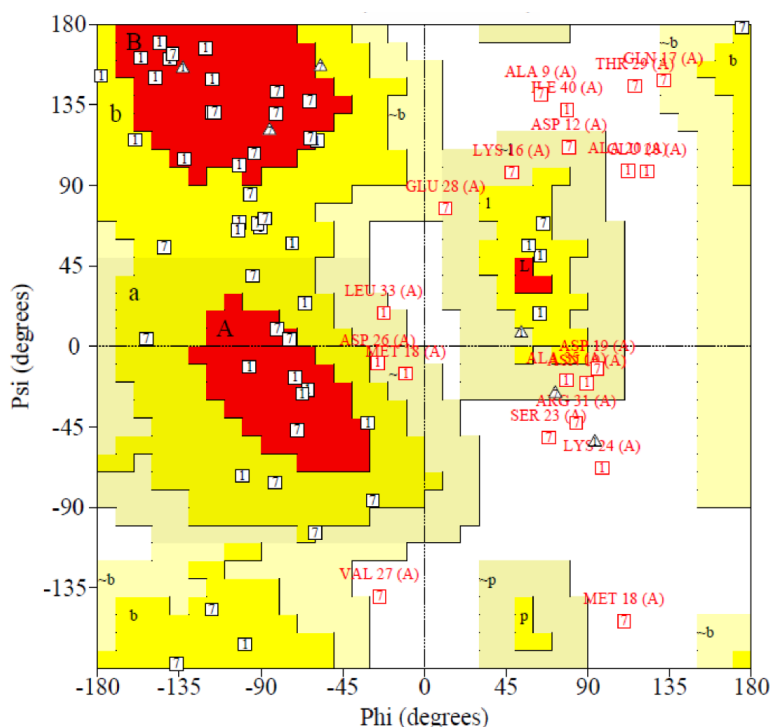


Figura 61 - Diagrama de Ramachandran dos modelos obtidos nas simulações 1 e 7. Proteína PDB ID: 2GP8. Cooperação via estruturas secundárias ativada, com peso = energia/10.

Desta vez utilizando como peso um décimo da energia atual para a cooperação via estruturas secundárias, foi possível notar certo avanço em tanto em termos estereoquímicos quanto em termos de medidas relacionadas à geometria.

Análise do mapa de Ramachandran disposto na Figura 61 contendo os melhores modelos segundo critérios de geometria revela que, assim como na cooperação com peso equivalente a 1, houve melhora em se tratando do percentual de resíduos em regiões não permitidas. Além disso, neste caso, houve significativa melhora em termos de RMSD, o que se refletiu em todos os outros critérios de similaridade. A seguir está disposto o modelo 228, proveniente da simulação 1, o qual obteve melhor desempenho no que diz respeito à posicionamento dos resíduos pertencentes às hélices.

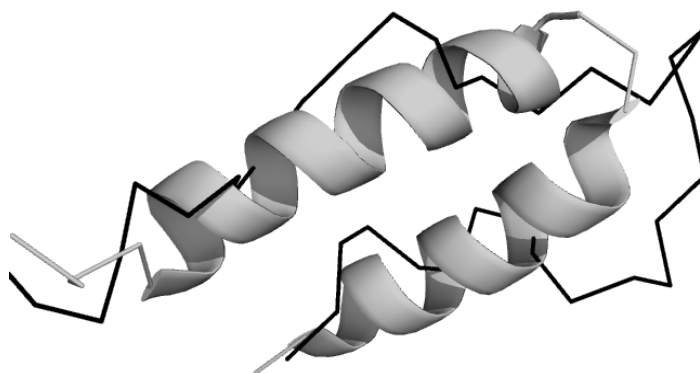


Figura 62 - Estrutura experimental (em cinza) e estrutura predita (em preto) pela simulação 1, proteína de PDB ID: 2GP8. Cooperação via estrutura secundária ativada com peso = energia/10.

### 6.3.3 Comparação

Nas Tabelas 25 e 26 é possível notar a diferença de desempenho entre a técnicas de cooperação utilizada por Bortolussi *et al.* e as utilizadas pelo PRO-SMART:

Tabela 24 – Valores (em Å) da comparação em termos de RMSD entre o PROSMART e o *framework* de Bortolussi *et al.* Cooperação via estruturas secundárias ativada com peso = 1.

PDB ID	PROSMART	Bortolussi <i>et al.</i>
1EDP	<b>5,99 ±0,75</b>	6,01 ±0,48
1ZDD	9,49 ±2,76	<b>6,83 ±1,24</b>
1VII	12,46 ±3,06	<b>8,03 ±1,13</b>
2GP8	12,86 ±3,93	<b>5,38 ±1,45</b>

Tabela 25 - Comparação em termos de energia entre o PROSMART e o *framework* de Bortolussi *et al.* Cooperação via estruturas secundárias ativada com peso = 1.

PDB ID	PROSMART	Bortolussi <i>et al.</i>
1EDP	<b>-16,43 ±7,75</b>	-4,72 0,81
1ZDD	<b>-19,76 ±16,88</b>	-10,70 ±0,83
1VII	<b>-43,85 ±18,73</b>	-9,02 ±1,29
2GP8	<b>-27,82 ±14,55</b>	-7,75 ±1,05

Tabela 26 - Valores (em Å) da comparação em termos de RMSD entre o PROSMART e o *framework* de Bortolussi *et al.* Cooperação via estruturas secundárias ativada com peso estipulado em 1/10 da energia.

PDB ID	PROSMART	Bortolussi <i>et al.</i>
1EDP	<b>5,77 (1,26)</b>	6,01 (0,48)
1ZDD	8,77 (1,63)	<b>6,83 (1,24)</b>
1VII	<b>7,85 (1,15)</b>	8,03 (1,13)
2GP8	8,39 (0,98)	<b>5,38 (1,45)</b>

Tabela 27 - Comparação em termos de energia entre o PROSMART e o *framework* de Bortolussi *et al.* Cooperação via estruturas secundárias ativada com peso estipulado em 1/10 da energia.

PDB ID	PROSMART	Bortolussi <i>et al.</i>
1EDP	<b>-9,42 ±9,82</b>	-4,72 ±0,81
1ZDD	10,32 ±13,16	<b>-10,70 ±0,83</b>
1VII	1,57 ±20,00	<b>-9,02 ±1,29</b>
2GP8	11,09 ±20,86	<b>-7,75 ±1,05</b>

Conforme já exposto no início da seção 6.4.1, existe uma grande diferença no fundamento das abordagens utilizadas pelo PRO-SMART e por Bortolussi *et al.* quanto à cooperação via estruturas secundárias e, tendo em mente que o PRO-SMART procurou se caracterizar como um preditor estritamente *ab initio*, a comparação entre as estratégias para tal se mostra como um grande desafio.

Análise das tabelas revela que a utilização de cooperação via estruturas secundárias se mostrou, de fato, algo bastante interessante e com potencial para futuramente ser ainda mais explorado pela ferramenta. Os resultados referentes às simulações envolvendo a mesma abordagem de Bortolussi *et al.* no que diz respeito ao peso dado à cooperação (peso=1) foram desanimadores pois, embora o desempenho do PRO-SMART tenha sido melhor para a proteína 1EDP, a cooperação diminuiu a qualidade das predições em comparação aos testes da 1EDP sem cooperação.

Em contrapartida, a nova estratégia utilizada para atribuir pesos para a cooperação via estrutura secundária desenvolvida por esta dissertação mostrou ser capaz de melhorar a qualidade das predições e chegou a obter melhores resultados em comparação a Bortolussi *et al.* para dois dos quatro polipeptídicos em que os testes foram efetuados.

No que diz respeito à energia, a abordagem que utiliza peso = 1 obteve valores significativamente menores, simbolizando uma melhor exploração do espaço conformacional por parte dos agentes. Quanto aos desvios padrão encontrados, assim como nos testes sem cooperação, houve uma maior variação em decorrência da utilização de clusterização. Quando analisada a Tabela 2 é possível perceber que os valores de energia obtidos pelas conformações do PRO-SMART aumentaram bastante em relação aos testes anteriores, certamente consequência das punições em termos de energia aplicadas pelo agente Ambiente, o que revela a maior importância dada às estruturas secundárias se comparado aos outros testes.



#### 6.4 Desempenho da Cooperação via Estruturas Secundárias

Como propósito de observar o desempenho do PRO-SMART utilizando-se das diferentes abordagens no que diz respeito à cooperação via estrutura secundária, decidiu-se criar as seguintes tabelas comparativas contendo os resultados obtidos pelas simulações tendo como alvo as proteínas de PDB ID: 1EDP, 1ZDD, 1VII e 2GP8.

O estudo dos resultados obtidos presentes nas tabelas a seguir revela-nos dois principais pontos merecedores de atenção.

Primeiramente, foi comprovado que os pesos estipulados para as punições em termos de energia impostas pelo agente Ambiente influenciam diretamente nas conformações encontradas, deixando claro que a utilização de peso=1 não beneficia o sistema, obtendo piores resultados em termos de RMSD em 100% dos casos e na grande maioria dos casos para as medidas TM-Score, MaxSub e GDT. Já a cooperação utilizando a estratégia de um décimo da energia mostrou-se mais efetiva, melhorando a qualidade das predições significativamente para as proteínas 1VII e 2GP8 e mantendo-se praticamente estável para as proteínas 1ZDD e 1EDP, embora tenha obtido melhores valores em termos de MaxSubs, por exemplo.

Em segundo lugar cabe a análise mais específica dos motivos pelos quais a utilização de cooperação não ter sido ainda melhor. No caso das simulações com peso=1 nota-se (via verificação dos gráficos de energia) que a ativação da cooperação (a qual ocorre no *tick* 100) não possui o efeito esperado pois, dependendo da energia atual do sistema, as punições de energia são superestimadas ou ignoradas. Neste quesito a incorporação de uma estratégia que se utiliza da energia atual do sistema para calcular a punição em termos de energia dada à conformação se mostrou bastante efetiva, resolvendo o problema da super ou subestimação. Ainda assim, o PRO-SMART alcançou resultados satisfatórios apenas no que diz respeito a duas das quatro proteínas testadas com a cooperação ativada e o motivo para tal, buscando entender os meandros que diferem os dois métodos sem descaracterizar o PRO-SMART, foi possível notar que a queda de rendimento do PRO-SMART acontece justamente quando as predições envolvem estruturas secundárias de tamanho grande. Se compararmos o tamanho médio das hélices presentes em 1EDP e 1VII (onde nosso sistema obteve melhor desempenho se comparado à Bortolussi *et al.*) chegamos ao número de, em média 6,2 resíduos por hélice, enquanto para 1ZDD e 2GP8 (onde Bortolussi *et al.* obteve vantagem) este número chega a 13,5 resíduos por hélice. Uma solução alternativa para tal seria a implementação de pseudo-

hélices menores representando uma hélice maior, algo que pode ser facilmente adicionado à ferramenta.

Em termos de tempo computacional, as simulações levam em média e levando-se em conta primeiramente as menores proteínas e posteriormente as maiores, entre 19 e 96 minutos sem cooperação, entre 28 e 260 minutos com cooperação peso=1 e entre 25 e 221 minutos com cooperação com peso relacionado à energia atual do sistema. A diferença no que diz respeito ao limiar superior de tempo encontrado nas simulações com diferentes pesos pode ser explicado pelo fato de, nas simulações com peso relacionado à energia atual da conformação, o sistema encontrar um maior número de conformações que melhoram a energia, assim evitando cálculos provenientes do método de Monte Carlo, o que reflete no tempo final de simulação.

Tabela 28 - Desempenho dos diferentes tipos de cooperação via estrutura secundária. Proteína de PDB ID: 1EDP. Melhores resultados em negrito.

Tipo de Cooperação	Sem Cooperação	Com Cooperação peso=1	Com Cooperação peso=energia/10
RMSD (Å)	5,95 ±1,39	5,99 ±0,75	<b>5,77 ±1,26</b>
TM-Score	0,14 ±0,04	<b>0,15 ±0,03</b>	<b>0,15 ±0,05</b>
MaxSub	0,37 ±0,05	0,36 ±0,04	<b>0,39 ±0,06</b>
GDT_TS	<b>0,52 ±0,05</b>	0,51 ±0,04	0,51 ±0,04
GDT_HA	<b>0,33 ±0,03</b>	<b>0,33 ±0,04</b>	<b>0,33 ±0,03</b>
Energia	<b>-28,84 ±11,01</b>	-16,43 ±7,75	-9,42 ±9,82
Tick	182 ±26	189 ±29	172 ±20

Tabela 29 - Desempenho dos diferentes tipos de cooperação via estrutura secundária. Proteína de PDB ID: 1VII. Melhores resultados em negrito.

Tipo de Cooperação	Sem Cooperação	Com Cooperação peso=1	Com Cooperação peso=energia/10
RMSD (Å)	9,33 ±1,20	12,46 ±3,06	<b>7,85 ±1,15</b>
TM-Score	<b>0,15 ±0,02</b>	0,14 ±0,01	<b>0,15 ±0,01</b>
MaxSub	<b>0,20 ±0,04</b>	<b>0,20 ±0,02</b>	<b>0,20 ±0,03</b>
GDT_TS	0,30 ±0,04	0,28 ±0,03	<b>0,33 ±0,03</b>
GDT_HA	<b>0,18 ±0,02</b>	<b>0,18 ±0,02</b>	<b>0,18 ±0,02</b>
Energia	<b>-54,97 ±15,50</b>	-43,85 ±18,73	1,57 ±20,00
Tick	188 ±19	198 ±11	156 ±40

Tabela 30 – Desempenho dos diferentes tipos de cooperação via estrutura secundária. Proteína de PDB ID: 1ZDD. Melhores resultados em negrito.

Tipo de Cooperação	Sem Cooperação	Com Cooperação Peso=1	Com Cooperação Peso=energia10
RMSD (Å)	<b>8,44 ±0,59</b>	9,49 ±2,76	8,77 ±1,63
TM-Score	<b>0,15 ±0,01</b>	0,14 ±0,02	<b>0,15 ±0,03</b>
MaxSub	0,19 ±0,02	0,21 ±0,02	<b>0,22 ±0,05</b>
GDT_TS	<b>0,33 ±0,03</b>	<b>0,33 ±0,04</b>	<b>0,33 ±0,04</b>
GDT_HA	<b>0,19 ±0,02</b>	0,18 ±0,01	<b>0,19 ±0,02</b>
Energia	<b>-42,95 ±18,68</b>	-19,76 ±16,88	10,32 ±13,16
Tick	192 ±20	186 ±19	167 ±39

Tabela 31 – Desempenho dos diferentes tipos de cooperação via estrutura secundária. Proteína de PDB ID: 2GP8. Melhores resultados em negrito.

Tipo de Cooperação	Sem Cooperação	Com Cooperação Peso=1	Com Cooperação Peso=energia10
RMSD (Å)	13,12 ±1,44	12,86 ±3,93	<b>8,39 ±0,98</b>
TM-Score	0,14 ±0,01	0,14 ±0,02	<b>0,15 ±0,02</b>
MaxSub	0,15 ±0,02	0,18 ±0,02	<b>0,19 ±0,03</b>
GDT_TS	0,25 ±0,03	0,27 ±0,02	<b>0,30 ±0,03</b>
GDT_HA	0,16 ±0,02	0,16 ±0,01	<b>0,17 ±0,02</b>
Energia	<b>-54,04 ±14,89</b>	-27,82 ±14,55	11,09 ±20,86
Tick	205 ±10,41	189 ±24	177 ±39

## 7. CONSIDERAÇÕES FINAIS

A abordagem deste trabalho, embora incomum na bioinformática atual, se mostrou capaz de alcançar bons resultados. Os três principais aspectos conceituais explorados pelo trabalho (nível de abstração, função de energia e cooperação) demonstraram que, se bem regulados, podem prover estruturas próximas às estruturas nativas, entretanto, a regulação destes aspectos é uma tarefa nada trivial, pois envolve uma gama muito grande de variáveis atuando em paralelo, o que prejudica a execução de testes e a obtenção de um conjunto otimizado de diretrizes.

No que diz respeito ao nível de abstração e à função de energia escolhida para testar o desempenho do PRO-SMART se verificou algo que já era esperado: O fato da função de energia possuir baixa resolução. Uma função de baixa resolução acarreta na obtenção de estruturas com energias muito baixas, mas que não possuem validade estereoquímica ou se distanciam das conformações nativas. Quanto aos agentes envolvidos, o único agente afetado por uma mudança em termos de função de energia(e conseqüentemente em termos de abstração) seria os agentes de mais baixo nível ou no caso agentes aminoácidos (C-Alfa e C-Beta).

A implantação de cooperação entre os agentes se mostrou um aspecto de suma importância para a melhoria da predição. Através de cooperação os agentes se mostraram capaz de alcançar melhores resultados se comparado aos testes em que atuaram sem cooperação. Entretanto, o esquema de cooperação mostrou também alta complexidade, sendo necessários estudos mais aprofundados para que seja possível a obtenção de melhorias ainda mais significativas.

Fica comprovado também, tendo em vista a análise dos resultados do PRO-SMART, que as medidas que tratam de similaridade geométrica entre proteínas devem ser utilizadas em conjunto, a fim de diminuir possíveis erros provenientes de avaliações restritas a apenas uma medida.

Em termos gerais o PRO-SMART se mostrou capaz de obter resultados regulares para prever estruturas de polipeptídicos e, embora se afaste do estado da arte no âmbito geral entre todos os programas existentes para tratamento do problema, quando comparado com trabalhos que utilizam abordagens de mesma fundamentação, os quais adentram o principal objetivo de trabalho da dissertação (como é o caso do arcabouço de Bortolussi *et al.*) os resultados são satisfatórios, com a obtenção de melhores valores para a 4 das 9 proteínas testadas.

## 7.1 Principais Contribuições

As principais contribuições resultantes desta dissertação no endereçamento do problema da predição de estruturas terciárias de proteínas são:

- Uma ferramenta capaz de predizer estruturas de proteínas em tempo hábil se comparado às abordagens usuais para predição. Ferramenta esta com capacitação para ser utilizada em qualquer sistema operacional;
- Uma interface gráfica pela qual o usuário pode adicionar aminoácido por aminoácido de sua proteína e, além disso, acompanhar o enovelamento da mesma, a partir de uma função de energia que escolher;
- Uma nova estratégia de busca de conformações tridimensionais baseada em Monte Carlo e arrefecimento simulado aliado à uma arquitetura hierárquica baseada na quantidade de movimentos efetuados por cada aminoácido, no caso representados por diferentes agentes autônomos e;
- A utilização de clusterização com a finalidade de obter-se o nicho mais acessado com baixa energia e não somente a última estrutura encontrada pela simulação mostrou-se uma alternativa interessante para tratar da escolha de modelos provenientes da simulação, abordagem a qual pode ter sua utilização estendida também à trajetórias de dinâmica molecular.

## 7.2 Trabalhos Futuros

Em se tratando de melhoria imediata, quatro importantes pontos são dignos de serem enfatizados. O primeiro ponto foi notado apenas após a finalização dos testes e análise dos resultados comparativos e diz respeito ao conjunto de proteínas escolhido para ser alvo dos testes aqui dispostos. Os testes foram feitos utilizando-se como alvo apenas as proteínas publicadas pelo trabalho *seed* de nossa abordagem e, assim sendo, o PRO-SMART se tornou restrito a alcançar bons resultados somente e exatamente nas proteínas onde o arcabouço de Bortolussi *et al.* obteve seus melhores resultados. Restrição essa que obviamente não modificaria o desempenho comparado da ferramenta, mas que impossibilitou uma melhor análise de desempenho geral do PRO-SMART. Em segundo lugar, no que está relacionado à melhoria imediata, está a utilização de uma função de energia que possua resolução mais

elevada. Como terceira melhoria imediata acredita-se que novas abordagens passíveis de utilização na implantação de cooperação por estruturas secundárias devem ser estudadas, principalmente no que diz respeito a folhas  $\beta$ . Além disso, elencamos também a adição de refinamento para as estruturas originárias do PRO-SMART. Dinâmicas moleculares de poucos nanossegundos podem, a partir de uma estrutura com topologia já estipulada, otimizar de melhor forma as cadeias laterais assim como efetuar pequenas mudanças no traço de carbonos alfas a fim de obtermos, sem a necessidade de dispendir muito tempo computacional, uma conformação de ainda maior qualidade. As figuras 63 e 64 a seguir representam um teste inicial onde o modelo conformacional obtido para a proteína 1EDP com cooperação via estrutura secundária com peso=1, submetido a uma dinâmica de 30ns ( $30 \times 10^{-9}$  segundos), conseguiu otimizar suas cadeias laterais que anteriormente se colocavam em posições não permitidas segundo o mapa de Ramachandran da Figura 55. Como resultado obteve-se a obtenção do padrão da hélice, além da diminuição do RMSD, de  $5,0 \text{ \AA}$  para  $2,62 \text{ \AA}$ .

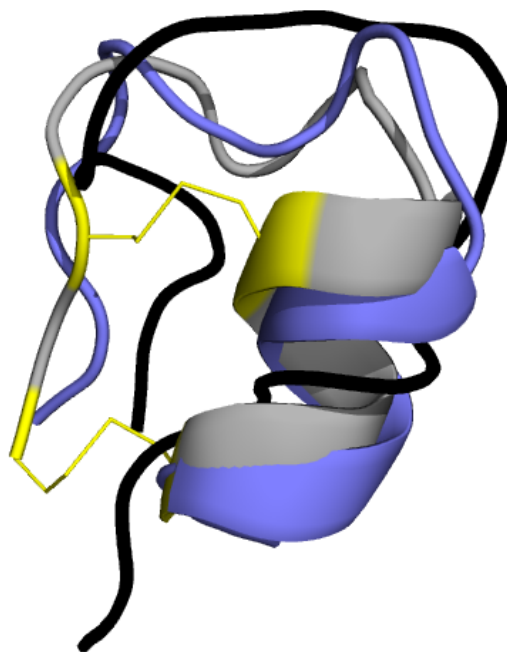


Figura 63 - Estrutura experimental (em cinza), estrutura predita (em preto) pela simulação 1 e estrutura refinada (em azul), proteína de PDB ID: 1EDP. Cooperação via estrutura secundária ativada com peso = energia/10. Ligações dissulfídicas em amarelo.

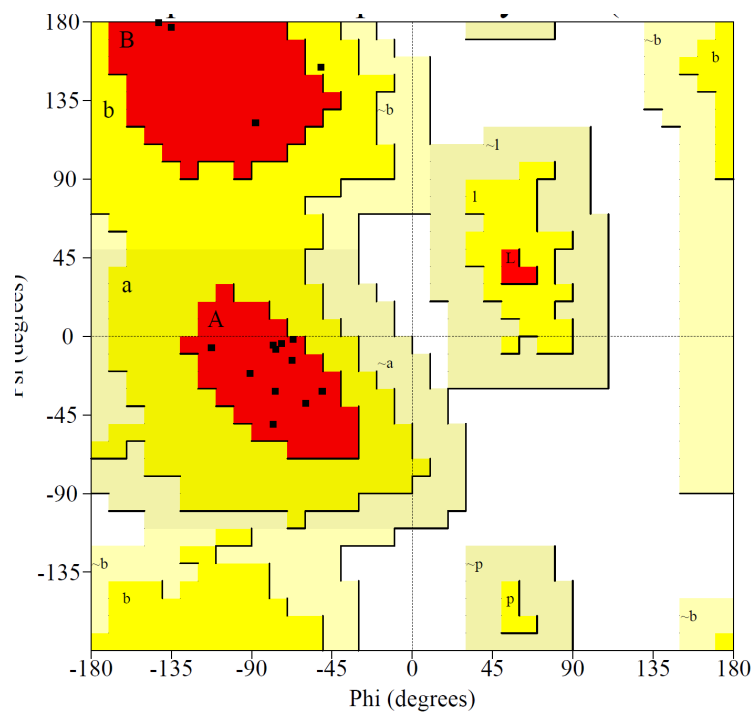


Figura 64 - Mapa de Ramachandran para a estrutura refinada proveniente da simulação 1, proteína de PDB ID: 1EDP. Cooperação via estrutura secundária ativada com peso =energia/10.



## REFERÊNCIAS

- [1] A. Addis, G. Armano, F. Mascia, E. Vargiu, "Protein Secondary Structure Prediction through a Cooperative MultiAgent Learning Approach," *Education*, vol. 2(1), pp. 122-125, 2007.
- [2] L. O. C. Alvares e J. S. Sichman, "Introdução aos Sistemas Multiagentes," em *Jornada De Atualização Em Informática*, 1997, pp. 1-37.
- [3] F. Amigoni e V. Schiaffonati, "Multiagent-Based Simulation in Biology," 2007, pp. 179-191.
- [4] N. Andersen, C. Chen, T. Marschner, S. J. Krystek, e D. Bassolino, "Conformational isomerism of endothelin in acidic aqueous media: a quantitative NOESY analysis.," vol. 5, 31 ed. *Biochemistry*, 1992, pp. 1280-95.
- [5] C. B. Anfinsen, "Principles that govern the folding of protein chains," *Science*, vol. 181, pp. 223-230, 1973.
- [6] G. Armano, G. Cherchi, A. Manconi e E. Vargiu, "PACMAS : A Personalized , Adaptive , and Cooperative MultiAgent System Architecture," 6th Workshop dagli Oggetti agli Agenti WOA 2005, pp. 54-60, 2005.
- [7] J. A. Bachman and P. Sorger, "New approaches to modeling complex biochemistry," *Nature Methods*, vol. 8, pp. 130-131, 2011.
- [8] D. A. Benson, I. Karsch-Mizrachi, K. Clark, D. J. Lipman, J. Ostell e E. W. Sayers, "GenBank," *Nucleic Acids Research*, vol. 40, pp. D48-D53, 2012.
- [9] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, *et al.*, "The Protein Data Bank," *Nucleic Acids Research*, vol. 28, pp. 235-242, 2000.
- [10] M. Berrera, H. Molinari e F. Fogolari, "Amino acid empirical contact energy definitions for fold recognition in the space of contact maps," *BMC Bioinformatics*, vol. 4, p. 8+, 2003.
- [11] G. D. Birkhoff, "What is the Ergodic Theorem?," *The American Mathematical Monthly*, vol. 49, pp. 222-226, 1942.
- [12] L. Bortolussi, A. Dovier e F. Fogolari, "Agent-based Protein Structure Prediction," *Multiagent and Grid Systems - Multi-agent systems for medicine, computational biology, and bioinformatic*, 2005.
- [13] L. Bortolussi, A. Dovier e F. Fogolari, "Multi-agent simulation of protein folding," in *International Workshop on Multi-Agent Systems for Medicine, Computational Biology, and Bioinformatics*, 2005.
- [14] L. Bortolussi, A. D. Palú, A. Dovier e F. Fogolari, "Protein folding simulation in CCP," in *In Proceedings of BioConcur*, 2004.

- [15] J. Boyle, "Lehninger principles of biochemistry (4th ed.): Nelson, D. e Cox, M.," *Biochemistry and Molecular Biology Education*, vol. 33, pp. 74-75, 2005.
- [16] N. Cannata, F. Corradini, E. Merelli, A. Omicini e A. Ricci, "An Agent-oriented Conceptual Framework for Biological Systems Simulation," *Models and Methaphors from Biology to Bioinformatics Tools*, pp. 167-180, 2004.
- [17] A. A. Canutescu, A. A. Shelenkov e R. L. Dunbrack, "A graph-theory algorithm for rapid protein side-chain prediction," *Protein science : a publication of the Protein Society*, vol. 12, pp. 2001-2014, 09/ 2003.
- [18] C. Clementi, "Coarse-grained models of protein folding: toy models or predictive tools?," *Current Opinion in Structural Biology*, vol. 18, pp. 10-15, 2// 2008.
- [19] A. G. Cochran, N. J. Skelton e M. A. Starovasnik, "Tryptophan zippers: Stable, monomeric  $\beta$ -hairpins," *Proceedings of the National Academy of Sciences*, vol. 98, pp. 5578-5583, 2001.
- [20] G. Colombo and C. Micheletti, "Protein folding simulations: combining coarse-grained models and all-atom molecular dynamics," *Theoretical Chemistry Accounts*, vol. 116, pp. 75-86, Aug 2006.
- [21] E. Costa and A. Simões, *Inteligência Artificial: Fundamentos e Aplicações*: FCA-Editora de Informatica, 2008.
- [22] P. Crescenzi, D. Goldman, C. Papadimitriou, A. Piccolboni e M. Yannakakis, "On the Complexity of Protein Folding," *Journal of Computational Biology*, vol. 5, pp. 597-603, 1998.
- [23] K. C. d. M. Dall'Agno, "Um estudo sobre a predição da estrutura 3D aproximada de proteínas utilizando o método CReF com refinamento.," Master, Pontificia Universidade Católica do Rio Grande do Sul, 2012.
- [24] S. Das, A. Abraham e A. Konar, "Swarm Intelligence Algorithms in Bioinformatics," *Computational Intelligence in Bioinformatics*, vol. 147, pp. 113-147, 2008.
- [25] K. A. Dill and J. L. MacCallum, "The Protein-Folding Problem, 50 Years On," *Science*, vol. 338, pp. 1042-1046, 2012.
- [26] Y. Duan and P. A. Kollman, "Computational protein folding: from lattice to all-atom," *IBM Systems Journal*, vol. 40, pp. 297-309, 2001.
- [27] A. V. Efimov, "Standard structures in proteins," *Progress in Biophysics and Molecular Biology*, vol. 60, pp. 201-239, 1993.
- [28] R. Fahrner, T. Dieckmann, S. Harwig, R. Lehrer, D. Eisenberg e J. Feigon, "Solution structure of protegrin-1, a broad-spectrum antimicrobial peptide from porcine leukocytes.," vol. 7, 3 ed. *Chemistry & Biology*, 1996, pp. 543-50.

- [29] J. Ferber, *Multi-Agent Systems - An Introduction to Distributed Artificial Intelligence*: Addison-Wesley, 1999.
- [30] C. A. Floudas, "Computational methods in protein structure prediction," *Biotechnology and Bioengineering*, vol. 97, pp. 207-213, 2007.
- [31] F. Fogolari, G. Esposito, P. Viglino e S. Cattarinussi, "Modeling of polypeptide chains as C alpha chains, C alpha chains with C beta, and C alpha chains with ellipsoidal lateral chains," *Biophysical Journal*, vol. 70, pp. 1183-1197, 1996.
- [32] A. C. B. Garcia and J. S. Sichman, "Sistemas Inteligentes - Fundamentos e Aplicações, de Rezende, S.O.; Prati, R," Manole, Ed., ed, 2003.
- [33] N. Gilbert and K. G. Troitzsch, *Simulation for the Social Scientist*: Open University Press, 2005.
- [34] D. Gront, S. Kmiecik e A. Kolinski, "Backbone building from quadrilaterals: a fast and accurate algorithm for protein backbone reconstruction from alpha carbon coordinates," *Journal of computational chemistry*, vol. 28, pp. 1593-1597, 07/ 2007.
- [35] G. Helles, "A comparative study of the reported performance of ab initio protein structure prediction algorithms," *Journal of the Royal Society Interface*, vol. 5, pp. 387-396, 2008.
- [36] A. Hinchliffe, "CHEMDRAW-PRO FOR WINDOWS," *Theochem-Journal of Molecular Structure*, vol. 120, pp. 335-336, Nov 1994.
- [37] W. S. Hlavacek, J. R. Faeder, M. L. Blinov, R. G. Posner, M. Hucka e W. Fontana, "Rules for Modeling Signal-Transduction Systems," *Science Signal Transduction Knowledge Environment*, vol. 2006, p. re6+, 2006.
- [38] W. Humphrey, A. Dalke e K. Schulten, "VMD: Visual molecular dynamics," *Journal of Molecular Graphics*, vol. 14, pp. 33-38, 1996.
- [39] P. Humphreys, "Computer Simulations," in *Proceedings of the Biennial Meeting of the Philosophy of Science Association*, 1990, pp. 497-506.
- [40] H. Jin and I.-C. Kim, "Plan-Based coordination of a multi-agent system for protein structure prediction," in *Proceedings of the 13th international conference on AI, Simulation, and Planning in High Autonomy Systems*, 2005, pp. 224-232.
- [41] D. T. Jones, W. R. Taylor e J. M. Thornton, "A new approach to protein fold recognition," *Nature*, vol. 358, pp. 86-89, 1992.
- [42] K. Karplus, "SAM-T08, HMM-based protein structure prediction," *Nucleic Acids Research*, vol. 37, pp. 492-497, 2009.
- [43] A. Kolinski and J. Skolnick, "Reduced models of proteins and their applications," *Polymer*, vol. 45, pp. 511-524, 2004.

- [44] R. A. Laskowski, M. W. Macarthur, D. S. Moss e J. M. Thornton, "PROCHECK: a program to check the stereochemical quality of protein structures," *J. Appl. Cryst.*, vol. 26, pp. 283-291, 1993.
- [45] A. Leach, *Molecular modelling : principles and applications*: Pearson Prentice Hall, 2001.
- [46] A. M. Lesk, *Introduction to Protein Architecture: The Structural Biology of Proteins*: Oxford University Press, 2000.
- [47] A. M. Lesk, *Introduction to bioinformatics*, 3rd ed. Oxford ; New York: Oxford University Press, 2008.
- [48] C. Levinthal, "Are there pathways for protein folding?," *Journal of Medical Physics*, vol. 65, pp. 44-45, 1968.
- [49] T. Lima, S. Faria, B. Soares Filho e T. Carneiro, "Modelagem de sistemas baseada em agentes: Alguns conceitos e ferramentas," in *Anais do XIV Simpósio Brasileiro de Sensoriamento Remoto*, 2009, pp. 5279-5286.
- [50] T. Lipinski-Paes and O. Norberto de Souza, "Cooperative multi-agent system for protein structure prediction," in 8th International Conference of the Brazilian Association for Bioinformatics and Computational Biology, Campinas - Brasil, 2012 p. 117.
- [51] A. Liwo, M. Khalili e H. A. Scheraga, "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, pp. 2362-2367, 02/ 2005.
- [52] M. Luck and E. Merelli, "Agents in bioinformatics," *The Knowledge Engineering Review*, vol. 20, pp. 117-125, 2005.
- [53] N. M. Luscombe, D. Greenbaum e M. Gerstein, "What is bioinformatics? A proposed definition and overview of the field," *Methods of information in medicine*, vol. 40, pp. 346-358, 2001.
- [54] M. A. Martí-Renom, A. C. Stuart, A. Fiser, R. SÁnchez, F. Melo e A. Sali, "Comparative protein structure modeling of genes and genomes," *Annual review of biophysics and biomolecular structure*, vol. 29, pp. 291-325, 2000.
- [55] C. McKnight, P. Matsudaira e P. Kim, "NMR structure of the 35-residue villin headpiece subdomain.," vol. 3, 4 ed. *Nature Structural Biology*, 1997, pp. 180-4.
- [56] E. Merelli, G. Armano, N. Cannata, F. Corradini, M. d'Inverno, A. Doms, *et al.*, "Agents in bioinformatics, computational and systems biology," *Briefings in bioinformatics*, vol. 8, pp. 45-59, 2007.
- [57] M. L. Mit Epistemology Group, "Introduction to StarLogo," ed, 2001.

- [58] L. Palopoli and G. Terracina, "A framework for improving protein structure predictions by teamwork," in *Proceedings of the First Asia-Pacific Conference on Bioinformatics*, 2003, pp. 163-171.
- [59] A. Pavlopoulou and I. Michalopoulos, "State-of-the-art bioinformatics protein structure prediction tools (Review)," *International Journal of Molecular Medicine*, vol. 28, pp. 295-310, 2011.
- [60] O. Pedreira, M. Piattini, M. R. Luaces e N. R. Brisaboa, "A systematic review of software process tailoring," *SIGSOFT Softw. Eng. Notes*, vol. 32, pp. 1-6, 2007.
- [61] G. Pollastri and A. McLysaght, "Porter: a new, accurate server for protein secondary structure prediction," *Bioinformatics*, vol. 21, pp. 1719-1720, 2005.
- [62] L.-H. Ren, Y.-S. Ding, Y.-Z. Shen e X.-F. Zhang, "Multi-agent-based bio-network for systems biology: protein-protein interaction network as an example," *Amino Acids*, vol. 35, pp. 565-572, 2008.
- [63] A. Roli and M. Milano, "MAGMA: A Multiagent Architecture for Metaheuristics," vol. 34, ed, 2004.
- [64] S. Romagnoli, R. Ugolini, F. Fogolari, G. Schaller, K. Urech, M. Giannattasio, et al., "NMR structural determination of viscotoxin A3 from *Viscum album L.*," vol. Pt 2, 350 ed: *Biochemical Journal*, 2000, pp. 569-577.
- [65] N. Siew, A. Elofsson, L. Rychlewski e D. Fischer, "MaxSub: an automated measure for the assessment of protein structure prediction quality," *Bioinformatics*, vol. 16, pp. 776-785, 2000.
- [66] K. T. Simons, R. Bonneau, I. Ruczinski e D. Baker, "Ab initio protein structure prediction of CASP III targets using ROSETTA," *Proteins: Structure, Function, and Genetics*, vol. 37, pp. 171-176, 1999.
- [67] N. J. Skelton, S. Russell, F. de Sauvage e A. G. Cochran, "Amino acid determinants of  $\beta$ -hairpin conformation in erythropoietin receptor agonist peptides derived from a phage display library," *Journal of Molecular Biology*, vol. 316, pp. 1111-1125, 3/8/ 2002.
- [68] C. D. Snow, E. J. Sorin, Y. M. Rhee e V. S. Pande, "How Well Can Simulation Predict Protein Folding Kinetics and Thermodynamics?," *Annual review of biophysics and biomolecular structure*, 2005.
- [69] M. A. Starovasnik, A. C. Braisted e J. A. Wells, "Structural mimicry of a native protein by a minimized binding domain," vol. 94, 19 ed. *Proceedings of the National Academy of Sciences of the United States of America*, 1997, pp. 10080-10085.
- [70] Y. Sun, M. H. Parker, P. Weigele, S. Casjens, P. E. Prevelige Jr e N. R. Krishna, "Structure of the coat protein-binding domain of the scaffolding protein from a double-stranded DNA virus," *Journal of Molecular Biology*, vol. 297, pp. 1195-1202, 4/14/ 2000.

- [71] J. Tisseau, *Virtual reality, in virtuo autonomy: Accreditation to Direct Research*, Université de Rennes 1, 2001.
- [72] S. Tisue and U. Wilensky, "NetLogo: A Simple Environment for Modeling Complexity," 2004.
- [73] V. Tozzini, "Coarse-grained models for proteins," *Current Opinion in Structural Biology*, vol. 15, pp. 144-150, 4// 2005.
- [74] A. Tramontano, "Integral and differential form of the protein folding problem," *Physics of Life Reviews*, vol. 1, pp. 103-127, 2004.
- [75] D. Voet and J. G. Voet, *Bioquímica*, 2006.
- [76] G. Weiss, *Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence*: MIT Press, 1999.
- [77] J. Xu and Y. Zhang, "How significant is a protein structure similarity with TM-score = 0.5?," *Bioinformatics*, vol. 26, pp. 889-895, 2010.
- [78] G. Weiss, *Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence*: MIT Press, 1999.
- [79] A. Zemla, "LGA: A method for finding 3D similarities in protein structures," *Nucleic acids research*, vol. 31, pp. 3370-3374, 07/ 2003.
- [80] Y. Zhang and J. Skolnick, "Scoring function for automated assessment of protein structure template quality," *Proteins: Structure, Function, and Bioinformatics*, vol. 57, pp. 702-710, 2004.
- [81] M. Zvelebil and J. Baum, *Understanding Bioinformatics*: Garland Science, 2007.

## APÊNDICE A: Protocolo de Mapeamento Sistemático

### 1. FORMULAÇÃO DA QUESTÃO

#### 1.1 Questões foco

O foco de interesse do mapeamento sistemático é sumarizar toda a informação existente no que se trata do tratamento do problema da predição de estrutura tridimensionais de proteína no âmbito de Sistemas Multi-agentes (SMAs), 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.

#### 1.2 Qualidade e amplitude da questão

##### 1.2.1 Problema

Predição de estrutura de proteínas através de sistemas Multiagentes.

##### 1.2.2 Questão

O que foi feito até então se utilizando de sistemas Multiagentes para o problema da predição de estrutura de proteínas?

##### 1.2.3 Palavras Chaves – Sinônimos

Para a predição de proteínas: “Protein Structure Prediction” e “Protein Folding”

Para sistemas multi-agente: “Multi-agent” e “Agent-based”.

##### 1.2.4 Intervenção

Será observado o tipo de abordagem multiagente utilizada pelo trabalho e os resultados obtidos em termos de RMSD (Root Mean Square Deviation), além de levar em conta as características específicas de cada trabalho como grau de abstração da representação das proteínas, linguagem utilizada e tamanho das proteínas.

##### 1.2.5 Controle

Artigos Seed: “Multi-Agent Simulation of Protein Folding” [2] e “Multiagent-Based Simulation” [3].

##### 1.2.6 Efeito

Descobrir quais abordagens alcançou mais sucesso até então. Quais características dos trabalhos realmente influenciaram os resultados obtidos.

##### 1.2.7 Outcome Measure

RMSD e tamanho das proteínas testadas, tempo de execução.

### 1.2.8 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 [5]. 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 novas drogas e a bioinformática possui o importante papel de acelerar o processo de evolução deste conhecimento [7].

A solução do problema PSP, ou avanços no seu tratamento, nos permitirá obter estruturas 3D de proteínas importantes, com aplicações relevantes na indústria biofarmacêutica. Ela nos permitirá compreender a estrutura de proteínas envolvidas em processos vitais, incluindo doenças como o câncer [4].

## 2. SELEÇÃO DE FONTES

### 2.1 Critério de seleção de fontes

Para a automação das pesquisas foi utilizada uma ferramenta avançada de pesquisa múltipla disponibilizada pela Pontifícia Universidade Católica do Rio Grande do Sul e baseada nas ferramentas MetaLib e SFX. Foram utilizadas 14 bases de dados e 1 catálogo On-line como alvo das palavras-chave, parte das bases de dados fortemente relacionadas à biologia e a outra parte fortemente relacionada à computação.

### 2.2 Linguagem

Inglês

### 2.3 Identificação de fontes

#### 2.3.1 Métodos de busca de fontes

Dado que as buscas são feitas através da interface de pesquisa múltipla da PUC (conforme explicado em 2.1), o método de busca é abstraído do usuário (fica a cargo das ferramentas que compõem a pesquisa múltipla). A única característica da busca por palavras-chave é a de que todas as buscas são feitas em todo o documento (não somente no abstract/palavras-chave).

#### 2.3.2 *Strings* de busca

Foram utilizadas combinações de palavras chave entre as duas grandes áreas alvo da pesquisa: Simulação Multi-agente e Predição de Proteínas. Para cada grande área foram escolhidas diferentes palavras chave, em inglês:

Para a predição de proteínas: “Protein Structure Prediction” e “Protein Folding”

Para sistemas multi-agente: “Multi-agent” e “Agent-based”



Gerando um total de 4 strings de busca:

“Protein Structure Prediction” AND “Multi-agent”

“Protein Structure Prediction” AND “Agent-based”

“Protein Folding” AND “Multi-agent” e

“Protein Folding” AND “Agent-based”

A fim de se evitar a obtenção de duplicatas, não serão feitas 4 diferentes pesquisas nas bases de dados, mas sim uma única pesquisa, contendo todas as 4 strings. A *string* será:

( ( “Protein Structure Prediction” AND “Multi-agent” ) OR ( “Protein Structure Prediction” AND “Agent-based” ) ) OR ( “Protein Folding” AND “Multi-agent” ) OR ( “Protein Folding” AND “Agent-based” ) )

### 2.3.3 Lista de fontes

O conjunto das 10 bases de dados está contido na Tabela 1, entretanto vale ressaltar que as bases de dados utilizadas são somente aquelas que passaram pelo aval do especialista na etapa de checagem de referências descrita em 2.5.

### 2.4 Seleção de fontes pós-avaliação

Bases de dados ou Catálogos Online que ao serem utilizadas na busca não retornaram nenhum resultado foram retiradas do protocolo final, são eles: Massachusetts Inst. of Technology (MIT) Libraries e Highwire Press.

### 2.5 Verificação de referências

Segundo Biolchini *et al.* em [1], a checagem da lista de bases de dados deve feita por especialista, com objetivo de retirar ou adicionar fontes . A checagem da lista de base de dados fica então, a cargo do especialista de domínio prof. Dr. Osmar Norberto de Souza.

## 3. SELEÇÃO DE ESTUDOS

### 3.1 Definição de estudos

#### 3.1.1 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 1.2.3 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 3.1.3 para serem considerados parte efetiva do conjunto de artigos que a revisão sistemática analisará.

### 3.1.2 Definição de Tipos de Estudos

Os estudos foram divididos de acordo com o tipo de abordagem multi-agente que utilizam e o nível estrutural de proteínas que possuem como alvo.

### 3.1.3 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.

1ª etapa: Escolha das palavras chave.

2ª etapa: Escolha das bases de dados.

3ª etapa: Pesquisa.

4ª etapa: Retirada de duplicatas.

5ª etapa: 1º Filtro: Leitura de Abstracts / palavras-chave.

6ª etapa: 2º Filtro: Leitura da Introdução.

7ª etapa: 3º Filtro: Leitura do artigo completo.

Tabela 1 - Lista Base de Dados

<b>Nome</b>	<b>Tipo</b>
Annual Reviews	Base de Dados
Catálogo On-line da PUCRS	Catálogo on-line
Electronic Journals (EBSCO)	Base de Dado
Google Acadêmico	Site de Busca
IEEE	Base de Dados
ProQuest (todas bases)	Base de Dados
SciELO.ORG	Base de Dados
SCOPUS (Elsevier)	Base de Dados
SpringerLink (MetaPress)	Base de Dados
ACM Digital Library (ACM)	Base de Dados
BioMed Central Journals	Base de Dados
BioOne (BioOne.Org)	Base de Dados
PubMed / Medline (NLM)	Base de Dados
Web of Science (ISI)	Base de Dados
Wiley Online Library Journals (Wiley)	Base de Dados

## REFERÊNCIAS

- [1] J . Biolchini, P. Mian, T. Conte, A. Natali e G. Travassos , “A Systematic Review Process for Software Engineering”, *ESELaw '05: 2nd Experimental Software Engineering Latin American Workshop*, 2005.
- [2] L. Bortolussi, A. Dovier e F. Fogolari, "Agent-based Protein Structure Prediction," *Multiagent and Grid Systems - Multi-agent systems for medicine, computational biology, and bioinformatic*, 2005.
- [3] L. Bortolussi, A. Dovier e F. Fogolari, "Multi-agent simulation of protein folding," in *International Workshop on Multi-Agent Systems for Medicine, Computational Biology, and Bioinformatics*, 2005.
- [4] Y. Duan and P. A. Kollman, "Computational protein folding: from lattice to all-atom," *IBM Systems Journal*, vol. 40, pp. 297-309, 2001.
- [5] A. M. Lesk, *Introduction to bioinformatics*, 3rd ed. Oxford ; New York: Oxford University Press, 2008
- [6] O. Pedreira, M. Piattini, M. R. Luaces e N. R. Brisaboa, "A systematic review of software process tailoring," *SIGSOFT Softw. Eng. Notes*, vol. 32, pp. 1-6, 2007.
- [7] M. Zvelebil and J. Baum, *Understanding Bioinformatics*: Garland Science, 2007.



## APÊNDICE B: Análise estereoquímica da proteína 1EDP.

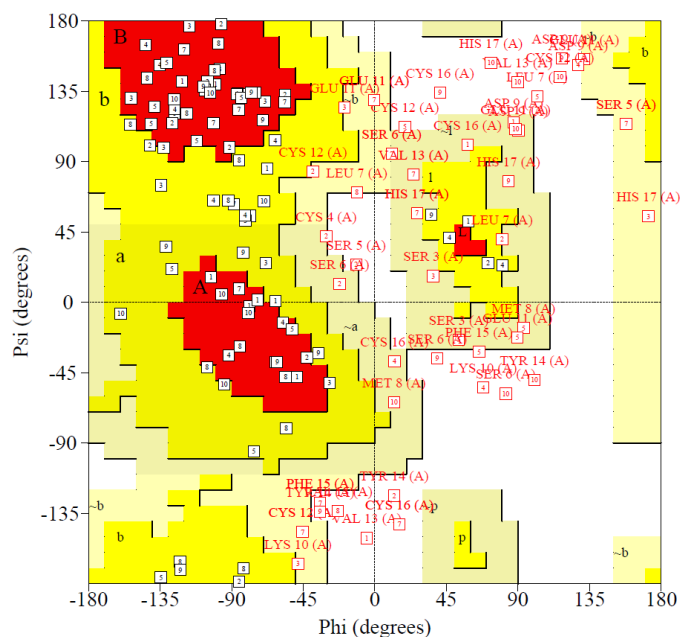


Figura 1 - Diagrama de Ramachandran contendo a análise estereoquímica das estruturas escolhidas como representantes das 10 diferentes simulações. Proteína de PDB ID: 1EDP

Tabela 1 – Proteína de PDB ID: 1EDP. Estatísticas obtidas pela análise dos diagramas de Ramachandran gerados por 10 simulações diferentes. Os números representam a quantidade de aminoácidos encontrados nas diferentes regiões do mapa de Ramachandran.

Região	Mais favorável	%	Adicionalmente permitida	%	Generosamente permitida	%	Não Permitidas	%
Simulação 1	8	53	3	20	3	20	1	7
Simulação 2	5	33	4	27	4	27	2	13
Simulação 3	6	40	5	33	4	27	0	0
Simulação 4	7	47	5	33	1	7	2	13
Simulação 5	6	40	3	20	3	20	3	20
Simulação 6	8	53	4	27	1	7	2	13
Simulação 7	6	40	1	7	6	40	2	13
Simulação 8	6	40	7	47	1	7	1	7
Simulação 9	5	33	5	33	1	7	4	27
Simulação 10	4	27	3	20	3	20	5	33