Pontifícia Universidade Católica do Rio Grande do Sul
Faculdade de Informática
Programa de Pós-Graduação em Ciência da Computação

Uma proposta para a predição computacional da estrutura 3D aproximada de polipeptídeos com redução do espaço conformacional utilizando análise de intervalos

Márcio Dorn

Dissertação apresentada como requisito parcial à obtenção do grau de mestre em Ciência da Computação

Orientador: Dr. Osmar Norberto de Souza

Porto Alegre
2008

Dados Internacionais de Catalogação na Publicação (CIP)

D713p Dorn, Márcio
Uma proposta para a predição computacional da estrutura 3D aproximada de polipeptídeos com redução do espaço conformacional utilizando análise de intervalos / Márcio Dorn. – Porto Alegre, 2008.
152 f.

Diss. (Mestrado) – Fac. de Informática, PUCRS.
Orientador: Prof. Dr. Osmar Norberto de Souza.

1. Bioinformática. 2. Computação Científica. 3. Proteínas – Estrutura 3D. I. Souza, Osmar Norberto de.

CDD 005.73

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

Pontifícia Universidade Católica do Rio Grande do Sul
FACULDADE DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

TERMO DE APRESENTAÇÃO DE DISSERTAÇÃO DE MESTRADO

Dissertação intitulada "**Uma Proposta para a Predição Computacional da Estrutura 3D Aproximada de Polipeptídeos com Redução do Espaço Conformacional Utilizando Análise de Intervalos**", apresentada por Márcio Dorn, como parte dos requisitos para obtenção do grau de Mestre em Ciência da Computação, Ciência da Computação, aprovada em 14/01/08 pela Comissão Examinadora:

Prof. Dr. Osmar Norberto de Souza – Orientador — PPGCC/PUCRS

Prof. Dr. Dalcidio Moraes Claudio — PPGCC/PUCRS

Prof. Dr. Carlos Augusto Prolo — FACIN/PUCRS

Homologada em 18/03/08, conforme Ata No. 05.08 pela Comissão Coordenadora.

Prof. Dr. Fernando Gehm Moraes
Coordenador.

Campus Central
Av. Ipiranga, 6681 – P32 – sala 507 – CEP: 90619-900
Fone: (51) 3320-3611 – Fax (51) 3320-3621
E-mail: ppgcc@inf.pucrs.br
www.pucrs.br/facin/pos

Agradecimentos

Primeiramente agradeço aos meus pais, Valdir e Marlise, a quem dedico este trabalho, por estarem presentes em todos os momentos e etapas da minha vida. Não há como mensurar o tamanho do carinho, amizade e apoio que vocês não só fizeram claramente transparecer nestes dois últimos anos, mas sim ao longo de toda a minha vida. Sem vocês nada seria possível. Vocês são um exemplo!. Aos meus avós Elga e Roberto por sempre estarem ao meu lado como um exemplo de pessoas. A minha madrinha Marlene por seu enorme carinho e amizade. Aos meus amigos, com os quais, sempre posso contar.

Ao professor Osmar pela orientação, pela dedicação, pela amizade, pelas palavras de apoio, pela confiança e pela compreensão. Agradeço a todos os professores do PPGCC, em especial ao professor Dalcídio pelo apoio, incentivo e por sempre estar de portas abertas para conversar. Por suas dicas e conselhos.

Aos integrantes do LABIO: Karina, André e Fúria pela amizade e pelo espírito de equipe. Agradeço também a toda a equipe do PPGCC/FACIN e ao CNPq, os quais, foram essenciais para o desenvolvimento deste trabalho.

“Das interessanteste an unseren Universum ist, dass man es verstehen kann.” *(Albert Einstein)*

Resumo

As proteínas são polipeptídeos formados por uma longa cadeia covalente de resíduos de aminoácidos que, em condições fisiológicas (ambiente nativo), adota uma topologia 3D única. Estas macromoléculas estão envolvidas na maior parte das transformações moleculares nas células vivas. A estrutura nativa dita a função bioquímica específica da proteína. Conhecer a estrutura 3D da proteína implica em também conhecer a sua função. Assim, conhecendo a sua estrutura é possível interferir ativando ou inibindo a sua função, como nas doenças onde os alvos dos fármacos são as proteínas. Experimentalmente, a estrutura 3D de uma proteína pode ser obtida através de técnicas de cristalografia por difração de raios X ou por ressonância magnética nuclear. Porém, devido às diversas dificuldades, incluindo o alto custo e o elevado tempo demandado por estas técnicas, a determinação da estrutura 3D de proteínas ainda é um problema que desafia os cientistas. Diversos métodos de predição *in sílico* foram criados durante os últimos anos buscando a solução deste problema. Estes métodos estão organizados em dois grandes grupos. Ao primeiro, pertencem os métodos de modelagem comparativa por homologia e métodos baseados em conhecimento como os de alinhavamento (*threading*). Ao segundo, pertencem os métodos *ab initio* e os métodos *de novo*. No entanto, estes métodos de predição possuem limitações: métodos baseados em modelagem comparativa por homologia e alinhavamento somente podem realizar a predição de estruturas que possuem seqüências idênticas ou similares à outras proteínas armazenadas no *Protein Data Bank* (PDB). Métodos *de novo* e *ab initio*, por sua vez, tornam possível a obtenção de novas formas de enovelamento. Entretanto, a complexidade e a grande dimensão do espaço de busca conformacional, mesmo para uma pequena molécula de proteína, torna o problema da predição intratável computacionalmente (*Paradoxo de Levinthal*). Apesar do relativo sucesso obtido por estes métodos para proteínas de pequeno tamanho, muitos esforços ainda são necessários para o desenvolvimento de estratégias para extração e manipulação de dados experimentais, bem como o desenvolvimento de metodologias que façam utilização destas informações com o propósito de predizer corretamente, a partir apenas da seqüência de aminoácidos de uma proteína, a sua estrutura 3D. Nesta dissertação é apresentada uma nova proposta para a predição *in sílico* da estrutura 3D aproximada de polipeptídeos e proteínas. Um novo algoritmo foi desenvolvido, baseando-se na análise de informações obtidas de moldes do PDB. Técnicas de mineração de dados, representação de intervalos e de manipulação das informações estruturais são utilizadas neste algoritmo. Os intervalos de variação angular de cada resíduo de aminoácido da cadeia polipeptídica são reduzidos como o objetivo de encontrar um intervalo fechado que contém a conformação com a menor energia potencial. Seis estudos de caso demonstram a aplicação do método desenvolvido.
Palavras-chave: Bioinformática, predição da estrutura 3D de proteínas, computação científica.

Abstract

Proteins are polypeptides formed by a long chain of amino acids residues which, in physiological conditions (native environment), adopt a unique three-dimensional (3-D) structure. These macromolecules are involved in most of the molecular transformations in the living cells. The native structure of a protein dictates its biochemical function. Hence, knowledge of a protein structure allows one to interfere with it, either by enhancing or inhibiting its function, such as in diseases in which the drug targets are proteins. Experimentally, the 3-D structure of a protein is obtained by techniques such as X-ray diffraction crystallography or nuclear magnetic resonance. However, due to the high cost and time demanded by these techniques, determination of the 3-D structure of a protein is a problem that still challenges the scientists. Many computational protein structure prediction methods have been proposed along the last years in order to address this problem. These methods are organized into two major groups. The first group comprehends comparative homology modelling and knowledge-based methods such as fold recognition via threading. The second group is made up by ab initio and de novo methods. However, these methods also have limitations: comparative homology modelling can only predict structures of proteins with amino acid sequences nearly identical or similar to other protein sequences of known structure in the protein Data Bank (PDB). Ab initio and de novo methods can predict new folds, but the complexity and high dimensionality of the search space, even for a small protein molecule, makes the problem computationally intractable. Despite the relative success of these prediction methods for small proteins and polypeptides, efforts are still needed to develop novel strategies for extracting and manipulating experimental data and to develop methods that use these data for correctly predicting a protein 3-D structure from its amino acid sequence only. In this dissertation we present a new computational method to predict approximate 3-D structure of polypeptides and proteins. A new algorithm was developed, based on information analysis obtained from PDB templates. Data mining techniques, intervals representation and treatment of experimental structural information are used in this algorithm. The polypeptide main chain torsion angles intervals, for each amino acid residue, are reduced with the objective to find a closed interval that contains the conformation with the lowest potential energy. Six case studies illustrate applications of the proposed method.

Keywords: Bioinformatics, three-dimensional protein structure prediction, scientific computation.

Lista de Figuras

Lista de Tabelas

Lista de Siglas

3D	Tridimensional	22
NMR	Ressonância Magnética Nuclear	22
PDB	*Protein Data Bank*	23
EMBL	Banco de Dados Europeu de Estruturas de Proteínas	38
PDBj	Banco de Dados Japonês de Estruturas de Proteínas	38
wwPDB	*Worldwide Protein Data Bank*	39
EP	Energia Potencial	41
DM	Dinâmica Molecular	41
LINUS	*Local Independently Nucleated Units of Structure*	41
BLAST	*Basic Local Alignment and Search Tool*	42
PSI-BLAST	*Position Specific Iterative - Basic Local Alignment and Search Tool*	42
CHARMM	*Chemistry at HARvard Molecular Mechanics*	46
AMBER	*Amber Molecular Dynamics Package*	46
CASP	*Critical Assessment of Techniques for Protein Structure Prediction*	49
RMSD	Desvio Médio Quadrático	58
XML	Linguagem Extensível de Formatação	62

Sumário

1 Introdução

1.1 Justificativa

As proteínas são macromoléculas envolvidas na maior parte das funções celulares. Estas funções podem ser reguladoras, de transporte, de catálise, entre outras. Uma proteína é constituída por uma seqüência de aminoácidos, comumente denominada estrutura primária, que forma uma cadeia polipeptídica por meio da polimerização representada por uma reação de condensação. A ligação CO-NH resultante, entre aminoácidos subseqüentes, é conhecida como ligação peptídica. A cadeia polipeptídica de uma proteína, em seu estado nativo, enovela-se assumindo uma conformação única. Esta conformação ou estrutura tridimensional (3D) determina a função que a proteína irá exercer na célula ou organismo [54].

Experimentalmente, a estrutura 3D de uma proteína pode ser obtida através de técnicas de cristalografia por difração de raios X ou por ressonância magnética nuclear (NMR, sigla em inglês). Porém, devido às diversas dificuldades, incluindo o alto custo e o elevado tempo demandado por estas técnicas, a determinação da estrutura 3D de proteínas ainda é um problema que desafia os cientistas. A dificuldade na determinação da estrutura 3D de proteínas gerou uma discrepância muito grande entre o volume de dados gerados por projetos genoma, os quais visam conhecer todos os genes de um determinado organismo, incluindo, mas não limitado às proteínas, e o número de estruturas 3D de proteínas que são conhecidas. Em 15 de outubro de 2007 havia aproximadamente 78 milhões de seqüências de proteínas no GenBank[1]. Dessas, aproximadamente seis milhões são consideradas únicas (não-redundantes ou NR). Por outro lado, no *Protein Data Bank* (PDB) [7], em 30 de outubro de 2007, havia 43.238 estruturas [2] 3D de proteínas. Como no caso das seqüências de proteínas, podemos eliminar redundância no PDB, filtrando proteínas cujos enovelamentos ou topologias são muito similares. Assim, temos apenas 1.056 topologias ou tipos de enovelamentos distintos [3]. Isto significa que apenas em torno de 0,02% das estruturas 3D de todas as seqüências únicas (NR) de proteínas são conhecidas. Esta discrepância tem motivado cientistas de áreas como ciência da computação, engenharias, física, química e matemática a construir modelos e desenvolver métodos que possam predizer e reproduzir a estrutura 3D destes polímeros a partir da sua seqüência de aminoácidos.

[1] GenBank: ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt
[2] Número de estruturas: http://www.rcsb.org/PDB/statistics/holdings.do
[3] Número de topologias: http://www.rcsb.org/PDB/statistics/contentGrowthChart.do?content=fold-scop

1.2 Motivação

A predição do enovelamento ou topologia [17] de uma proteína é atualmente um dos maiores problemas da Bioinformática Estrutural. O principal desafio é compreender como a informação codificada na seqüência linear de aminoácidos ou estrutura primária se traduz na estrutura 3D ou estrutura terciária de uma proteína e, a partir deste conhecimento, desenvolver metodologias computacionais que possam predizer, de forma correta, a estrutura nativa e funcional da proteína. Muitas metodologias e algoritmos foram propostos ao longo dos anos como solução a este complexo problema [12, 63, 68, 91]. Essas metodologias podem ser classificadas em dois grandes grupos. Ao primeiro grupo pertencem os métodos de modelagem comparativa por homologia e métodos baseados em conhecimento. Na modelagem comparativa, uma seqüência alvo é alinhada com a seqüência de uma proteína molde com estrutura 3D conhecida e armazenada no PDB. Quando homologia é detectada, geralmente com mais de 30% de identidade entre as seqüências alinhadas, em toda a sua extensão, a modelagem pode ocorrer realizando-se a cópia das coordenadas 3D do molde ou obtendo-se a média entre múltiplos moldes e substituindo esta na proteína-alvo. Pode-se ainda, utilizar distâncias inter-atômicas de regiões de alinhamento dos moldes como restrições para a modelagem [12, 47]. A modelagem comparativa por homologia é o método de predição com maior acurácia nos resultados finais [42, 58, 91]. Em contraste, métodos baseados em conhecimento (*knowledge-based*) utilizam potenciais estatísticos derivados da análise de padrões de enovelamento de proteínas com estruturas 3D conhecidas e armazenadas em uma base de dados [12]. A partir destas estatísticas, uma proteína alvo pode ser predita mesmo não havendo estruturas homólogas na base de dados [47]. Os métodos de reconhecimento de enovelamento estrutural via alinhavamento (*threading*) são os melhores exemplos desta técnica [42, 91]. Este método permite a detecção de homologia entre seqüências e estruturas de proteínas que não seria possível com métodos de alinhamento par a par de seqüências, como os utilizados na modelagem comparativa por homologia.

Ao segundo grupo pertencem os métodos *ab initio* e os métodos *de novo*, os quais possuem a capacidade de predizer novas formas de enovelamento para proteínas que não possuem homólogas no PDB. Os métodos *ab initio* são fundamentados na termodinâmica estatística e se baseiam no fato de que a estrutura nativa de uma proteína corresponde ao mínimo global de sua energia livre [91]. Esta metodologia simula o espaço conformacional da proteína utilizando uma função de energia potencial, a qual descreve a energia interna da proteína e suas interações com o meio no qual está inserida. O objetivo é encontrar um mínimo global de energia livre que corresponda ao estado nativo ou funcional da proteína [68, 91].

No entanto, estas metodologias de predição possuem limitações: métodos baseados em modelagem comparativa por homologia somente podem realizar a predição de estruturas que possuem seqüências quase idênticas ou similares em um banco de dados de estruturas conhecidas. Métodos *de novo* e métodos *ab initio*, por sua vez, tornam possível a obtenção de estruturas

cujas formas de enovelamento ainda não são conhecidas. Entretanto, a complexidade e a grande dimensão do espaço de busca conformacional [65], mesmo para uma pequena molécula polipeptídica ou de proteína, tornam o problema da predição intratável computacionalmente (*Paradoxo de Levinthal*) [44, 53].

A complexidade e a grandeza deste espaço de busca conformacional podem ser estimadas pelo tempo necessário para encontrar o estado nativo de enovelamento de uma proteína. Este tempo é obtido por meio do produto do número de conformações que uma cadeia polipeptídica pode assumir pelo tempo necessário para encontrar cada uma destas conformações. Este número pode ser estimado da seguinte maneira: se assumirmos que cada um dos aminoácidos em uma proteína pode assumir cinco configurações diferentes - este número é bastante inferior aos números reais - o número de conformações possíveis para uma proteína com 100 aminoácidos seria igual a $5^{100} = 10^{70}$. Se o tempo estimado para se calcular cada conformação diferente da cadeia polipeptídica for igual a 10^{-11} segundos (10 x 10^{-12} s = 10 picossegundos) então, o tempo estimado para se encontrar o estado nativo de uma proteína com 100 resíduos de aminoácidos seria de cerca de 10^{59} segundos, ou aproximadamente 10^{52} anos [44]. Considerando que a idade da Terra é de $4,6$ x 10^9 anos, é claramente impossível percorrer todo o espaço conformacional para esta proteína de 100 aminoácidos apenas (isso necessitaria 10^{43} vezes a idade da Terra!). Sabe-se, porém, que as proteínas se enovelam em tempos que vão de milissegundos a segundos [81]; este é o paradoxo de Levinthal [80]. No início da década de 1990 este paradoxo foi descrito na linguagem da complexidade computacional [65, 92] demonstrando que a busca aleatória do espaço conformacional das proteínas, como proposto por Levinthal [53], é um problema NP-completo. Como se pode notar, a complexidade de um problema desta natureza cresce à medida que aumenta o tamanho da seqüência da proteína. O tamanho médio de uma proteína é de 250 aminoácidos.

1.3 Objetivos e contribuições almejadas

Apesar do relativo sucesso na predição de novas formas de enovelamento para proteínas de pequeno tamanho (inferior a 100 aminoácidos) pelos métodos de predição *ab initio* ou *de novo*, ainda é necessário o desenvolvimento de estratégias para extração e manipulação de dados experimentais, bem como o desenvolvimento de metodologias que façam a utilização destas informações com o propósito de predizer corretamente, a partir da seqüência de aminoácidos de uma proteína, a sua estrutura 3D correspondente.

Sendo assim, o desenvolvimento de métodos computacionais de predição que reduzam o esforço computacional e que permitam a predição de maneira rápida e correta de novas formas de enovelamento se apresenta como um dos maiores desafios da Bioinformática Estrutural e da biologia molecular do século XXI. Neste trabalho é apresentada uma nova proposta para a predição *in sílico* da estrutura 3D de proteínas e polipeptídeos. Um novo algoritmo baseado na

análise de informações extraídas de fragmentos de proteínas com estrutura 3D conhecida busca a construção de estruturas 3D aproximadas de proteínas.

Dentre as principais objetivos e contribuições almejadas com a realização deste trabalho é possível citar:

- A proposta e desenvolvimento de um novo algoritmo para, a partir da seqüência de aminoácidos de uma proteína-alvo, produzir com o mínimo esforço computacional, estruturas 3D aproximadas de proteínas;
- O estudo e utilização de técnicas de agrupamento para manipular dados de proteínas com estrutura 3D conhecida experimentalmente;
- O estudo e desenvolvimento de formas de representação da estrutura 3D de proteínas na forma de intervalos de variação angular;
- O desenvolvimento de técnicas para redução de intervalos das estruturas 3D representadas na forma de intervalos de variação angular;
- A realização de experimentos utilizando o método desenvolvido na predição da estrutura 3D aproximada de diferentes classes de proteínas.

1.4 Organização do trabalho

Esta dissertação está organizada em seis capítulos. O primeiro, compreendendo este capítulo, apresenta a justificativa, a motivação e os objetivos com a realização deste trabalho de pesquisa. No Capítulo 2 é realizada uma breve introdução à Bioinformática Estrutural, aos conceitos básicos sobre proteínas, aminoácidos, ligações peptídicas, níveis hierárquico estruturais de proteínas e banco de dados de estruturas.

No Capítulo 3 são apresentadas as principais metodologias para predição *in sílico* da estrutura 3D de proteínas. No Capítulo 4 é apresentada uma nova proposta para predição *in sílico* da estrutura 3D aproximada de proteínas e polipeptídeos modelando e representando uma cadeia polipeptídica em termos de intervalos. No Capítulo 5 são apresentados os experimentos realizados utilizando o método de predição desenvolvido. Finalmente no Capítulo 6 são apresentadas as considerações finais sobre o trabalho realizado e as perspectivas para trabalhos futuros.

2 As proteínas

2.1 Introdução

As proteínas (polipeptídeos) são macromoléculas envolvidas na maior parte das transformações moleculares em uma célula viva. Quimicamente as proteínas são biopolímeros lineares formados por um alfabeto de 20 tipos diferentes de aminoácidos. A seqüência de aminoácidos, ou estrutura primária, forma, através de um processo de condensação, a cadeia polipeptídica da proteína. A ligação CO-NH resultante, uma ligação amida, é conhecida como ligação peptídica. Esta cadeia polipeptídica quando em condições fisiológicas (ambiente nativo), adota uma única estrutura tridimensional (3D) ou conformação nativa. Isto é, quando uma proteína é sintetizada ela se dobra para que parte da cadeia principal e da lateral, fundamentais para desempenhar a sua função, sejam postas em posição geométrica precisa. Esta dobra nativa, adotada pela cadeia polipeptídica, não sofre variação [29], sendo única para uma dada seqüência de aminoácidos. A estrutura nativa determina a função bioquímica específica da proteína na célula [6, 10], a qual pode ser de catálise, de ligação, de transporte, entre outras [10]. Conhecer a estrutura 3D da proteína implica em também conhecer a sua função. A partir deste conhecimento é possível influênciar, através do desenvolvimento de compostos químicos, fármacos ou drogas, a ação que a proteína exerce no organismo.

Experimentalmente, a estrutura 3D de uma proteína pode ser obtida através de técnicas de cristalografia por difração de raio X ou de ressonância magnética nuclear (NMR) [6, 91]. No entanto, o elevado custo e o alto grau de complexidade se apresentam como fatores negativos destas técnicas. Estes, motivaram a realização de novos estudos para desenvolvimento de técnicas que pudessem predizer de forma eficaz e eficiente a estrutura 3D de uma proteína.

O processo pelo qual uma seqüência de aminoácidos atinge sua conformação em estado nativo é chamado de enovelamento ou dobramento [10]. A seqüência de aminoácidos e o ambiente em que estes estão inseridos são os fatores que, durante o processo de enovelamento, fazem com que a proteína assuma determinada conformação. Experimentos realizados por Anfinsen [4] demonstraram que a molécula de uma proteína quando desnaturada [1] (estado desenovelado) por rompimento das condições em seu ambiente pode enovelar-se novamente em sua estrutura nativa quando as condições fisiológicas são restauradas. A partir desta constatação, assume-se que a seqüência de aminoácidos contém toda a informação necessária para determinar a estrutura

[1] Proteína desnaturada: quando aquecidas ou sujeitas a fortes ácidos ou bases, as proteínas perdem a sua estrutura terciária específica e podem formar coágulos insolúveis. A perda da estrutura resulta na perda da função.

nativa da proteína. Buscou-se então, o desenvolvimento de técnicas que pudessem para uma dada seqüência de aminoácidos, determinar a estrutura nativa e funcional da proteína correspondente. A computação, com seu crescimento, motivou a pesquisa e o desenvolvimento de algoritmos que podussem realizar tal tarefa. O problema da predição da estrutura 3D de uma proteína, somente a partir da seqüência de aminoácidos, tornou-se então um dos principais e maiores problemas ainda não resolvidos da biologia molecular estrutural [10] e da Bioinformática. As principais metodologias para predição *in sílico* da estrutura 3D de uma proteína são descritas em capítulo à parte (Capítulo 3).

Este capítulo, tem como principal objetivo, realizar uma breve descrição dos conceitos básicos sobre estrutura de proteínas, os quais, servirão de apoio para o entendimento dos capítulos seguintes. São apresentadas as proteínas, seus elementos constituíntes: os aminoácidos, seu processo de formação: a ligação polipeptídica, seus níveis estruturais: estrutura primária, estrutura secundária, estrutura terciária e estrutura quaternária, e os banco de dados que armazenam informações estruturais de proteínas (PDB).

2.2 Aminoácidos

Os aminoácidos são as unidades básicas que formam as proteínas. Um aminoácido é quimicamente composto por um átomo de carbono denominado Cα, o qual possui quatro diferentes ligantes: um grupo amino (-NH_2), um grupo carboxílico (-COOH), um átomo de hidrogênio (H) e um grupo orgânico R também chamado de cadeia lateral ou radical livre. A Figura 1 apresenta a estrutura química padrão de um aminoácido.

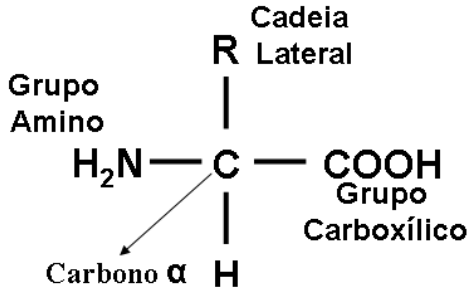

Figura 1 – Representação gráfica da estrutura química de um aminoácido: C (carbono α), COOH (grupo carboxílico), H_2N (grupo amino), H (átomo de hidrogênio), R (cadeia lateral ou radical livre).

Existem 20 tipos diferentes de aminoácidos que se diferenciam por suas cadeias laterais. Estas, podem possuir apenas alguns átomos ou anéis aromáticos complexos. O grupo R de cada aminoácido caracteriza as suas propriedades físico-químicas [91]. Os aminoácidos, por convenção internacional, são identificados por abreviações de três letras (derivadas do seu nome

em inglês) ou por um símbolo de uma letra [49]. A Tabela 1 relaciona os 20 aminoácidos existentes e seu respectivo código abreviado de três letras e o símbolo de uma letra.

Tabela 1: Relação dos 20 aminoácidos e seus respectivos códigos de três e de uma letra.

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

Em um cadeia polipeptídica[2], a região N-terminal é aquela que possuí um grupo amino livre e a região C-terminal é aquela que possuí na cadeia polipeptídica um grupo carboxílico livre. Peptídeos e polipeptídeos serão abordados nas seções seguintes.

Os aminoácido podem ser classificados pela natureza química de seus grupos R. Segundo Lehninger [49], os aminoácidos podem ser divididos em cinco classes: grupos R apolares e alifáticos (1), grupos R aromáticos (2), grupos R não-carregados e polares (3), grupos R carregados positivamente ou básicos (4) e grupos R carregados negativamente ou ácidos (5). Conhecer as propriedades fisico-químicas de cada aminoácido é fundamental para entender a bioquímica das proteínas. Estas propriedades são importantes, pois são elas que contribuem para que uma proteína encontre a sua estabilidade fisico-química representando o seu estado nativo.

[2] A cadeia polipeptídica é formada pela combinação linear de vários aminoácidos através de uma ligação peptídica.

2.2.1 Grupos R apolares e alifáticos

A primeira classe compreende os aminoácidos com cadeia lateral estritamente hidrofóbica e apolar, isto é, que não se dissolve em água. Esta classe engloba os aminoácidos: alanina (ala), valina (val), leucina (leu), isoleucina (ile), glicina (gly) e prolina (pro). A cadeia lateral destes aminoácidos contribuí para a estabilização da estrutura da proteína pela promoção de interações hidrofóbicas em seu interior [49]. A glicina (gly), apesar de ser um aminoácido apolar, não contribuí efetivamente para a existência de interações hidrofóbicas. O grupo amino secundário dos resíduos da prolina é mantido em uma conformação rígida que leva à redução da flexibilidade estrutural de regiões polipeptídicas em que este aminoácido ocorre [49]. A Figura 2 apresenta a estrutura química dos aminoácidos classificados como apolares e alifáticos.

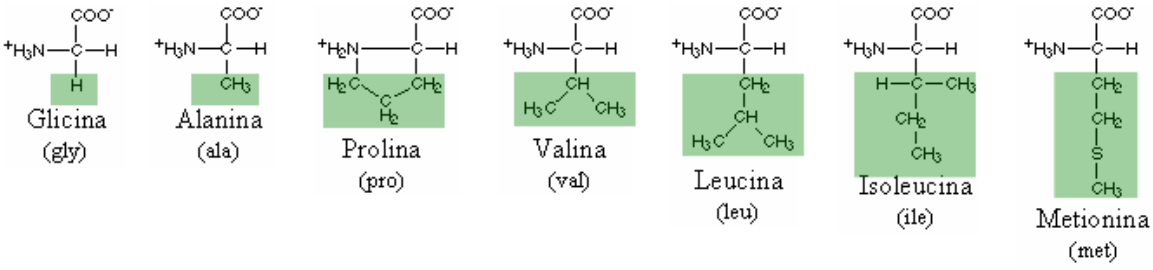

Figura 2 – Representação química dos aminoácidos apolares e alifáticos: o grupamento R dos aminoácidos alanina (ala), valina (val), leucina (leu), isoleucina (ile), metionina (met), glicina (gly) e prolina (pro) aparece destacado.

2.2.2 Grupos R aromáticos

A segunda classe de aminoácidos compreende os aminoácidos que podem participar de interações hidrofóbicas. São eles: a fenilamina (phe), a tirosina (tyr) e o triptofano (trp). Estes aminoácidos com suas cadeias laterais aromáticas são relativamente apolares ou hidrofóbicos [49]. A Figura 3 apresenta a estrutura química dos aminoácidos classificados como aromáticos.

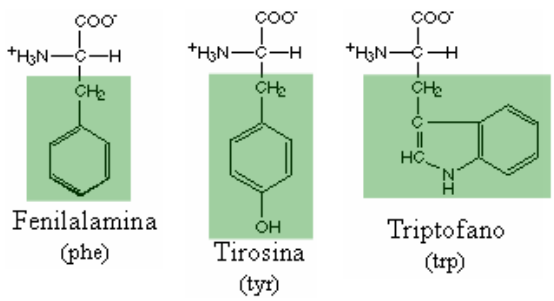

Figura 3 – Representação química dos aminoácidos aromáticos: o grupamento R dos aminoácido fenilamina (phe), tirosina (tyr) e triptofano (trp) aparece destacado.

2.2.3 Grupos R não carregados e polares

A terceira classe de aminoácidos abrange os aminoácidos não carregados, mas polares. Fazem parte deste grupos a serina (ser), a treonina (thr), a cisteína (cys), a asparagina (asn) e a glutamina (gln). Estes aminoácidos são mais solúveis em água do que os aminoácidos não-polares, isto porque contêm grupos funcionais que formam ligações de hidrogênio com a água [49]. A Figura 4 apresenta a estrutura química dos aminoácido com grupamento R não carregado e polar.

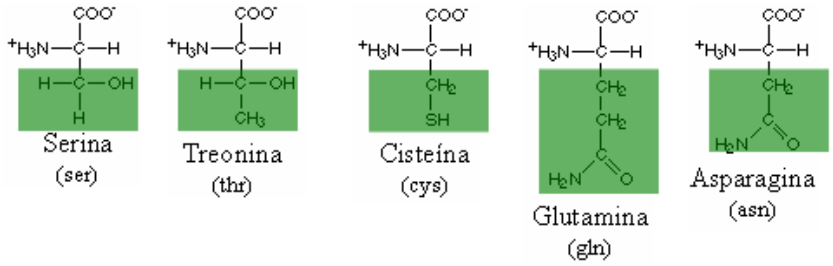

Figura 4 – Representação química dos aminoácidos não carregado e polares: o grupamento R dos aminoácidos serina (ser), treonina (thr), cisteína (cys), asparagina (asn) e glutamina (gln) aparece destado.

2.2.4 Grupos R carregados positivamente ou básicos

Os grupamentos R mais hidrofílicos são aqueles que são positivamente ou negativamente carregados. Os aminoácidos pertencentes a quarta classe são aqueles nos quais os grupos R têm uma carga positiva líquida em pH 7 [49]. Esta classe abrange os aminoácidos como a lisina (lis), a arginina (arg) e a histidina (his). A Figura 5 apresenta a estrutura química dos aminoácidos com grupamento R carregado positivamente.

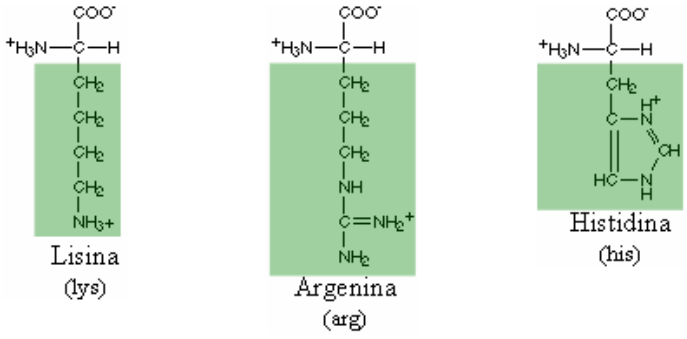

Figura 5 – Representação química dos aminoácidos básicos: o grupamento R dos aminoácidos lisina (lys), argenina (arg) e histidina (his) aparece destacado.

2.2.5 Grupos R carregados negativamente ou ácidos

Os aminoácidos que possuem o grupamento R com uma carga negativa em pH 7 pertencem à quinta classe. Fazem parte desta classe os aminoácidos conhecidos como ácido aspártico e ácido glutâmico. A Figura 6 apresenta a estrutura química dos aminoácidos com grupamento R carregado negativamente.

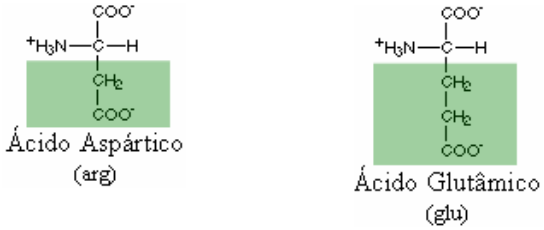

Figura 6 – Representação química dos aminoácidos carregados negativamente: a cor verde destaca o grupamento R dos aminoácidos ácido aspártico (asp), ácido glutâmico (glu).

2.3 Ligações peptídicas e polipeptídeos

Os aminoácidos, durante a síntese das proteínas, se ligam covalentemente de forma seqüencial, formando um polímero ou cadeias polipeptídicas. Esta ligação recebe o nome de ligação peptídica e se forma entre o átomo de carbono (C) do grupo carboxílico de um aminoácido e o átomo de nitrogênio (N) do grupo amina de outro aminoácido. Os elementos que compõem a água são removidos como um co-produto da reação. A água (H_2O) se forma a partir do -OH do grupo carboxila de um dos aminoácidos e de um átomo de H do grupo -NH_2 do outro aminoácido. A Figura 7 esquematiza a formação de uma ligação peptídica entre dois resíduos de aminoácidos [3].

[3]Resíduo de aminoácido: devido à desidratação ocorrida com a ligação de dois aminoácidos, estes passam a ser chamados de resíduos (sobras) de aminoácidos.

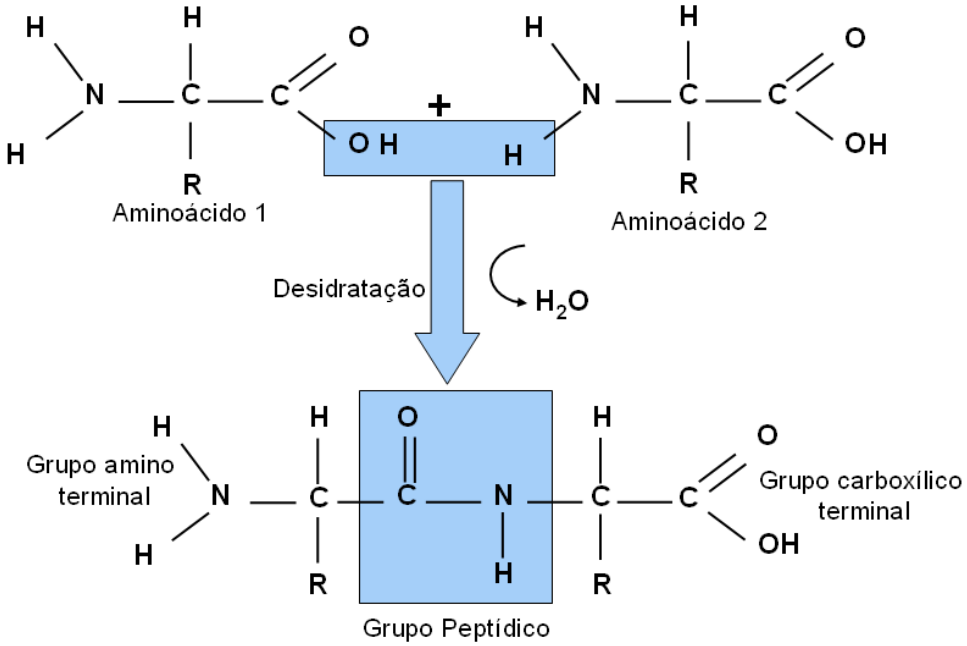

Figura 7 – Ligação peptídica: representação esquemática do processo de formação de uma ligação peptídica entre dois resíduos de aminoácidos.

2.3.1 Propriedades das ligações peptídicas: torções da cadeia principal

Quando ocorre a ligação de muitos aminoácidos, o polímero resultante recebe o nome de polipeptídeo. A repetição do conjunto -N-C_α-C- em uma proteína é chamado de cadeia principal da proteína (ou *backbone* em inglês). Este padrão se repete ao longo da cadeia polipeptídica sem sofrer alterações. A direção da cadeia polipeptídica é determinada a partir do grupo amino terminal (grupo N-terminal) até o grupo carboxila terminal (grupo C-terminal) em um polipeptídeo.

A ligação C-N tem um caráter parcial de dupla ligação, com o átomo de N alcançando uma carga positiva parcial e o O uma carga negativa parcial, não permitindo que a molécula normalmente gire sobre esta ligação. A rotação somente é permitida sobre as ligações N-C_α e C_α-C. Desta forma o esqueleto da cadeia polipeptídica pode ser representando como uma série de planos rígidos com planos consecutivos compartilhando um ponto em comum de rotação em C_α. Esta ordenação planar e rígida é o resultado da estabilização por ressonância da ligação peptídica. Isso impõem restrições importantes no número de conformações que uma proteína pode adotar. A Figura 8 representa os planos sucessivos da cadeia principal de um polipeptídeo.

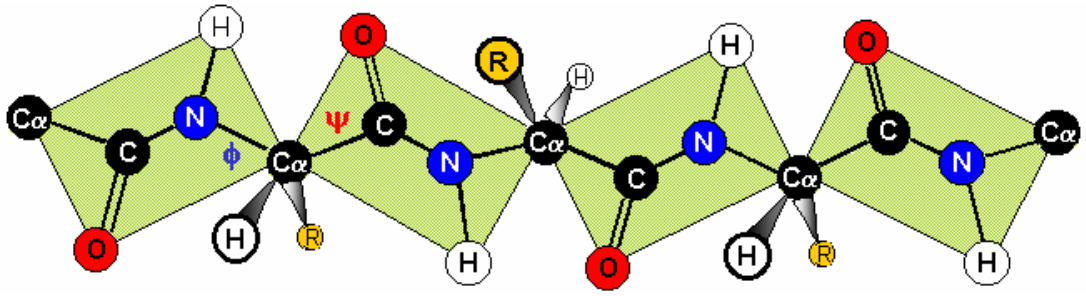

Figura 8 – Estrutura planar de um polipeptídeo: representação gráfica do esqueleto do polipeptídeo representado como uma série de planos sucessivos (Figura adaptada de [49]).

A rigidez e a planaridade de uma ligação peptídica é representada pelo ângulo diedro ω (ômega). A rotação em torno da ligação entre o carbono (C) da carboxila e o nitrogênio (N) da amina define o ângulo omega (ω). Este ângulo não é livre para rotacionar. Devido a esta restrição, o ângulo ω tem suas liberdades de torção variando próximas a 0°(cis) ou a 180°(trans). A ligação entre o N e o C_α forma o ângulo ϕ e a ligação entre o C_α e o C da carboxíla forma o ângulo ψ. São os ângulos ϕ e ψ os maiores responsáveis pelas torções da cadeia principal de uma proteína.

Quando os ângulos ϕ e ψ são fixados em 180°, o polipeptídeo adota uma conformação totalmente extendida [49]. A conformação de uma proteína pode ser representada unicamente pelos seus ângulos de torção (ϕ, ψ e ω) da cadeia principal. A Figura 9 representa a cadeia principal de um polipeptídeo destacando os seus ângulos de torção.

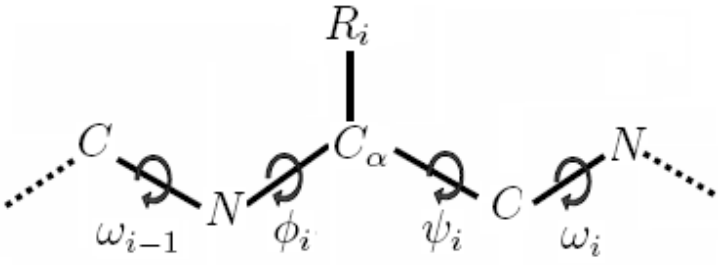

Figura 9 – Representação esquemática de um modelo de peptídeo identificando os ângulos de torção ϕ, ψ e ω da cadeia principal.

Os ângulos diedros ϕ, ψ e ω formam um tripleto de ângulos de torção, porém o dupleto formado pelos ângulos ϕ e ψ é o que contribuí efetivamente, devido às suas liberdades, para as conformaçãoes da cadeia principal.

2.3.2 Ângulos de torção da cadeia lateral

As cadeias laterais também possuem ângulos de torção. Os ângulos de torção χ das cadeias laterais ocorrem em número diferente e dependem exclusivamente do tipo de resíduo de aminoácido presente no polipeptídeo. Estes ângulos não interferem drasticamente na conformação assumida pela cadeia principal do polipeptídeo, porém contribuem para a estabilidade da molécula e a formação de ligações de hidrogênio[4]. Cadeias laterais com ângulos de torção χ incorretos podem causar choques estereoquímicos entre átomos da cadeia principal e átomos da cadeia lateral ou mesmo entre átomos de cadeias laterais vizinhas. E isto, pode contribuir para que a proteína assuma uma conformação incorreta.

A Tabela 2 apresenta o número de ângulos de torção χ de cada um dos 20 resíduos de aminoácidos existentes.

Tabela 2: Número de ângulos χ presente em cada resíduo de aminoácido.

Resíduos	Número de ângulos χ
GLY, ALA, PRO	Cadeia principal
SER, CYS, THR, VAL	χ_1
ILE, LEU, ASP, ASN, HIS, PHE, TYR, TRP	χ_1, χ_2
MET, GLU, GLN	χ_1, χ_2, χ_3
LYS, ARG	$\chi_1, \chi_2, \chi_3, \chi_4$

2.3.3 Mapa de Sasisekharan-Ramakrishnan-Ramachandran

Os ângulos ϕ e ψ podem ter qualquer valor entre -180°e +180°, porém, muitas combinações de ϕ e ψ são proibidas por interferências estéricas entre átomos no esqueleto principal da cadeia polipeptídica e entre átomos da cadeia lateral dos aminoácidos. Os valores permitidos e proibidos para os ângulos de torção ϕ e ψ são graficamente demonstrados pelo mapa de Sasisekharan-Ramakrishnan-Ramachandran [74], ou simplesmente mapa de Ramachandran.

A Figura 10 apresenta o mapa de Ramachandran, destacando as regiões permissíveis para a combinação dos ângulos ϕ e ψ. Conforme será discutido nas seções seguintes, as regiões no mapa de Ramachandran representam, em termos de enovelamento, padrões de torção da cadeia polipeptídica (folhas β, hélices α).

[4]Ligações de hidrogênio: surgem quando dois grupos polares interagem. Os dois grupos polares devem ser de tipos específicos, isto é, um deve ser doador de próton e outro deve ser aceptor de próton.

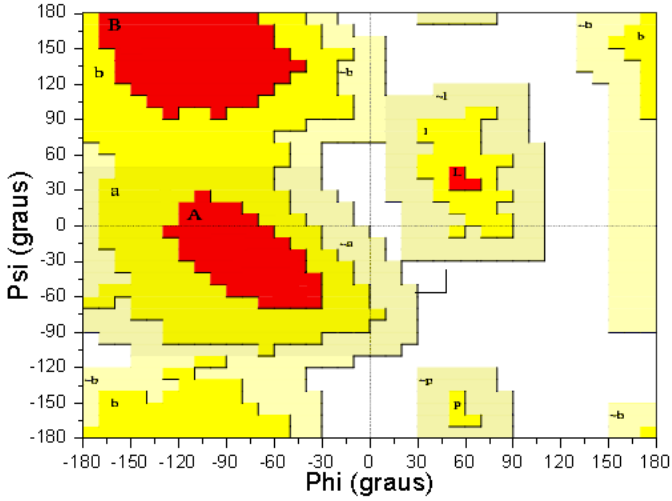

Figura 10 – Mapa de Ramachandran: a região mais favorável é apresentada em vermelho, a região permitida é apresentada em amarelo, a região ainda aceitável é apresentada em amarelo claro e a região não permitida em branco. A região em vermelho no canto superior esquerdo representa a região de folhas β paralelas e anti-paralelas. A região em vermelho no centro esquerdo, e no centro direito representam a região de hélices α a direita e a esquerda respectivamente (modelo adotado por Thornton e colaboradores [48]).

2.4 Níveis de organização estrutural

Para facilitar a tarefa de descrever e entender a estrutura de uma proteína, a mesma é estudada em 4 níveis hierárquico estruturais: a estrutura primária, a estrutura secundária, a estrutura terciária e a estrutura quaternária [49]. Cada um destes níveis estruturais é apresentado e discutido nas próximas seções.

2.4.1 Estrutura primária

A estrutura primária de uma proteína é descrita por sua seqüência linear de resíduos de aminoácidos. Cada aminoácido se liga à outro aminoácido através de uma ligação peptídica (Seção 2.3). O início da estrutura primária de uma proteína corresponde a sua região N-terminal e o final da estrutura primária é determinada pela região C-terminal (Figura 7).

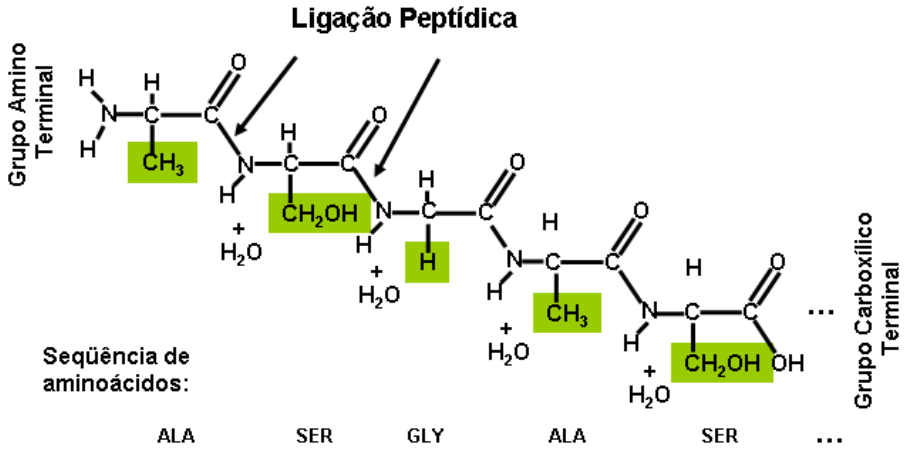

Figura 11 – Representação esquemática da estrutura primária de uma proteína. As regiões destacadas representam a cadeia lateral do polipeptídeo.

2.4.2 Estrutura secundária

Apesar das proteínas serem polímeros lineares, as estruturas das mesmas não são aleatórias, ou seja, elas apresentam em alguns casos certa regularidade. São arranjos particularmente estáveis de resíduos de aminoácidos formando padrões estruturais ou regulares. São estas conformações regulares do polipeptídeo que representam a estrutura secundária da proteína. A regularidade na conformação espacial é mantida graças às interações intermoleculares (ligações de hidrogênio) entre os hidrogênios dos grupos amino e os átomos de oxigênio dos grupos carboxílicos de outros aminoácidos. Alguns tipos de estruturas secundárias são particularmente estáveis e de alta freqüência em proteínas. As conformações mais proeminentes são as hélices α e as folhas β. Além destas, há também, estruturas irregulares tal como, a volta e a alça. Estas estruturas irregulares, são estruturas aleatórias que tem a função de fazer a conexão entre as estruturas secundárias regulares.

A seguir são descritas em maiores detalhes cada uma das estruturas secundárias:

Hélice α: em uma cadeia polipeptídica, o grupo NH da cadeia principal pode formar uma ligação de hidrogênio com o grupo CO do quarto aminoácido mais próximo, a repetição i$\rightarrow$ i+4 define um padrão regular conhecido como hélice α [50, 71]. As hélices α possuem em média 3,6 resíduos de aminoácidos por volta. A estrutura é estabelizada por uma ligação de hidrogênio ligando o átomo de nitrogênio de uma ligação peptídica e o átomo de oxigênio da carbonila do quarto aminoácido da região N-terminal, daquela ligação peptídica [49]. Cada volta sucessiva de hélice α é presa às voltas adjacentes por três ou quatro ligações de hidrogênio. São estas ligações de hidrogênio, que quando combinadas, garantem a estabilidade da estrutura helicoidal (Figura 12B).

Os resíduos de aminoácidos da estrutura regular em forma de hélice α têm seus ângulos diêdros (ϕ e ψ) variando no mapa de Ramachandran em torno de -30°a -120°para ϕ e -60°a 20°para ψ (Figura 10). A Figura 12A apresenta uma hélice α representada pelo arranjamento dos átomos de sua cadeia principal. A Figura 12B apresenta uma hélice α representada através dos átomos de sua cadeia principal juntamente com a sua representação na forma de *ribbon* [5].

Diferentes seqüências de aminoácidos têm propriedades diferentes para formar hélices α. Estas propriedades resultam do formato de sua cadeia lateral [10]. Os resíduos de aminoácidos como a metionina, a alanina, a leucina, a glutamina e a lisina são mais propícias a formar hélices α. Os resíduos de aminoácidos glicina, serina, prolina e tirosina geralmente não participam da formação de hélices α [10]. A glicina tende a interromper a regularidade presente em uma estrutura em forma de hélice α devido a sua grande flexibilidade conformacional. A prolina tende a quebrar ou torcer helices α devido ao fato de não poder formar uma ligação de hidrogênio e por interferência estérica da sua cadeia lateral [49]. O número de aminoácidos presentes em uma hélice α é bastante variável e pode estar no intervalo de 5 à 40 resíduos de aminoácidos, sendo mais freqüente hélices α com 10 resíduos de aminoácidos [71].

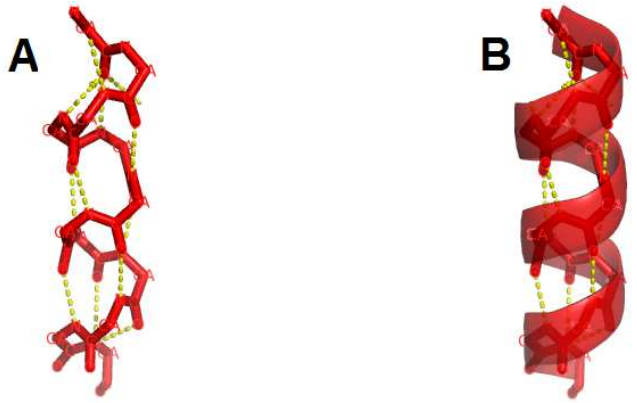

Figura 12 – Representação gráfica da estrutura secundária regular do tipo hélice α: (A) uma hélice α vista através dos átomos de sua cadeia principal, (B) uma hélice α vista através dos átomos de sua cadeia principal e através de sua representação do tipo *ribbon*. As linhas tracejadas representam as ligações de hidrogênio.

Folha β: no segundo tipo de estrutura secundária, o esqueleto da cadeia polipeptídica é estendido em ziguezague em vez de uma estrutura helicoidal como nas hélices α. Quando as cadeias polipeptídicas estão arranjadas lado a lado formam uma estrutura que se assemelha a uma série de fitas, trata-se da folha β [49]. Neste arranjo, as ligações de hidrogênio são formadas entre segmentos adjacentes da cadeia polipeptídica. No mapa de Ramachandran, as estruturas secundárias em forma de folha β ocupam a região que varia de -180°a -45°para ϕ e 45°a 225°para ψ (Figura 10).

As cadeias polipeptídicas adjacentes em uma folha β podem ser paralelas ou antiparalelas.

[5] Representação na forma de *Ribbon*: a proteína é representada como uma superfície lisa e densa.

Nas folhas paralelas o sentido da folha é da região N-terminal para a região C-terminal. Nas folhas antiparalelas, o sentido da folha é da região C-terminal para a região N-terminal. Os padrões das ligações de hidrogênio das folhas paralelas e antiparalelas são diferentes [70]. A Figura 13A apresenta uma folha β paralela através dos átomos de sua cadeia principal. A Figura 13B mostra a representação do tipo *ribbon* de uma folha β paralela. A Figura 14A apresenta uma folha β antiparalela através dos átomos de sua cadeia principal. A Figura 14B mostra a representação do tipo *ribbon* de uma folha $beta$ antiparalela.

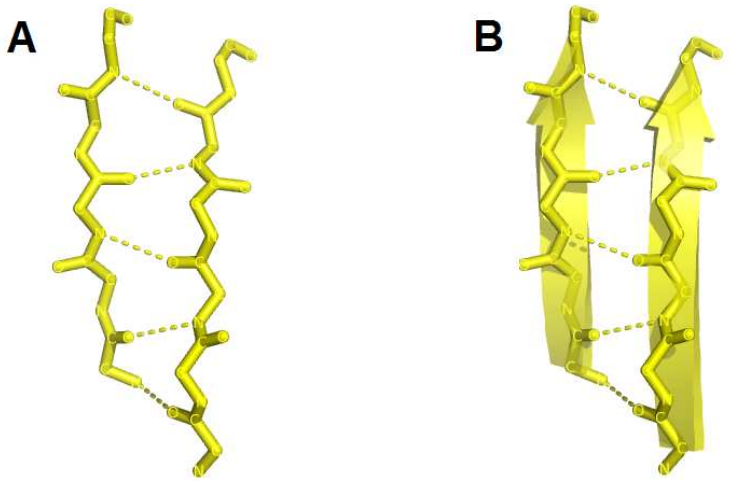

Figura 13 – Representação gráfica da estrutura secundária regular do tipo folha β paralela: (A) uma folha β paralela através dos átomos de sua cadeia principal, (B) uma folha β paralela através dos átomos de sua cadeia principal juntamente com a sua representação do tipo *ribbon*. As linhas tracejadas representam as ligações de hidrogênio.

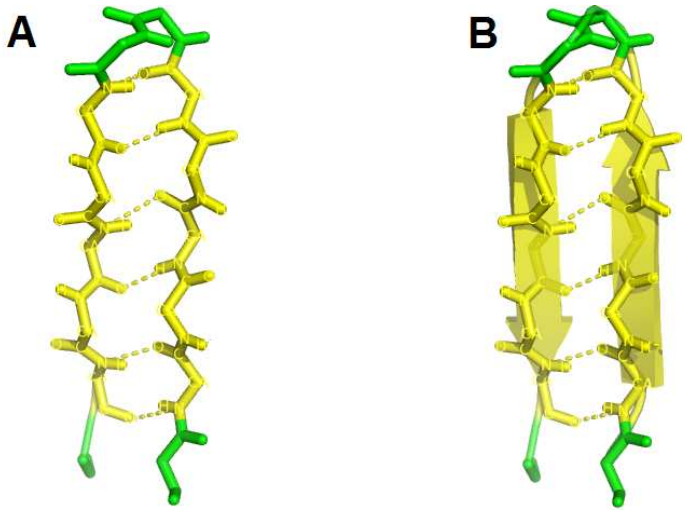

Figura 14 – Representação gráfica da estrutura secundária regular do tipo folha β antiparalela: (A) uma folha β antiparalela através dos átomos de sua cadeia principal, (B) uma folha β antiparalela através dos átomos de sua cadeia principal juntamente com a sua representação do tipo *ribbon*. As linhas tracejadas representam as ligações de hidrogênio.

Voltas e alças: O terceiro tipo de estrutura trata de uma estrutura secundária irregular, denominada por volta ou dobras. As voltas são formadas em regiões onde o polipeptídeo muda a sua direção, ou seja, após uma estrutura secundária regular em forma de hélice α e folhas β. As voltas são os elementos estruturais que unem sucessivas estruturas secundárias regulares. A Figura 15A, apresenta uma volta. Esta volta tem um tamanho maior quando comparada com a alça da Figura 15B.

Por serem estruturas irregulares não existe uma região específica para voltas no mapa de Ramachandran. A combinação de ângulos ϕ e ψ podem ocupar qualquer região neste mapa de Ramachandran, isto incluí regiões de folhas β, de hélices α. Devido a esta particularidade, as voltas são difíceis de serem preditas por métodos computacionais (conforme será discutido nos capítulos seguintes).

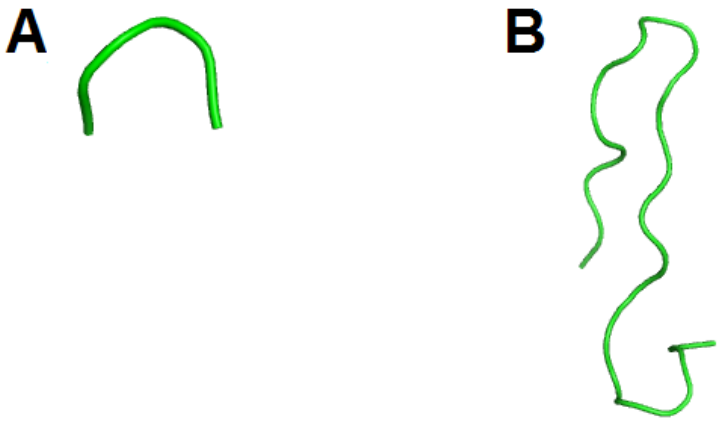

Figura 15 – Representação gráfica de estruturas secundárias irregulares e aleatórias: (A) representa uma volta, (B) representa uma alça.

2.4.3 Estrutura terciária

A estrutura terciária de uma proteína é representada pela distribuição de suas estruturas secundárias no espaço 3D. A forma tridimensional assumida pela proteína é também chamada de estrutura nativa da proteína ou estrutura funcional. A estrutura nativa da proteína tem como principal fator de formação a variação dos fatores termodinâmicos, ou seja, interações covalentes, ligações de hidrogênio, interações hidrofóbicas, interações hidrofílicas, interações eletrostáticas, forças de van der Walls e forças repulsivas [29].

Através da estrutura terciária de uma proteína é possível analisar ou prever a função que a mesma exerce no organismo. É possível, através de seu estudo, identificar o sítio ativo de enzimas, sítios de ligação em um receptor, ou um local de recombinação para a ação de outra proteína [49].

A Figura 16A, apresenta a estrutura terciária da Crambina[6] [90] (código PDB: 1CRN), composta por 2 estruturas secundárias em forma de hélice α e duas estruturas em forma de fitas β formando uma folha β atiparalela, conectadas por estruturas irregulares do tipo volta [64]. A Figura 16B apresenta a estrutura terciária de um peptídeo de escorpião [9] (código PDB: 1ACW), composta por uma estrutura regular em forma de hélice α e uma folha β atiparalela, conectadas por estruturas irregulares do tipo volta [64].

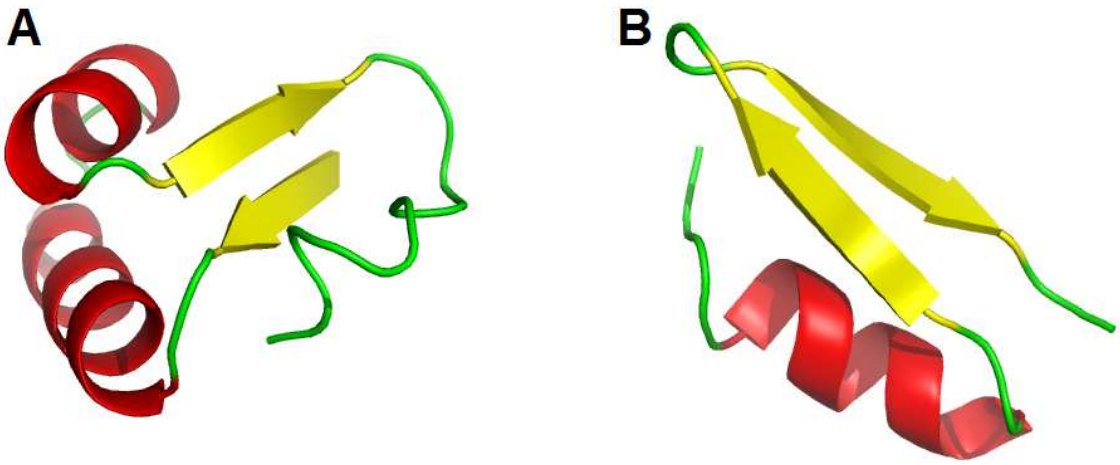

Figura 16 – Representação do tipo *Ribbon* de estruturas terciárias. (A) apresenta a estrutura terciária da proteína Crambina, cujo código PDB é 1CRN; (B) apresenta a estrutura terciária do peptídeo de escorpião, cujo código PDB é 1ACW. As pontes de sulfeto formadas entre os resíduos de aminoácido CIS10 e CIS26, CIS06 e CIS24 e CIS3 e CIS19 da proteína Crambina e as cadeias laterais e a cadeia principal de ambas as proteínas não são mostradas para facilitar a visualização.

A estrutura terciária de uma proteína está relacionada com a sua topologia (ou enovelamento). A topologia de uma proteína é dada pelo tipo de sucessão de estruturas secundárias que estão conectadas e a partir da forma na qual estas estruturas estão organizadas no espaço 3D. A Seção 2.5 apresenta as diferentes classes de topologias.

2.4.4 Estrutura quaternária

Uma proteína pode apresentar diversas cadeias (ou subunidades) polipeptídicas formando uma estrutura quaternária. A estrutura quaternária de uma proteína é o arranjamento de várias estruturas terciárias. Esta estrutura é mantida pelas mesmas forças que determinam as estruturas secundárias e terciárias (ligações de hidrogênio, interações hidrofóbicas, interações hidrofílicas). A Figura 27 apresenta a estrutura quaternária da Hemoglobina [45] (código PDB: 1A00).

[6]Crambina: pequena proteína globular presente em sementes de plantas.

A estrutura quaternária da Hemoglobina apresenta 4 cadeias: A, B, C e D. Cada uma destas subunidades é uma estrutura em nível terciário.

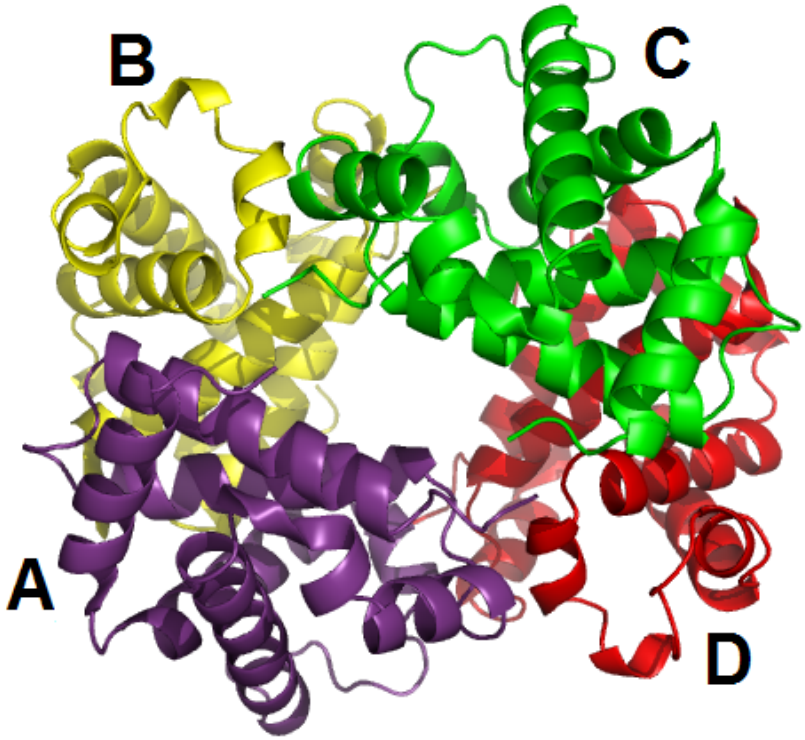

Figura 17 – Representação do tipo *Ribbon* da estrutura quaternária da Hemoglobina (Codigo PDB: 1A00 [45]) sem o grupo Heme, identificando as suas quatro subunidades. Roxa: subunidade A da hemoglobina, amarelo: subunidade B da hemoglobina, verde: subunidade C da hemoglobina, vermelho: subunidade D da hemoglobina.

2.5 Classificação de estruturas de proteínas

As proteínas podem ser agrupadas e classificadas segundo seus padrões de enovelamento [51]. Entre proteínas com padrões de enovelamento similar, existem famílias que compartilham características em suas estruturas, seqüências e funções [51]. A classificação mais geral de famílias de estruturas de proteínas é baseada nas suas estruturas secundárias e terciárias:

- Hélice α: estrutura secundária composta totalmente ou em sua maioria por hélices α;
- Folha β: estrutura secundária composta totalmente ou em sua maioria por folhas β;
- α + β: hélices α e folhas β separadas em partes diferentes da molécula; ausência da estrutura supersecundária β-α-β;
- α/β: hélices e folhas dispostas a partir de unidades β-α-β;

- α/β linear: uma linha passando pelos centros das fitas das folha é aproximadamente linear;
- Barril α/β: uma linha passando pelos centros das fitas da folha é aproximadamente circular.

Existem diversos bancos de dados que realizam a classificação e o agrupamento das estruturas de proteínas. São exemplos: o SCOP [64], o CATH [67] e o FSSP/DDD [36].

2.6 Determinação experimental da estrutura 3D das proteínas

A estrutura 3D de uma proteína pode ser obtida, experimentalmente, através de técnicas de cristalografia por difração de raio X ou por ressonância magnética nuclear (NMR). A cristalografia por difração de raio X é o mais antigo e mais preciso método para determinação da estrutura de uma proteína. A técnica permite determinar a estrutura 3D de proteínas, não existindo limites para o tamanho das moléculas em estudo, porém, as amostras (cristais) sofrem com danos causados pela radiação aplicada, não podendo ser analisada a dinâmica das interações entre proteínas, substratos e solventes [6]. A ressonância magnética nuclear, por sua vez, é uma técnica mais nova, apresentando vantagens referentes a possibilidade de estudo da estrutura e da dinâmica da molécula em estado líquido ou em um ambiente fisiológico. A principal desvantagem dos métodos experimentais para a determinação da estrutura 3D de proteínas está relacionada ao alto custo dos experimentos e ao elevado grau de complexidade. Isto, motivou ciêntistas da computação, físicos, biólogos e matemáticos a trabalharem no desenvolvimento de novas metodologias que pudessem predizer de forma correta a estrutura terciária de uma proteína, dada unicamente a sua seqüência de aminoácidos. No Capítulo 3 são abordadas as principais metodologias para a predição *in sílico* da estrutura 3D de polipeptídeos.

2.7 Banco de dados de estruturas de proteínas

Os bancos de dados estruturais são a base da Bioinformática Estrutural, pois fornecem os ingredientes para a predição, a análise e o estudo da estrutura de biomoléculas. O mais conhecido banco de dados de estruturas 3D de proteínas é o *Protein Data Bank* (PDB) [8], o qual tem como principal propósito armazenar, organizar e distribuir estruturas de macromoléculas. O PDB não é o único banco de dados de informações estruturais de proteínas, existem ainda outras bases de dados que organizam e distribuem este tipo de informações. São exemplos destas bases de dados: EMBL (banco de dados de estruturas de proteínas localizado na europa) [93] e PDBj (banco de dados de estruturas localizado no japão). Esforços entre pesquisadores do

PDB, EMBL e PDBj estão sendo realizados para criar um banco de dados único de estruturas 3D para a comunidade de pesquisadores, trata-se do wwPDB [7].

A possibilidade de acesso a informações sobre a estrutura de proteínas é uma peça chave nos métodos de predição de estruturas 3D de proteínas. Estes utilizam as informações destas bases para busca de padrões, construções de modelos e validação das estruturas 3D preditas.

2.7.1 Biliotecas de rotâmeros

As bibliotecas de rotâmeros são conjuntos de ângulos diedros usados especificamente para a otimização das posições dos ângulos diedros das cadeias laterais. Uma biblioteca de rotâmeros é constituída a partir de informações das torções de estruturas semelhantes, obtidas experimentalmente. Devido a restrições estéricas nas torções dos ângulos χ da cadeia lateral, as cadeias laterais dos aminoácidos de um polipeptídeo assumem apenas algumas conformações [29].

As bibliotecas de rotâmeros podem ser dependentes ou independentes da estrutura primária da proteína. As bibliotecas independentes classificam todos os valores de ângulos χ de um determinado aminoácido sem considerar se o mesmo ocorre em uma estrutura regular de hélice α ou folha β. Em contrapartida as bibliotecas dependentes classificam os aminoácidos de acordo com a sua presença em estruturas secundárias regulares. A biblioteca mais comum e mais utilizada é a biblioteca de rotâmeros de Dunbrack [22].

2.8 Resumo do capítulo

Neste Capítulo foram apresentados os conceitos básicos da Bioinformática Estrutural. De forma mais detalhadas foram discutidas as proteínas, as suas unidades básicas (os aminoácidos), a formação de cadeias polipetídicas e os níveis hierárquico estruturais em que as proteínas são estudadas. Foram também apresentados os banco de dados de estruturas de proteínas e as biliotecas de rotâmeros.

Este capítulo serve como base para a compreensão dos demais tópicos desta Dissertação. No capítulo 3 são abordadas as principais metodologias para predição *in sílico* da estrutura 3D de proteínas.

3 Predição *in sílico* da estrutura tridimensional de proteínas

3.1 Introdução

A predição da estrutura tridimensional (3D) de proteínas, baseada unicamente em sua seqüência de aminoácidos, é um problema que vem, ao longo dos últimos 40 anos, desafiando cientistas da computação, engenheiros, matemáticos e biólogos [6]. A predição do enovelamento [17] de uma proteína é um dos maiores problemas na Bioinformática Estrutural. O principal desafio é compreender como a informação codificada na seqüência linear de aminoácidos (estrutura primária) traduz-se na estrutura 3D (estrutura terciária) de uma proteína, e a partir deste conhecimento, desenvolver metodologias computacionais que possam predizer, de forma correta, a estrutura nativa da proteína.

Muitas metodologias e algoritmos foram propostos ao longo dos anos como solução para este problema complexo [12, 63, 68, 91]. Essas metodologias podem ser classificadas em dois grandes grupos. Ao primeiro grupo pertencem os métodos de modelagem comparativa por homologia e métodos baseados em conhecimento [58]. Ao segundo grupo pertencem os métodos *ab initio* e os métodos *de novo* [76,86], os quais possuem a capacidade de predizer novas formas de enovelamento para proteínas que não possuem homólogas no PDB.

Na modelagem comparativa, uma seqüência alvo é alinhada contra a seqüência de uma proteína molde com estrutura 3D conhecida e armazenada no PDB [8]. Quando a homologia é detectada, geralmente com mais de 30% de identidade, a modelagem pode proceder realizando a cópia das coordenadas 3D de um molde ou obtendo a média entre múltiplos moldes e substituíndo estas na proteína-alvo. Ainda, é possível utilizar distâncias inter-atômicas de regiões de alinhamento dos moldes como restrições para a modelagem [12, 47]. A modelagem comparativa por homologia de seqüência se apresenta como o método de predição com maior acurácia nos resultados finais [42, 58, 91]. Em contraste, métodos baseados em conhecimento (*knowledge-based*) utilizam potenciais estatísticos derivados da análise de padrões de enovelamento de proteínas com estruturas 3D conhecidas e armazenadas em uma base de dados [12]. A partir destas estatísticas, uma proteína-alvo pode ser predita mesmo não havendo estruturas homólogas na base de dados [47]. Os métodos de reconhecimento estrutural via alinhavamento (*threading*) são os melhores exemplos desta técnica [42, 91]. Estes métodos permitem a detecção de homologia entre seqüências e estruturas de proteínas que não seria possível com métodos de alinhamento par a par de seqüências, como os utilizados na modelagem comparativa por homologia.

Os métodos *ab initio* são fundamentados na termodinâmica e baseiam-se no fato de que a estrutura nativa de uma proteína corresponde ao mínimo global de sua energia livre [91]. Esta metodologia simula o espaço conformacional da proteína utilizando uma função de energia potencial (EP), a qual descreve a energia interna da proteína e suas interações com o meio em que está inserida. O objetivo é encontrar um mínimo global de energia livre que corresponda ao estado nativo ou funcional da proteína [68, 91]. Técnicas estocásticas e determinísticas como Monte Carlo e simulações por Dinâmica Molecular (DM), respectivamente, são as principais metodologias empregadas em predições *ab initio* [68]. ROSETTA [76, 84] e LINUS [85–87] são exemplos de métodos de predição *de novo*, os quais através de um conjunto de funções de classificação (*scoring functions*) e de funções especiais para cálculo de energia potencial, derivadas de métodos puramente *ab initio*, buscam a predição *de novo*s enovelamentos. Os métodos *de novo* são os que, atualmente, apresentam os melhores resultados em predições [63].

No entanto, estas metodologias de predição possuem limitações: métodos baseados em modelagem comparativa por homologia somente podem realizar a predição de estruturas que possuem seqüências idênticas ou bastante similares em um banco de dados de estruturas conhecidas. Métodos *de novo* e métodos *ab initio*, por sua vez, tornam possível a obtenção de estruturas cujas formas de enovelamento ainda não são conhecidas. Entretanto, a complexidade e o tamanho do espaço dimensional [65], mesmo para uma pequena molécula polipeptídica, tornam o problema da predição intratável computacionalmente [44, 53] (Paradoxo de *Levinthal*), apesar da existência de plataformas computacionais com alto poder de processamento, algumas das quais, como a série BlueGene [2,60], custam dezenas ou até centenas de milhões de doláres.

Neste capítulo são abordadas as principais metodologias existentes para a predição *in sílico* da estrutura 3D de polipeptídeos. Incialmente são descritos os métodos de modelagem comparativa por homologia e os métodos baseados em conhecimento. Em seguida, são descritos os métodos *de novo* e *ab initio*.

3.2 Modelagem comparativa por homologia

Na modelagem comparativa por homologia uma seqüência de resíduos de aminoácidos de uma proteína (seqüência alvo) é alinhada contra a seqüência de aminoácidos de outra proteína com estrutura conhecida e armazenada no PDB [8] (seqüência e estrutura-molde). Caso a seqüência alvo seja bastante similar à seqüência de estrutura conhecida, utiliza-se esta estrutura como um molde para a modelagem da estrutura da proteína [42] [58].

Segundo Sternberg [89] um método de predição baseado em homologia possui seis passos básicos:

1. Encontrar um conjunto de proteínas, determinadas experimentalmente, que são homólogas à seqüência da proteína-alvo;

2. Estabelecer um alinhamento de seqüência entre a seqüência-alvo e as proteínas com estruturas determinadas experimentalmente;

3. Identificar segmentos da cadeia principal da estrutura desconhecida que são conservados em relação a estruturas conhecidas;

4. Modelar as regiões variáveis [1];

5. Modelar as cadeias laterais da proteína-alvo;

6. Refinar a estrutura predita através de métodos de refinamento[2], tendo como base a energia potencial.

A técnica de modelagem comparativa por homologia pode ser aplicada toda vez que for possível detectar uma relação evolucionária entre a proteína-alvo e a proteína-modelo, a qual, tem a estrutura 3D conhecida [12, 42, 58]. A relação evolucionária entre proteínas é uma fator fundamental nos métodos de modelagem comparativa por homologia, pois, parte-se do preceito de que proteínas alvo podem ser moldadas a partir de proteínas homólogas com estrutura 3D determinada experimentalmente. A estrutura destas proteínas são similares no sentido de que aminoácidos com características fisico-químicas idênticas ocupam posições iguais em proteínas homólogas. Os ângulos de torção da cadeia principal também preservam um certo padrão em seus valores.

Segundo Tramontano [91], a análise comparativa por homologia é o método mais utilizado para predição de estruturas de proteínas, e isto se deve a duas razões: primeiro à qualidade dos modelos preditos, que possuem uma razoável relação evolucionária, apresentam-se com uma acurácia maior do que aquelas produzidas com técnicas diferentes. O segundo motivo, se refere ao fato de que a confiabilidade do modelo pode ser estimada a *priori*. Desta forma, é possível estimar a qualidade da estrutura predita.

A Figura 18 esquematiza um processo típico de modelagem comparativa por homologia. Inicialmente, seqüências similares à seqüência-alvo são coletadas usando ferramentas de busca em banco de dados [91], tais como, FASTA [72], BLAST ou PSI-BLAST [3]. As seqüências encontradas são realinhadas usando um programa para alinhamento múltiplo de seqüências, como por exemplo CLUSTALW [35] ou T-COFFEE [66]. Este alinhamento entre a seqüência da proteína-alvo e a(s) seqüência(s) de proteína(s) com estrutura(s) conhecida(s) irão formar a base do modelo. Em seguida, procede-se construíndo a cadeia principal das regiões homólogas à estrutura da cadeia principal de regiões divergentes e por último as cadeias laterais. A avaliação final do modelo é feita levando-se em consideração toda a informação disponível da proteína de interesse [91]. Segundo Baxevanis [6], o passo mais crítico na modelagem por homologia

[1] Região variável: região de estruturas irregulares.

[2] Métodos de refinamento: após a construção do modelo da proteína-alvo é necessário otimizá-lo. As interações desfavoráveis entre átomos não-ligados, bem como as energias de ângulos torcionais e de ligação são otimizadas por métodos de mecânica molecular.

é o alinhamento. Um alinhamento incorreto pode produzir um efeito de distorção dos demais passos, gerando um modelo de estrutura final distorcido e incorreto.

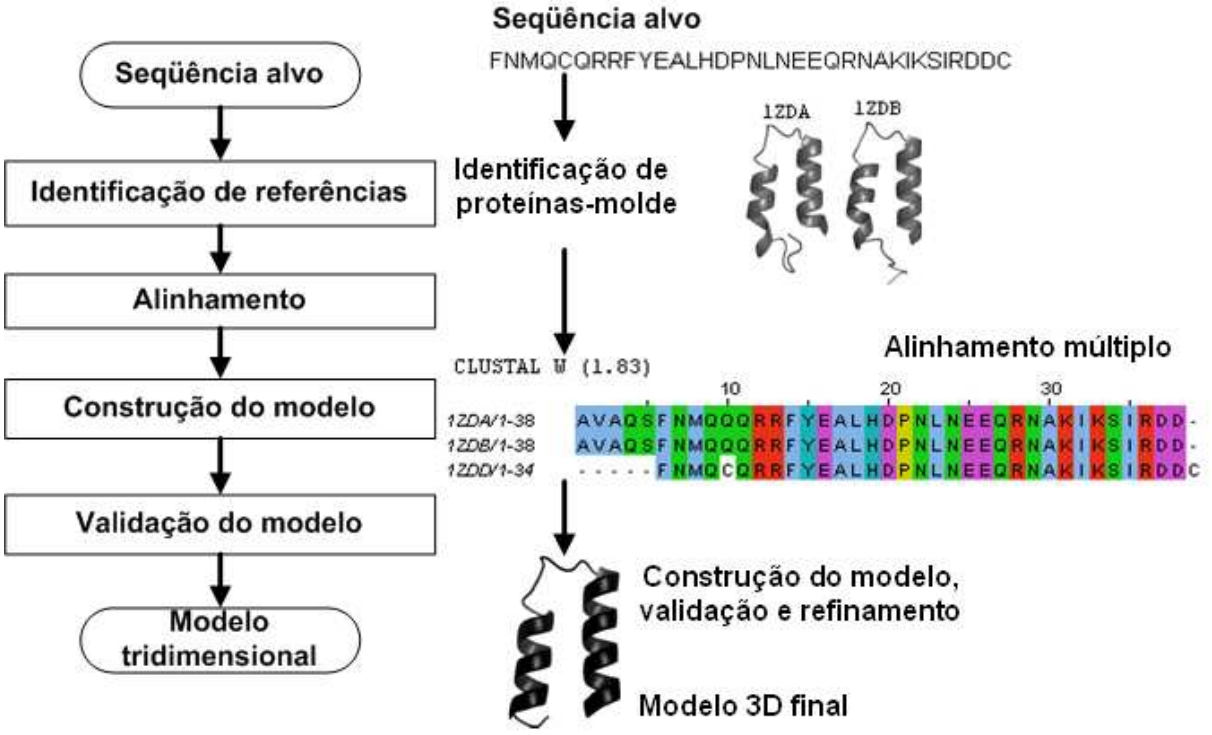

Figura 18 – Representação esquemática de um processo típico de modelagem comparativa por homologia: inicialmente são identificadas proteínas-molde. Em seguida a seqüência da proteína-alvo é alinhada contra a seqüência das proteínas-moldes e, posteriormente, um modelo é construído e validado, obtendo-se ao final, a estrutura 3D da proteína-alvo. Caso necessário a estrutura final pode passar por um processo de refinamento. Adaptado de [58].

Apesar da alta qualidade em suas predições, a técnica de modelagem comparativa por homologia apresenta limitações. A primeira limitação diz respeito à impossibilidade de realizar a predição de novas formas de enovelamento. Isto é explicado pelo fato de tal metodologia estar presa às formas de enovelamento conhecidas e armazenadas no banco de dados de estruturas (PDB). A segunda limitação está no fato de não ser possível estudar o processo de envovelamento da proteína, ou seja, o caminho que a proteína percorre do seu estado desenovelado até o seu estado enovelado e funcional (estado nativo).

3.3 Modelagem baseada em conhecimento: reconhecimento de padrões de enovelamento via alinhavamento

Os métodos baseados em conhecimento (*knowledge-based*) utilizam, para realizar as suas predições, potenciais estatísticos derivados da análise de padrões de enovelamento de proteínas

com estruturas 3D conhecidas e armazenadas em uma base de dados de estruturas. Métodos de reconhecimento de motivos estruturais via alinhavamento ou *threading* são um exemplo de método pertencente a esta metodologia. Estes métodos se baseiam na observação de que uma larga porcentagem de proteínas adota um número limitado de formas de enovelamento. Existem aproximadamente 10 diferentes formas de enovelamento em 50% das estruturas conhecidas [79]. Através da detecção de similariedades estruturais, as quais não podem ser detectadas unicamente pela similariedade entre as seqüências de aminoácidos, são construídos os modelos 3D. A Figura 19 esquematiza um método genérico de predição de estruturas baseado no reconhecimento de padrões de enovelamento.

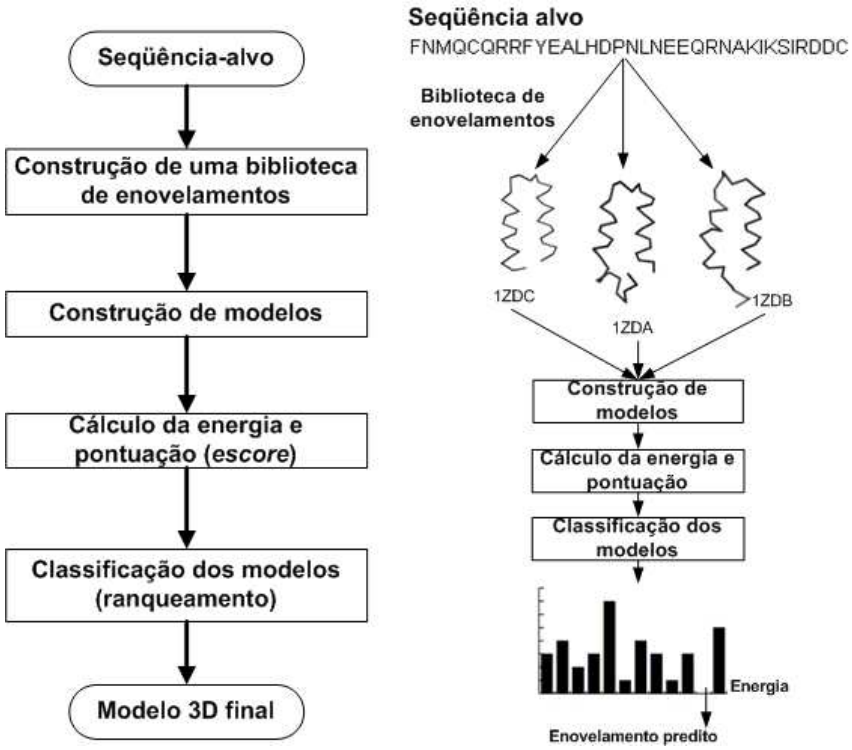

Figura 19 – Representação esquemática de um processo típico de modelagem baseada no reconhecimento de padrões de enovelamento via alinhavamento (*Threading*): inicialmente é construída uma biblioteca de enovelamentos, em seguida são construídos modelos, a energia de cada modelo é calculada e as conformações são pontuadas (*escore*), os modelos são classificados (ranqueadas). Figura adaptada de [35]

.

Inicialmente, para uma dada seqüência de resíduos de aminoácidos, é construída uma biblioteca de padrões de enovelamento. Se fragmentos da seqüência da proteína-alvo se ajustam bem à estas formas de enovelamento, é possível deduzir um alinhamento, mesmo que não haja informação suficiente para construir um modelo 3D completo. Em um segundo momento, a partir das informações obtidas de proteínas com estruturas conhecidas, são construídos modelos estruturais. Com base no valor retornado de uma função objetivo, estes modelos estruturais são pontuados (*escore*). A partir da pontuação obtida por cada modelo estrutural todas as con-

formações construídas são classificadas (ranqueadas) e o modelo 3D final é escolhido. O alinhavamento é freqüentemente utilizado para identificar homologias que não podem ser descobertas por um alinhamento par a par de seqüências de proteínas.

3.4 Predição *ab initio*

Os métodos *ab initio* são fundamentados na termodinâmica e baseiam-se no fato de que a estrutura nativa de uma proteína corresponde ao mínimo global de sua energia livre [91]. Esta metodologia simula o espaço conformacional da proteína utilizando uma função de energia potencial, a qual, descreve a energia interna da proteína e suas interações com o meio em que está inserida. O objetivo é encontrar um mínimo global de energia livre que corresponda ao estado nativo ou funcional da proteína [68, 91].

A predição de estruturas 3D de proteínas, por meio de métodos *ab initio*, pode ser dividida em dois sub-problemas: o primeiro diz respeito ao cálculo da energia de uma dada conformação e, o segundo, diz respeito à estratégia de busca utilizada para encontrar todas as possíveis conformações [91]. O cálculo da energia potencial de uma conformação trata da representação da conformação através de parâmetros da física clássica. A estratégia de busca representa um método pelo qual o espaço de busca conformacional da proteína é percorrido, através de alterações na conformação, buscando encontrar aquela que apresente a menor energia potencial [68].

Os métodos *ab initio* conseguem prever novas formas de enovelamento, pois, não estão limitados à proteínas que possuem sua estrutura já conhecida como na modelagem comparativa por homologia e no alinhavamento. No entanto, os métodos *ab initio* apresentam problemas no que se refere à dimensão do espaço de busca conformacional [44, 53]. Este problema é conhecido e freqüentemente referenciado por vários autores como o *Paradoxo de Levinthal*, devido aos estudos realizados por Cyrus Levinthal em 1968 [53], onde este verificou que é impossível, computacionalmente, reproduzir o processo de enovelamento de uma proteína, devido ao grande número de estados conformacionais que tal molécula pode assumir.

Técnicas estocásticas e determinísticas como Monte Carlo e simulações por Dinâmica Molecular (DM), respectivamente, são exemplos de metodologias para busca de conformações usando como parâmetro a energia potencial [68]. Nas seções seguintes serão apresentadas, brevemente, um modelo de função para cálculo da energia potencial (EP) de uma conformação e as duas metodologias para busca de conformações (Monte Carlo e DM).

3.4.1 Funções de energia potencial

A energia da conformação de uma proteína pode ser descrita através de uma função de energia potencial (EP) composta pelas interações entre átomos da proteína. Esta função é formada

pela soma das interações covalentes e não-covalentes e pode ser descrita como:

$$E_C = \sum_{lig} K_b(b_c - b_{eq})^2 + \sum_{ang} K_\Theta(\Theta_C - \Theta_{eq})^2 + \sum_{die} \frac{K\phi}{2}[1 + cos(n\phi_C - \Upsilon)] + \sum_{lnc}[\frac{A_{ij}}{r_{ij,C}^{12}} - \frac{B_{ij}}{r_{ij,C}^{6}} + k_e\frac{q_1 q_2}{r_{,C}^{2}}]$$

.

Esta função (E_C) retorna o valor aproximado da energia potencial de uma conformação da proteína, através da soma dos potenciais de ligação (*lig*), dos potenciais angulares (*ang*), potenciais de torção (*die*) e da energia potencial dos átomos não ligados (*lnc*). Existem diversos programas que foram desenvolvidos para realizar o cálculo da energia potencial de moléculas, cada qual apresentando suas particularidades, porém seguindo a estrutura de função de energia acima descrita. Exemplos destes softwares são: CHARMM [11, 56], AMBER [13] e GROMMOS [82].

A seguir é descrita cada uma das interações presentes na função de energia potencial E_C. Inicialmente são descritas as interações covalentes e em seguida as interações não-covalentes.

Interações covalentes: as ligações covalentes ocorrem quando átomos compartilham um par de elétrons [54]. As ligações covalentes alteram a natureza dos átomos envolvidos, no caso das proteínas, por exemplo, as ligações covalentes são responsáveis por manter unidos um aminoácido ao outro através de ligações peptídicas da cadeia principal [29]. Fazem parte das interações covalentes a energia potencial liberada pela ligação de dois átomos, o potencial angular e o potencial dos ângulos diedros.

A energia potencial de uma ligação entre dois átomos pode ser descrita pela Equação 3.1. Onde E_{lig} é a energia potencial da ligação química, *b* é o comprimento da ligação, b_{eq} o comprimento de equilíbrio da ligação, e K_b a constante.

$$E_{lig} = K_b(b - b_{eq})^2 \tag{3.1}$$

O potencial angular é representado pela Equação 3.2. Onde E_{ang} é a energia de variação dos ângulos de ligação.

$$E_{ang} = K_\Theta(\Theta - \Theta_{eq})^2 \tag{3.2}$$

E_{die} é a energia potencial dos ângulos diedros, onde *N* é o número de mínimos e γ o ângulo de fase (Equação 3.3).

$$E_{die} = \sum_{n=1}^{N} K_\phi[1 + cos(n\phi - \gamma)] \tag{3.3}$$

Interações entre átomos não ligados covalentemente: são todas aquelas interações entre átomos que não estão conectados por ligações covalentes, são forças mais fracas do que as forças covalentes, porém, contribuem de forma muito expressiva para a estabilidade da molécula. Fazem parte das interação não covalentes as forças elétrostáticas, as energias de dispersão, as energias de repulsão (força de *van der Waalls*).

Nas interações eletrostáticas, um núcleo e seus elétrons interagem de acordo com a lei de *Coulomb* [83], conforme descrito pela Equação 3.4. Onde F_e é a força eletrostática, q_1 e q_2 são as cargas parciais das duas partículas, medidas em unidades chamadas *coulombs (C)*, e k_e ($k_e = 8,99 \times 10^9 N.m^2/C^2$) é a constante de *Coulomb* e r é a distância em metros. A equação de *Coulomb* afirma que a força de atração ou repulsão entre duas cargas (q_2 e q_2) é diretamente proporcional ao produto dos valores absolutos das duas cargas e inversamente proporcional ao quadrado da distância r entre elas.

$$F_e = k_e \frac{q_1 q_2}{r^2} \tag{3.4}$$

As energias de dispersão são representadas pela Equação 3.5. Onde E_{dis} é a energia de dispersão, B_{ij} depende do par de átomos envolvidos e é usualmente estimado empiricamente dos dados derivados de raios X de estruturas cristalinas.

$$E_{dis} = -\frac{B_{ij}}{r_{ij}^6} \tag{3.5}$$

Um outro fato que precisa ser levado em consideração é que o orbital de um átomos não pode ser sobreposto por outro devido ao princípio de exclusão de *Pauli*, em que o estado de dois elétrons não podem possuir o mesmo estado quântico. Este efeito pode ser observado ao se assumir que cada átomo é uma esfera rígida com um raio específico (o raio de *van der Waalls*) e dois átomos não podem ter seus raios sobrepostos. Esta energia representa a energia de repulsão descrita pela Equação 3.6. Onde E_{rep} é a energia de repulsão, A_{ij} é o termo determinado empiricamente.

$$E_{rep} = \frac{A_{ij}}{r_{ij}^{12}} \tag{3.6}$$

A união das forças covalentes e das forças não-covalentes representa a energia da conformação da molécula em um dado momento. São estas forças as componentes de uma função de energia potencial (EP) que descreve por meio de parâmetros da física clássica, o estado fisico-químico da molécula. A conformação de uma proteína em seu estado nativo é aquela que possuí a menor energia potencial livre.

3.4.2 Métodos para busca de conformações

Diversos métodos para busca por conformações com menor energia potencial foram desenvolvidos nos últimos anos. Estes métodos são classificados em dois grupos: métodos estocásticos e métodos determinísticos. Algoritmos baseados em técnicas estocásticas como Monte Carlo pertencem ao primeiro grupo e métodos determinísticos, como Dinâmica Molecular (MD), pertencem ao segundo grupo.

Dinâmica Molecular: em uma simulação por DM, uma trajetória de um sistema molecular é gerada pela solução simultânea da equação de movimento de *Newton* para todos os átomos do sistema molecular. Cada átomo de uma proteína possui uma energia potencial e sente conseqüentemente uma força externa igual à derivação espacial desta energia. Através da segunda lei de *Newton* é possível calcular a força de ação de cada átomo, conforme pode ser observado na Equação 3.7.

$$F_i = m_i \frac{\alpha^2 x}{\alpha t^2} \tag{3.7}$$

A partir da força de ação de cada átomo é possível calcular a posição de cada átomo ao longo de uma série de curtos passos de tempo, e isto, resulta numa série de estruturas ao longo do tempo, a qual é comumente chamada de trajetória. Na prática é selecionada uma temperatura *T*, a partir da qual é computada uma velocidade inicial de distribuição dos átomos conforme a energia cinética total do sistema de acordo com a Equação 3.8.

$$f = (\frac{m}{2\pi KT})^{2/3} e^{-\frac{mv^2}{2KT}} \tag{3.8}$$

Esta equação, proporciona a fração f de partículas de massa m e velocidade v em um sistema com temperatura T. Para integrar a equação de *Newton*, calcula-se a aceleração, a velocidade, e a posição de cada átomo em cada passo de tempo, conforme descrito nas Equações 3.9, 3.10 e 3.11.

$$a_i(t) = \frac{F_i}{m_i} \tag{3.9}$$

$$v_i(t + \Delta t) = v_i(t) + a_i(t)\Delta t \tag{3.10}$$

$$r_i(t + \Delta t) = r_i(t) + v_i(t + \Delta t)\Delta t \tag{3.11}$$

Segundo Tramontano [91], o passo de tempo (*timestep*) precisa ser suficientemente pequeno para garantir que a aceleração seja praticamente constante durante o mesmo. Simulações por

Dinâmica Molecular a uma temperatura relativamente alta podem ser usadas para explorar uma fração mais ampla do espaço conformacional da proteína.

Monte Carlo: métodos por DM simulam o espaço conformacional de uma proteína através de movimentos ao longo de uma trajetória. Porém esta não é a única opção, uma outra alternativa é mover-se de uma conformação à outra por meio de amostragens de aceitação ou rejeição da nova conformação de acordo com uma única função de energia [91]. Métodos baseados na técnica de Monte Carlo [80] iniciam o processo de busca por conformações de menor energia potencial, partindo de uma conformação inicial com energia potencial E e realizando alterações na conformação obtendo novas conformações [80]. A energia E' de uma nova conformação é calculada. Se E' é menor que E o movimento realizado na conformação é aceito, e assim procede o processo de busca. Nos métodos de Monte Carlo aceita-se uma configuração que reduz a energia do sistema, porém não necessariamente exclui-se as demais conformações que não obtiveram valores de energia menores. O não descarte destes valores, fará com que, durante o processo de busca, o método de Monte Carlo não fique preso em um mínimo local [91].

3.5 Predição *de novo*

Os métodos *de novo* são aqueles métodos de predição que, através de um conjunto de funções de classificação (*scoring functions*) e de funções especiais para cálculo de energia potencial (EP), derivadas de métodos puramente *ab initio*, buscam a predição de novas formas de enovelamento. Os métodos *de novo* são os que, atualmente, apresentam os melhores resultados nas predições realizadas no CASP[3] [63]. São exemplos de métodos *de novo*: ROSETTA (ROBETTA [4]) [75, 76, 84], LINUS [85–87] e FRAGFOLD [42].

Os métodos *de novo* buscam realizar a predição de novas formas de enovelamento baseando-se em moldes. Esta concepção surge a partir da observação de que quando um novo enovelamento é descoberto, este é composto por motivos estruturais comuns de fragmentos ou de estruturas supersecundárias de proteínas com estruturas conhecidas [91]. Desta forma, se existem fragmentos da proteína que se enovelam em estruturas similares, então é possível utilizar esta informação ou estes fragmentos na construção de novos modelos estruturais 3D de proteínas. Esta é a base dos métodos *de novo* baseados em fragmentos. Nestes métodos, a conformação de uma dada seqüência alvo é construída com base em informações extraídas de fragmentos de proteínas com estruturas 3D conhecidas [91]. A conformação de uma proteína passa a ser vista

[3]CASP (*Critical Assessment of Techniques for Protein Structure Prediction*): evento realizado a cada dois anos para avaliar a qualidade das predições realizadas pelos atuais métodos. O mesmo reúne vários grupos de pesquisa que desenvolvem métodos para predições de 3D estruturas de proteínas. Atualmente o CASP está em sua sétima edição. Maiores informações podem ser obtidas em http://predictioncenter.gc.ucdavis.edu/.

[4]ROBETTA (*Full-chain Protein Structure Prediction Server*): servidor de predição de estruturas 3D de proteínas que utiliza o método ROSETTA. Disponível em: http://robetta.bakerlab.org/index.html

como um conjunto de vários fragmentos da seqüência de aminoácidos representando motivos estruturais diversos.

A estratégia adotada pelos métodos baseados em fragmentos é coletar as estruturas locais assumidas por pequenos segmentos de fragmentos em estruturas 3D já conhecidas e realizar a combinação destas estruturas locais para produzir um número de prováveis modelos 3D de uma proteína-alvo, onde o modelo final é selecionado levando-se em consideração a energia potencial mínima (derivada de métodos *ab initio*) [91].

Devido ao fato de que seqüências locais semelhantes nem sempre assumem a mesma estrutura 3D, por motivo do efeito do grande número de interações ocorrentes na estrutura terciária da proteína, os métodos de predição baseados em fragmentos não podem simplesmente fragmentar a seqüência de aminoácidos da proteína-alvo e através de consulta em bases de dados de estruturas de proteínas-molde, obter as informações de enovelamento do fragmento em questão e realizar a junção destes fragmentos sem nenhum critério. É necessário que sejam estabelecidos critérios de relação entre fragmentos de forma que se possa determinar os fragmentos com maior probabilidade de inserção durante a construção do modelo final da seqüência-alvo. Estes critérios surgem da idéia de que existem interações não-covalentes entre os átomos da molécula e de que desta forma uma determinada região da proteína é influenciada por interações que ocorrem em outra região estrutural [91].

Em métodos baseados em fragmentos, a cadeia principal da proteína-alvo é enovelada através da alteração em seus ângulos de torção com base em ângulos de estruturas de proteínas-molde ou ainda a partir de informações de posicionamento de átomos no espaço 3D.

Um aspecto importante destes métodos de predição está relacionado à forma e aos termos utilizados por uma função que realiza o classificação (ranqueamento) e exploração de todas as possíveis combinações de fragmentos para construção das estruturas. A classificação dos fragmentos visa estabelecer uma relação de ordem no conjunto de fragmentos, buscando organizar aqueles que são mais aptos a ocupar determinada região na cadeia principal da proteína-alvo. Esta classificação é realizada por meio da análise do grau de similaridade entre os fragmentos e através da análise das interações de outros fragmentos da proteína que podem influenciar na forma que este fragmento irá enovelar-se. O método ROSETTA [75, 76, 84] utiliza um complexo conjunto de funções (lista de fórmulas utilizadas pelo Método ROSETTA na tabela 1 e 2 de Baker [75]).

De forma geral, um método de predição baseado em análise e combinação de fragmentos é composto por quatro etapas distintas onde, dada a seqüência completa de aminoácidos de uma proteína, este procede:

- Dividindo a seqüência alvo em fragmentos;
- A partir de cada fragmento, realiza-se a busca por seqüências (dos fragmentos) similares em um banco de dados de estruturas conhecidas;
- Os fragmentos-molde são classificados (ranqueamento);

- A partir dos fragmentos-molde e com a utilização de uma técnica de combinação, a estrutura tridimensional é construída;
- A conformação é refinada.

A Figura 20 apresenta o diagrama esquemático do processo de um método *de novo*.

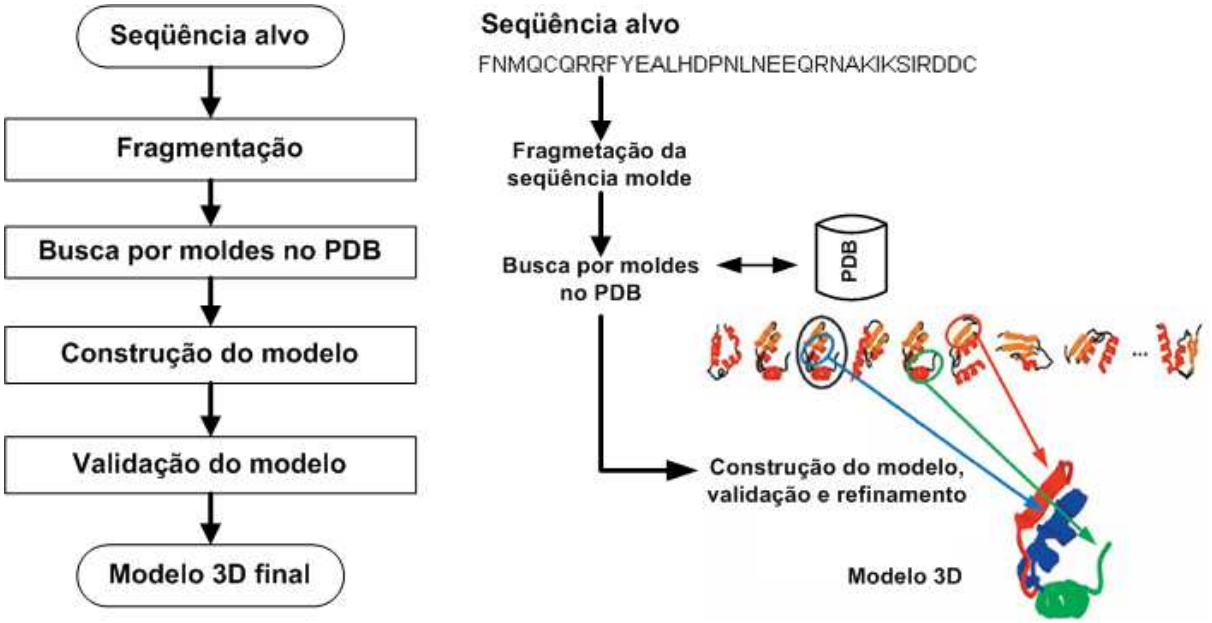

Figura 20 – Representação esquemática de um método *de novo* utilizando combinação de fragmentos: a seqüência alvo é fragmentada, moldes são obtidos do PDB, os fragmentos são classificados, a conformação é construída e caso necessário a conformação é refinada.

Na seção seguinte são brevemente descritos os três principais métodos *de novo* existentes.

3.5.1 Métodos *de novo*

Nesta seção são descritos brevemente os três principais métodos de predição existentes, os quais, fazem uso da análise e combinação de fragmentos para a predição de estruturas 3D de proteínas. O método LINUS é considerado um método *de novo*, no entanto, não faz uso de combinação de fragmentos.

Primeiramente descreve-se o método ROSETTA [75, 76, 84] e em seguida o método LINUS [85–87] e, por fim, o método *FRAGFOLG* [41].

1. ROSETTA [75, 76, 84]: este método implementa um modelo de predição de enovelamento, onde pequenos fragmentos da cadeia principal da proteína-alvo são alterados pela utilização de informações estruturais de proteínas-molde [12, 51]. A probabilidade de uma conformação em particular ser assumida é baseada na similaridade entre a seqüência

local da proteína-alvo e as proteínas-modelo. O primeiro passo do método ROSETTA consiste em fragmentar a seqüência alvo em fragmentos com 3 e 9 resíduos aminoácidos. Em seguida, são obtidas proteínas com seqüência homóloga no PDB. O método ROSETTA explora as possíveis combinações de fragmentos utilizando simulações de Monte Carlo [75]. A função de energia possui termos que refletem a compactação, fitas β emparelhadas e o isolamento interior de resíduos hidrofóbicos em proteínas [51]. O método ROSETTA realiza 1000 simulações independentes, com a estrutura inicial selecionada a partir do padrão de distribuição de conformações dos fragmentos, gerado previamente. Estas estruturas geradas, são agrupadas, e aquelas conformações localizadas no centro dos grupos maiores são as definidas como as predições da estrutura 3D da proteína-alvo [51]. Segundo [12], a principal diferencial entre o método ROSETTA e os demais métodos baseados na combinação de fragmentos se encontra no seu caráter estocástico e iterativo da utilização de múltiplos e pequenos fragmentos de proteínas diferentes para a construção das estruturas 3D. O método ROBETTA consegue gerar estruturas de proteínas completamente desconhecidas graças a forma que trata a inserção e a busca de partes da estrutura que não são conhecidas. Atualmente, ROSETTA é o método de predição que apresentou os melhores resultados no CASP. Na versão do CASP VII (2006), o método ROSETTA, para proteínas de até 200 resíduos de aminoácidos, obteve-se resultados de RMSD entre o C_α da estrutura predita e da estrutura experimental, na faixa de 1.40Åà 5.10Å(Figura 5 - *Examples of successful free modelling predictions* encontrada em [19]).

2. LINUS [85–87]: este método se baseia no fato de que as proteínas estão organizadas hierarquicamente. A formação destas hierarquias sugere que cadeias de regiões vizinhas interagem entre si para formar módulos primitivos de enovelamento. O enovelamento da proteína ocorre através de uma condensação hierárquica. Nesta condensação a estrutura secundária da proteína é codificada em quatro categorias: hélice α, folha β, alça e volta. Um enovelamento complexo pode ser decomposto utilizando estes quatro elementos base. LINUS, inicia a predição da estrutura terciária da proteína-molde com a cadeia polipeptídica totalmente estendida. Em seguida utilizando pré-definições geométricas e uma função de energia realiza alterações na cadeia polipeptídica, através de mudanças em seus ângulos de torção. As conformações são geradas aleatóriamente. Um algoritmo baseado em Monte Carlo é utilizado para decidir se a estrutura gerada será aceita ou se o algortimo procede retornando à estrutura anterior. Estas alterações na conformação são realizadas seqüêncialmente, isto é, aminoácido por aminoácido. A hierarquia é estabelecida através do reconhecimento de conformações favoráveis em cada ciclo de execução (um ciclo representa as alterações em todos os resíduos de aminoácidos). LINUS realiza um grande de número de passos. Onde, amostra periodicamente as conformações dos resíduos para acumular informações estatísticas de preferências estruturais. Segundo Lesk [51], a representação do processo de enovelamento de uma proteína, no método de

predição LINUS, é realística, apesar de ser aproximada. Nela, todos os átomos da proteína, exceto os hidrogênios, são modelados, mas a função de energia é aproximada e a dinâmica simplificada. A função de energia do LINUS considera: a repulsão estérica que impede a sobreposição dos átomos, o agrupamento dos resíduos hidrofóbicos isolados interiores, as ligações de hidrogênio e as pontes salinas.

3. FRAGFOLD [41]: este método se baseia na idéia de que existem motivos estruturais comuns nas estruturas das proteínas, sendo que estas estruturas somente diferem entre si na forma em que estes motivos estruturais estão combinados. O método FRAGFOLD pré-seleciona fragmentos estruturais de estruturas de proteínas conhecidas. A predição das estruturas secundárias é realizada através da avaliação do ajuste das seqüências nos enovelamentos conhecidos, e desta forma é selecionado o modelo que apresenta a melhor compatibilidade com o fragmento da proteína-alvo. Este método utiliza uma lista de fragmentos construída a partir de todos os fragmentos tripeptídeos, tetrapeptídeos e pentapeptídeos de proteínas-molde. Para cada posição da seqüência da proteína-alvo é computado o potencial de utilização de cada fragmento. Posteriormente, o método FRAGFOLD obtém a estrutura predita da proteína-alvo, realizando a combinação de fragmentos através de um algoritmo genético ou de Simulated Annealing, onde a metade dos movimentos correspondem à inserção da estrutura supersecundária pré-selecionada e a outra metade envolve a escolha livre de qualquer fragmento de tamanho pequeno [12]. As regiões da proteína-alvo, que não foram moldadas, são identificadas através de técnicas de clusterização, predizendo desta forma, o provável enovelamento da cadeia polipeptídica.

Os métodos de predição *de novo* apresentam vantagens em relação aos outros métodos de predição. A primeira vantagem se refere à capacidade de predição de novas formas de enovelamento, o que não é possível de ser realizado pelos métodos baseados em análise comparativa por homologia. A segunda vantagem se refere à redução do espaço de busca conformacional, o que em métodos de predição *ab initio* é um grande problema e que demanda grande esforço computacional. Esta redução do espaço conformacional se deve ao fato que em uma simples substituição de um fragmento na proteína-alvo, está-se movendo de uma região de uma proteína um fragmento que já possui uma estrutura com mínima energia potencial. No entanto, apesar de reduzir o espaço de busca conformacional, os métodos *de novo*, que utilizam fragmentos de proteínas-molde, ainda possuem duas principais limitações. A primeira está relacionada ao desafio de tratar o grande espaço de busca conformacional originado pelas diferentes formas de combinação de fragmentos molde. A segunda, se refere ao desafio de reduzir a energia potencial da conformação nas regiões onde ocorre a combinação dos fragmentos. No CASP VII, o método ROSETTA, utilizando a rede de computação distribuída ROSETTA@home[5], contava com uma

[5]ROSETTA@home: é um projeto de computação distribuída, baseado na oferta voluntária de recursos de processamento de pessoas de todo o mudo. Este projeto procura determinar a estrutura 3D de proteínas que podem ser usadas na pesquisa de medicamentos que possam levar à cura de doenças como a AIDS, a Malária e a Alzheimer. Maiores detalhes podem ser encontrados em: http://boinc.bakerlab.org/rosetta/.

infraestrutura computacional de 140.000 computadores, com aproximadamente 65.000 computadores para serem utilizados em um mesmo tempo. Esta estrutura, representa uma capacidade de processamento de 37 TFlops, o que possibilitou que as predições de estruturas 3D de cada proteína testada fossem feitas em aproximadamente 500.000 horas de processamento [19]. A necessidade desta estrutura computacional retrata o problema da complexidade presente nas funções de *score*, de combinação e refinamento dos métodos *de novo*.

3.5.2 Resumo do capítulo

Neste capítulo foram brevemente descritas as principais metodologias existentes para a predição *in sílico* da estrutura 3D de proteínas. Os principais métodos de predição foram divididos em dois gandes grupos. O primeiro grupo abrigando os métodos baseados em homologia e àqueles baseados em conhecimento. O segundo grupo englobando os métodos *ab initio* e *de novo*.

Os métodos de predição do primeiro grupo obtêm bons resultados em suas predições. No entanto, estas, estão limitadas à informações estruturais de proteínas já conhecidas. Desta forma, não é possível predizer novas formas de enovelamento. O segundo grupo apresenta vantagens no que se refere à predição de novas formas de enovelamento. No entanto, estes possuem como maior desafio tratar o grande espaço conformacional. Atualmente, os métodos *de novo* são os que apresentam os melhores resultados nas predições de novas formas de enovelamento.

Este capítulo elucida os esforços realizados ao longo dos anos para a desenvolvimento de métodos eficazes e eficiêntes para a predição *in sílico* da estrutura 3D de proteínas. A predição da estrutura 3D de uma proteína, a partir de sua seqüência de aminoácidos, é uma problema complexo e ainda sem solução, apesar dos grandes avanços realizados nos últimos anos.

Levando em consideração as vantagens e desvatagens apresentadas, pelos atuais métodos de predição, no Capítulo 4 é apresentada uma nova proposta para a predição *in sílico* da estrutura 3D de proteínas. No método proposto, desenvolve-se uma nova metodologia que ameniza o problema gerado pela grandeza do espaço de busca conformacional e que ainda possa predizer novas formas de enovelamento.

4 Proposta para a predição da estrutura tridimensional de polipeptídeos utilizando cálculo de intervalos

4.1 Introdução

Nos capítulos anteriores foram apresentados os principais conceitos relacionados à Bioinformática Estrutural, mais especificamente à predição de estruturas 3D de polipeptídeos. No Capítulo 2 foram apresentadas os aminoácidos, as proteínas, as cadeias polipeptídicas, sua organização hierárquica estrutural e os bancos de dados de estruturas 3D de proteínas. No Capítulo 3 foram apresentadas as principais métodologias para a predição *in sílico* da estrutura 3D de polipeptídeos, suas vantagens e limitações. Neste capítulo será apresentada uma nova proposta para a predição da estrutura 3D de proteínas. O método desenvolvido caracteriza-se como um método *de novo*, no sentido de que busca se beneficiar da alta acurácia encontrada em métodos de predição baseados em modelagem comparativa por homologia e da capacidade de predição de novas formas de enovelamento presentes em métodos *ab initio*.

Na próxima seção será descrita a estrutura geral de um método para a predição da estrutura 3D de proteínas. Esta estrutura, serve como base para o método de predição desenvolvido neste trabalho. Nela, são apresentados aspectos relacionados à forma de representação de uma cadeia polipeptídica, à utilização de uma função de energia como parâmetro para escolha de conformações, à um método de busca para percorrer o espaço conformacional e uma função para análise de conformações preditas. Na medida em que cada uma destas etapas é apresentada e discutida, são descritos aspectos relacionados ao método de predição desenvolvido.

4.2 Estrutura geral de um método para predição de estruturas 3D de polipeptídeos

No desenvolvimento de um método de predição de estruturas 3D de polipeptídeos, precisam ser levados em consideração quatro fatores. Estes abrangem aspectos de representação e metodologia para busca e análise de conformações preditas [18], os quais, tratam especificamente de:

- Formas de representação da cadeia polipeptídica;
- Uma função de energia potencial que atue como uma função de custo, representando a

conformação de uma proteína de um ponto de vista fisico-químico;

- Métricas para avaliar a similaridade entre a conformação nativa e as conformações preditas de uma proteína; e
- Uma metodologia para a busca por conformações no espaço tridimensional.

Estes quatro fatores são a base do método de predição de estruturas 3D desenvolvido e apresentado neste capítulo. Cada um destes fatores é descrito em maiores detalhes nas próximas seções.

4.2.1 Representação da cadeia polipeptídica

A conformação da cadeia poliptídica de uma proteína pode ser representada de diferentes maneiras. As formas mais comuns de representação são:

1. Coordenadas cartesianas de todos os átomos;
2. Coordenadas cartesianas de átomos pesados (*heavy atoms*);
3. Coordenadas cartesianas dos átomos da cadeia principal e centróides da cadeia lateral;
4. Coordenadas cartesianas do carbono α; ou
5. Ângulos de torção da cadeia principal e da cadeia lateral.

No método desenvolvido nesta dissertação utiliza-se os ângulos de torção da cadeia principal e o ângulos de torção das cadeias laterais para representar uma conformação. Esta escolha se baseia no fato de que a conformação da cadeia polipeptídica pode, simplesmente, ser descrita pelos valores dos ângulos diedros da estrutura principal da proteína. São estes, o ângulo diedro ϕ (*phi*) em torno da ligação entre o nitrogênio (N) e o carbono α (Cα), o ângulo diedro ψ (*psi*) em torno da ligação entre o carbono alfa (Cα) e o carbono (C) e o ângulo diedro ω (*omega*) em torno da ligação peptídica entre o carbono (C) e nitrogênio (N) (conforme descrito na seção 2.3: ligação peptídica). Cada ângulo diedro possui graus de liberdade e de restrição em suas torções: estas liberdades são visíveis pelo mapa de Ramachandran (conforme descrito na Seção 2.3.3, Figura 10) e representam regiões onde podem ocorrer combinações de ângulos diedros (*phi* e *psi*) nas quais não ocorrem choques estereoquímicos entre átomos das cadeias laterais.

Uma caracterísitica importante da representação geométrica baseada em ângulos de torção está no fato de que pequenas mudanças nos ângulos diedros ϕ e ψ de um peptídeo podem induzir mudanças ignificantes na conformação, pois, estes ângulos possuem maior liberade em suas torções (conforme descrito na Seção 2.3.1). Os ângulos χ da cadeia lateral ocorrem em número diferente, e dependem exclusivamente do tipo de resíduo de aminoácido, conforme pode ser observado na Tabela 2 e descrito na Seção 2.3.2.

As cadeias laterais influenciam as torções dos ângulos diedros da cadeia principal. O posicionamento incorreto das cadeias laterais pode ocasionar choques estereoquímicos entre átomos da cadeia lateral e átomos da cadeia principal, ou ainda choques entre átomos de cadeias laterais vizinhas, o que restringe as torções dos ângulos diedros e conseqüentemente, influencia na estrutura 3D adotada pela cadeia polipeptídica.

A utilização de ângulos de torção para representar a conformação de uma cadeia polipeptídica tem ainda uma outra grande vantagem, o número de variáveis para manipulação é menor do que o número de variáveis presentes em representações baseadas em coordenadas cartesianas. Isto torna mais eficiente de um ponto de vista computacional, o tratamento destas variáveis.

4.2.2 Função de custo

Após definida a forma de representação de uma cadeia polipeptíca, é necessário poder avaliar se uma dada conformação é melhor do que outra. Para isto, é necessário um critério de avalição, ou uma função de custo que sirva como parâmetro para seleção de modelos conformacionais. Uma função de custo representa este critério de avaliação.

A função de custo mais utilizada na literatura se trata da análise de conformações por meio de parâmetros da física clássica. São freqüentemente referenciadas como funções de energia potencial (EP) [91]. Estas funções retornam um valor de energia baseando-se na conformação da molécula, provendo, desta forma, a informação necessária para determinar quais conformações são melhores, do ponto de vista termodinâmico, em comparação com outras. A maior parte das funções de energia tem a forma

$$E(C) = \sum_{\text{ligações}} B(C) + \sum_{\text{ângulos}} A(C) + \sum_{\text{torções}} T(C) + \sum_{\text{não-ligados}} N(C)$$

Onde, C é uma conformação e E(C) é a energia potencial da conformação obtida pela soma dos potenciais de ligação (B), potenciais angulares (A), potenciais das torções (T) e do potencial de energia dos átomos não ligados (N). O número de funções de energia presentes na literatura é grande. Dentre as principais se pode citar: AMBER [13], CHARMM [11, 56] e GROMMOS [82].

Neste trabalho, foi utilizado a função de energia do CHARMM(*Chemistry Harvard Macromolecular Mechanics*) [11, 56] versão 27 para cálculo da energia potencial das conformações geradas. A função de energia do CHARMM é composta pela soma de diversas funções da mecânica molecular agrupadas em dois grupos principais: átomos ligados covalentemente (potencial de vibração do ângulo, estiramento da ligação química, torção dos ângulos diedros, *Urey-Bradley* (UB) e impróprios) e átomos não ligados covalentemente (potencial de van der Waalls e potencial elétrostático). A Função de energia do CHARMM tem a forma:

$$E(CHARMM) = \underbrace{\sum_{\text{ligações}} K_b(b-b_0)^2}_{E_1} + \underbrace{\sum_{\text{UB}} k_{UB}(S-S_0)^2}_{E_2} + \underbrace{\sum_{\text{ângulos}} K_\theta(\theta-\theta_0)^2}_{E_3} +$$
$$\underbrace{\sum_{\text{diedros}} k_\chi[1+\cos(n\chi-\delta)]}_{E_4} + \underbrace{\sum_{\text{impróprios}} k_{imp}(\phi-\phi_0)^2}_{E_5} +$$
$$\underbrace{\sum_{\text{não-ligados}} \varepsilon_{ij}[(\frac{Rmin_{ij}}{r_{ij}})^{12} - (\frac{Rmin_{ij}}{r_{ij}})^6]}_{E_6} + \underbrace{[\frac{q_i q_j}{er_{ij}}]}_{E_7}.$$

onde,

- b é o tamanho da ligação, b_0 é a distância de equilíbrio de ligação e K_b é a constante de força de ligação;
- S é a distância entre dois átomos separados por duas ligações covalentes, S_0 é a distância de equilíbrio e K_{UB} é a constante de força de *Urey-Bradley*;
- θ é o ângulo de valência, θ_0 é o ângulo de equilíbrio e k_θ é a constante de força do ângulo de valência;
- χ é o ângulo de torção, k_χ é a constante de força do ângulo de torção, n é a multiplicidade e δ é o ângulo fase.
- ϕ é o ângulo impróprio, ϕ_0 é o equilíbrio do ângulo impróprio e k_{imp} é a constante de força do impróprio; e
- ε_{ij} é a energia de atração/repulsão de Lennard Jonnes, r_{ij} é a distância entre um átomo i e um átomo j, $Rmin_{ij}$ é o raio mínimo de interação, q_i é a carga atômica parcial e e é a constante dielétrica.

Para realizar o cáculo da energia potencial de conformações usando a função de energia do CHARMM, no método de predição desenvolvido foram utilizadas rotinas do pacote de modelagem molecular TINKER (rotinas PDBxyz e analyse) [73].

4.2.3 Métricas de avaliação

Métricas são um conjunto de ferramentas para avaliar o grau de similaridade entre uma conformação predita e a conformação nativa de uma proteína. A métrica mais utilizada para avaliar a similaridade entre estruturas é o desvio médio quadrático (RMSD). A Equação 4.1 calcula o RMSD entre duas estruturas, onde r_{ai} e r_{bi} são, respectivamente, a posição de um

átomo i de uma estrutura a e de uma estrutura b, e onde as estruturas a e b são superpostas de forma otimizada.

$$RMSD(a,b) = \sqrt{\frac{\sum_{i=1}^{n} |r_{ai} - r_{bi}|^2}{n}} \tag{4.1}$$

Neste trabalho utilizou-se o programa ProFit (Grupo Dr. Andrew C. R. Martim) para cálculo do desvio médio quadrático entre as estruturas preditas e a estrutura nativa de uma proteína. O programa ProFit é baseado no algoritmo de subreposição proposto por McLachlan [59] que busca sobrepor de forma otimizada os átomos de duas cadeias polipeptídicas.

4.2.4 Métodologia de busca de conformações e predição de estruturas 3D

O quarto fator presente em um método de predição de estruturas 3D diz respeito a uma metodologia para busca de conformações no espaço conformacional de uma proteína. Conforme descrito no Capitulo 3, diversas metodologias para predição da estrutura 3D de polipeptídeos foram desenvolvidas ao longo dos últimos anos, sendo que o principal empecilho para o sucesso destes métodos têm sido o tratamento eficiente da grandeza do espaço de busca conformacional [53, 65] e a capacidade de predição de novas formas de enovelamento [91]. Esta complexidade pode ser mensurada pelo tempo necessário para encontrar o estado nativo de enovelamento de uma proteína (exemplo apresentado na Seção 1.2).

Devido ao problema da grandeza do espaço de busca conformacional, os métodos atualmente desenvolvidos para predição de estruturas 3D de polipeptídeos buscam moldar conformações usando informações oriundas de proteínas as quais possuem sua estrutura 3D conhecida experimentalmente. Métodos *de novo* como ROSETTA [75, 76, 84], FRAGFOLD [41] e LINUS [85–87], conseguem obter bons resultados em suas predições, porém a elevada complexidade de suas funções de pontuação (escore) exigem em elevado tempo de processamento (meses dependendo do tamanho da proteína). Neste trabalho, buscou-se manter a capacidade de predizer novas formas de enovelamento e desenvolver um novo método de predição que pudesse, de uma maneira rápida, predizer a partir da seqüência de aminoácidos a estrutura 3D aproximada de uma proteína. O método aqui desenvolvido para busca de conformações e predição de estruturas 3D é baseado na construção de intervalos de variação para as torções da cadeia principal de um polipeptídeo alvo, a partir de informações de ângulos diedros obtidos de proteínas molde do PDB. Um intervalo fechado de variação angular, para cada ângulo diedro de cada aminoácido, reduz o espaço de busca conformacional que o polipeptídeo pode assumir. A redução deste intervalo é feita objetivando encontrar o menor intervalo fechado que contenha a conformação de menor energia potencial. Na seção seguinte é formalizado e descrita cada etapa do método de predição desenvolvido.

4.3 O método desenvolvido

Após apresentada a forma adotada para representação da cadeia polipeptídica, a métrica para avaliação de estruturas 3D de proteínas e a função de custo utilizada, é apresentado o método de predição desenvolvido. No método proposto, a conformação C de uma proteína é representada na forma de um vetor $C = \{x_1, x_2, \ldots, x_n\}$, onde x_i é um tripleto de ângulos de torção ω (*omega*), ϕ (*phi*) e ψ (*psi*) (Figura 21) de cada resíduo de aminoácido presente na estrutura primária desta proteína. O conjunto de tripletos consecutivos representa as rotações internas da cadeia principal da proteína.

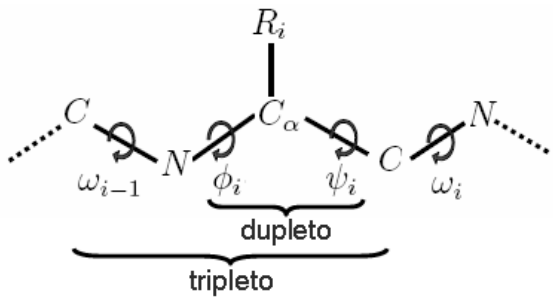

Figura 21 – Representação esquemática de um modelo de peptídeo identificando um dupleto e um tripleto de ângulos de torção da cadeia principal.

As ligações ϕ (entre o grupo NH e o C_α) e ψ (C_α e o grupo C) não possuem restrições quanto à liberdade de rotação na cadeia principal, porém existem algumas combinações não permitidas devido a restrições estéreoquímicas, ou seja, combinações que podem causar colisões desfavoráveis entre átomos da cadeia principal ou das cadeias laterais. Os ângulos ω, por sua vez, são fixados em valores próximos ou iguais a 0° ou +180°. Com isto, é possível concluir que devido a liberdade presente nos ângulos de torção ϕ e ψ, estes são os principais responsáveis pela forma de organização da estrutura 3D de um polipeptídeo. Desta forma, no método proposto são considerados somente os dupletos de ângulos de torção formados pelos ângulos ϕ e ψ (Figura 21) na representação da cadeia principal de um polipeptídeo. No método proposto, os ângulos ômega são mantidos inalterados permanecendo com valor igual à 180°.

O método desenvolvido é composto por nove etapas: (1) inicialmente a seqüência alvo é fragmentada; (2) fragmentos moldes (*templates*) são obtidos do banco de dados de estruturas 3D (PDB); (3) ângulos de torção do dupleto correspondente ao aminoácido central são calculados; (4) os ângulos de torção dos dupletos dos fragmentos molde são agrupados; (5) cada grupo é classificado segundo as regiões conformacionais do mapa de Ramachandran; (6) os ângulos de torção são representados na forma de intervalos; (7) a estrutura secundária da proteína é predita; (8) uma conformação inicial, representada na forma de intervalos, é obtida; (9) o intervalo da conformação é reduzido. A Figura 23 apresenta a estrutura básica e a seqüência de passos do método proposto. Estas etapas são descritas nas próximas seções.

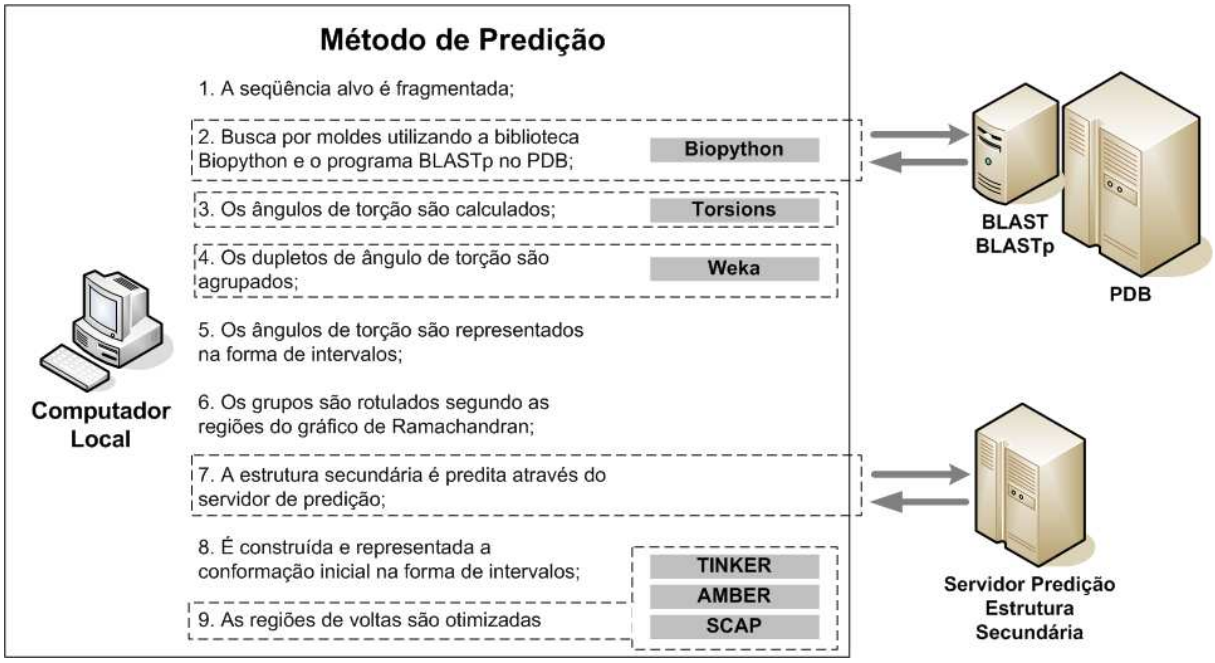

Figura 22 – Representação esquemática dos nove passos do método de predição desenvolvido. Os passos 2 e 7 utilizam servidores acessados remotamente. Os passos 1, 3-6, 8-9 são executados localmente.

4.3.1 Etapa 1: fragmentação da seqüência alvo

Nesta etapa, a seqüência alvo K de uma proteína é fragmentada em pequenos e subseqüentes fragmentos s_i com l resíduos de aminoácidos cada (Figura 23). O conjunto de todos os possíveis fragmentos de tamanho l de uma seqüência alvo K é representado por $S = \{s_i, s_{i+1}, \ldots, s_p\}$, onde s_i e s_p correspondem, respectivamente, ao primeiro e ao último fragmento. Sendo n o número de aminoácidos de uma seqüência alvo K e l um número ímpar para o tamanho de cada s_i fragmento, então o número p de possíveis fragmentos obtidos a partir de uma seqüência alvo é dada pela Equação 4.2.

$$p = [n - (l - 1)] \tag{4.2}$$

Um fragmento s_i inicia com o *i*-ésimo resíduo de aminoácido e termina com o *j*-ésimo resíduo formando um conjunto de tripletos consecutivos de ângulos de torção $\{(\omega_{i-1}, \phi_i, \psi_i), \ldots, (\omega_{j-1}, \phi_j, \psi_j)\}$ (Figura 21). A Figura 23 esquematiza a forma de fragmentação de uma seqüência alvo K com n resíduos de aminoácidos obtendo todos os seus subseqüentes s_i fragmentos com tamanho $l = 5$.

Considerando que o dupleto do aminoácido central é influenciado pelos aminoácidos vizinhos, define-se um valor ímpar para l. Esta escolha está diretamente ligada às fases de utilização de moldes do PDB para construção de intervalos de variação angular apresentados ao longo da descrição do método. Somente a informação dos dupletos do resíduo de aminoácido central dos fragmentos molde são considerados.

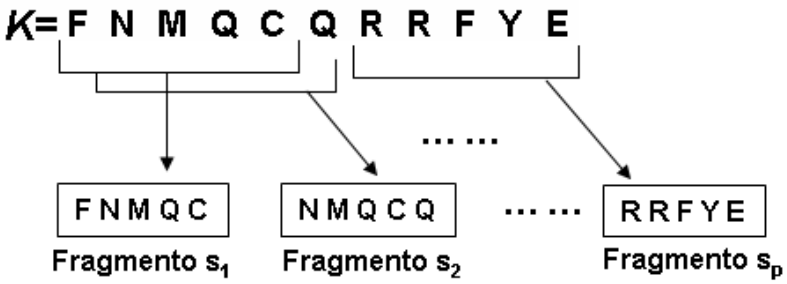

Figura 23 – Representação esquemática da forma de fragmentação de uma seqüência alvo K em p subseqüentes fragmentos s_i.

4.3.2 Etapa 2: busca por proteínas molde

Nesta etapa, para cada fragmento s_i de S são buscados fragmentos molde na base de dados do PDB (Figura 23, passo 2). Utiliza-se a versão *web* do programa BLASTp [3] para identificar fragmentos homólogos (moldes) ao fragmento alvo s_i. Somente os moldes de tamanho *l* igual ao fragmento s_i de proteínas molde, e que não possuam nenhuma relação evolucionária[1] com a seqüencia alvo K, são considerados para análise futura. Utilizou-se no programa BLASTp a matriz de substituição[2] BLOSUM 62[3] [34] para o cálculo da pontuação de cada alinhamento.

Para automatizar esta fase modificou-se a biblioteca BioPython[4] [14], através de desenvolvimento de funções para conexão com o PDB e análise de arquivos XML retornados de consultas ao PDB. Ao final desta etapa têm-se uma lista de códigos de acesso PDB correspondentes a cada fragmento s_i. Todos os arquivos pdbs retornados da busca são obtidos do banco de dados de estruturas 3D (PDB) e são armazenados localmente.

4.3.3 Etapa 3: cálculo dos ângulos de torção dos dupletos

Para cada fragmento alvo s_i de S, um conjunto de arquivos pdb molde é obtido do PDB (Figura 23, passo 3). A partir deste ponto são calculados, através do programa *Torsions* (Grupo do Dr. Andrew C. R. Martim)[5], os ângulos de torção do aminoácido central de cada fragmento-molde associados a um fragmento s_i. Cada dupleto de ângulos de torção é representado por uma tupla $t_i = (\phi, \psi)$, compostas pelos ângulos *phi* e *psi*, respectivamente.

[1]Considera-se como relacionadas evolucionariamente aquelas proteínas que apresentem homologia de seqüência igual ou superior à 70% em relação a seqüência alvo K.

[2]Matriz de Substituição: contém valores proporcionais à probabilidade de um aminoácido i mutar a um aminoácido j para todos os pares de aminoácidos. Ela fornece uma medida quantitativa para a substituição de aminoácidos em um alinhamento de seqüências.

[3]BLOSUM62: ("BLocks SUbstitution Matrix at 62%") foi a primeira matriz de logaritmos de probabilidades derivada das substituições de aminoácidos entre segmentos de seqüências de identidade igual ou superior à 62%.

[4]BioPython: o projeto BioPython trata-se de uma associação internacional de desenvolvedores de ferramentas de código aberto na linguagem Python para a biologia molecular computacional e a Bioinformática.

[5]Grupo do Dr. Andrew C. R. Martim: http://www.bioinf.org.uk/

A partir desta fase, cada fragmento s_i passa a ser visto como um conjunto de dupletos de ângulos de torção molde $s_i = \{t_i, t_{1+1} \ldots, t_p\}$, onde t_i e t_p correspondem ao primeiro e ao último dupleto de ângulos de torção do aminoácido central de um fragmento molde, respectivamente. Ao final desta etapa têm-se $S = \{s_i = \{t_i, t_{i+1}, \ldots, t_p\}, s_{i+1} = \{t_i, t_{i+1}, \ldots, t_p\}, \ldots, s_p = \{t_i, t_{i+1}, \ldots, t_p\}\}$ representando o conjunto S de todos os fragmentos alvos e seus correspondentes dupletos molde.

4.3.4 Etapa 4: agrupamento de dupletos

Nesta etapa todas as tuplas t_i de um fragmento s_i são submetidos a um processo de agrupamento. São identificas as regiões, onde estão concentradas as tuplas-molde no mapa de Ramachandran (Figura 24). O agrupamento é realizado através do método probabilístico EM (Expectation Maximization), o qual se baseia na análise de diferentes distribuições de probabilidades, uma para cada grupo, onde se busca identificar o conjunto de grupos mais favoráveis dado um conjunto de dados. O algoritmo EM (algoritmo não supervisionado) inicia realizando o agrupamento dos dupletos tendo como base o algoritmo de k-médias [57], obtendo, assim, uma solução inicial. O algoritmo k-médias minimiza uma função de erro quadrático conforme descrito na Equação 4.3, onde existem f grupos k_i, $i = 1, 2, .., f$ e $m(k_i)$ é o ponto médio de todos os dupletos em $t_j \in k_i$. Após determinada a solução inicial, as probabilidades de cada grupo são calculadas para cada dupleto t_i (fase de *Expectation*). A partir das probabilidades são calculados os parâmetros de distribuição, a partir dos quais, é realizada a "maximização"(*Maximization*) das distribuições de probabilidades. Uma descrição mais detalhada do algoritmo EM pode ser encontrada em [94].

$$V = \sum_{i=1}^{f} \sum_{t_j \in k_i} |t_j - m(k_i)|^2 \tag{4.3}$$

O valor médio $m(k_i)$ dos n dupletos t_j de um grupo k_i, $j = 1, 2, \ldots, n$ é obtido pela Equação 4.4.

$$m(k_i) = \frac{1}{n} \sum_{j=1}^{n} t_j \tag{4.4}$$

Após identificados os f k_i grupos para o conjunto de dupletos t_i de s_i é calculada a média e o desvio padrão estimado de cada ângulo diedro (*phi*, *psi*) em um grupo k_i. Sendo θ um dos ângulos diedros *phi* ou *psi*, a média ($m(k_i, \theta)$) entre os ângulos $\theta \in k_i$ é obtida através da Equação 4.5, onde n é o número de tuplas no grupo k_i.

$$m(k_i, \theta) = \frac{1}{n} \sum_{j=1}^{n} t[\theta]_j \tag{4.5}$$

O desvio padrão estimado ($\sigma(k_i, \theta)$) dos ângulos $\theta \in k_i$ é calculado a partir da média aritmética dos seus t_j elementos através da Equação 4.6. Um valor elevado de desvio padrão indica que os elementos analisados estão distantes da média $m(k_i, \theta)$. Um valor baixo indica que os elementos do grupo estão concentrados em uma região próxima à média.

$$\sigma(k_i, \theta) = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(t[\theta]_j - m(k_i, \theta))^2} \tag{4.6}$$

Durante a fase de agrupamento, utiliza-se um valor de $f = 4$ para o número de grupos a serem identificados. Este número de grupos é fixado levando em consideração o fato de, que apesar de os ângulos ϕ e ψ possuírem uma alta liberdade de rotação, existem combinações entre estes que não são favoráveis estéreoquimicamente. Estas regiões favoráveis são descritas pelo mapa de Ramachandran (Seção 2.3.3, Figura 10) [74] e ocorrem em número igual a quatro (em um dos 4 quadrantes em que o mapa é dividido). A primeira região, corresponde à região de hélices α. A segunda região, é a região de folhas β. A terceira região é a região de hélices com sentido à esquerda. O restante da área corresponde a 36% da área total do mapa, porém abriga somente 1.9% dos aminoácidos, sendo esta a quarta região [37, 49]. Esta região abriga geralmente os resíduos de aminoácidos presentes em estruturas irregulares como voltas e alças (Seção 2.3.3, Figura 10). No entanto, por serem estruturas secundárias irregulares, as voltas também podem apresentar resíduos de aminoácidos presentes em regiões de estruturas secundárias regulares como hélice α e folhas β.

Como regra geral utilizou-se um padrão de 4 regiões (4 grupos) para atender o caso em que as tuplas associadas a um fragmento pudessem estar ocupando, simultâneamente, estas regiões no mapa de Ramachandran. Para o caso em que todas as tuplas t_i estejam ocupando uma única destas 4 regiões no mapa de Ramachandran, são igualmente identificados 4 grupos. Com isto, busca-se identificar a sub-região em que há uma maior concentração de tuplas.

Ao final desta etapa, têm-se para cada fragmento s_i, um conjunto de 4 ($f = 4$) grupos k_i descritos cada um por uma $4-upla$, $k_i = (m\phi, \sigma\phi, m\psi, \sigma\psi)$ e $s_i = \{k_i = (m\phi, \sigma\phi, m\psi, \sigma\psi), \ldots, k_f = (m\phi, \sigma\phi, m\psi, \sigma\psi)\}$, onde $i = 1, 2, \ldots, f$ em k, e $m\phi$, $m\psi$, $\sigma\phi$, $\sigma\psi$ representam, respectivamente, a média e o desvio padrão estimado dos ângulos ϕ e ψ de um i-ésimo grupo de s_i. Para automatizar o processo de agrupamento utilizou-se o pacote *Weka* [94] para mineração de dados .

A Figura 24 ilustra a identificação de 4 grupos k_i no conjunto de tuplas t_i pertencentes a um fragmento s_i (Fragmento FNMQC (A), NMQCQ (B) e MQCQR (C)).

A informação obtida de cada grupo $k_i \in s_i$ (média e desvio padrão estimado) é utilizada para obtenção de intervalos de variação angular que representarão cada grupo grupo k_i, conforme será discutido na próxima seção.

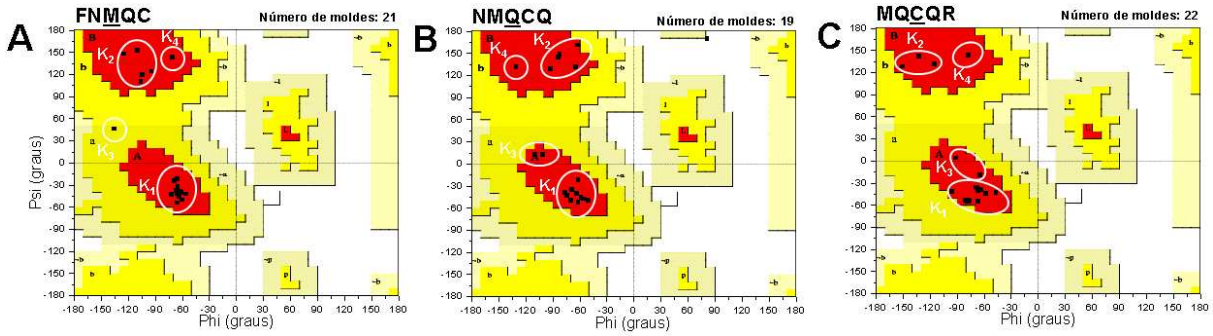

Figura 24 – Representação esquemática mostrando a identificação de k_i grupos no mapa de Ramachandran: (A) grupos identificados a partir das tuplas obtidas do resíduo central de aminoácido dos fragmentos-molde retornados para o fragmento alvo FNMQC; (B) grupos identificados a partir das tuplas obtidas do resíduo central de aminoácido dos fragmentos-molde retornados para o fragmento alvo NMQCQ; (A) grupos identificados a partir das tuplas obtidas do resíduo central de aminoácido dos fragmentos-molde retornados para o fragmento alvo MQCQR.

4.3.5 Etapa 5: representação dos ângulos de torção na forma de intervalos

Nesta etapa, são construídos intervalos de variação para os ângulos de torção de cada grupo k_i . Um intervalo fechado X nos números $\mathbb{R}$ é representado na forma $[X] = [\underline{x}, \overline{x}]$, onde $\underline{x}$ e $\overline{x}$ representam, respectivamente, o limite inferior e o limite superior do intervalo $[X]$. Um valor a está presente no intervalo $[X]$ quando $[a \in \mathbb{R} | \underline{x} \leq a \leq \overline{x}]$ [1]. Operações podem ser realizadas sobre estes conjuntos de intervalos. Se $\mathbf{x} = [\overline{x}, \underline{x}]$ e $\mathbf{y} = [\overline{y}, \underline{y}]$, então as quatro operações básicas para a aritmética intervalar obedecem $\mathbf{x} \, op \, \mathbf{y} = \{x \, op \, y \, | \, x \in \mathbf{x} \, e \, y \in \mathbf{y}\}$ para $op \in \{+, -, \times, \div\}$, sendo elas representadas da seguinte forma:

Subtração: $\mathbf{x} - \mathbf{y} = [\underline{x} - \overline{y}, \overline{x} - \underline{y}]$;
Adição: $\mathbf{x} + \mathbf{y} = [\underline{x} + \underline{y}, \overline{x} + \overline{y}]$;
Multiplicação: $\mathbf{x} \times \mathbf{y} = [min\,\{\underline{xy}, \underline{x}\overline{y}, \overline{x}\underline{y}, \overline{xy}\}, max\{\underline{xy}, \underline{x}\overline{y}, \overline{x}\underline{y}, \overline{xy}\}]$;
Divisão: $\mathbf{x} \div \mathbf{y} = \mathbf{x} \times \frac{1}{\mathbf{y}}$.

A partir do valor médio $m(k_i, \theta)$ e do valor de desvio padrão estimado $\sigma(k_i, \theta)$ calculado a partir dos dupletos (*phi* e *psi*) de um grupo k_i são construídos intervalos de variação angular. O intervalo de um ângulo diedro é representado por $[\theta] = [\underline{\theta}, \overline{\theta}]$, onde θ representa o intervalo para o ângulo ϕ ou o ângulo ψ, $\underline{\theta}$ e $\overline{\theta}$ representam, respectivamente, o limite inferior e o limite superior do intervalo de um ângulo de torção $[\theta]$.

O limite inferior $\underline{\theta}$ para um intervalo $[\theta]$ de um grupo k_i é obtido pela diferença entre a média $m(k_i, \theta)$ e o desvio padrão estimado $\sigma(k_i, \theta)$ (Equação 4.7). O limite superior $\overline{\theta}$ para um intervalo $[\theta]$ de um grupo k_i é obtido pela soma da média $m(k_i, \theta)$ e do desvio padrão estimado $\sigma(k_i, \theta)$ (Equação 4.8) de seus elementos t_j.

$$\underline{\theta} = m(k_i, \theta) - \sigma(k_i, \theta) \tag{4.7}$$

$$\overline{\theta} = m(k_i, \theta) + \sigma(k_i, \theta) \tag{4.8}$$

O tamanho w de um intervalo $[\theta]$ é dado pela Equação 4.9. O valor central c de um intervalo θ é obtido pela Equação 4.10.

$$w([\theta]) = \overline{\theta} - \underline{\theta} \tag{4.9}$$

$$c([\theta]) = \underline{\theta} + \frac{w([\theta])}{2} \tag{4.10}$$

Para cada grupo k_i de s_i são construídos intervalos para $[\phi]$ (*phi*) e $[\psi]$ (*psi*). Cada fragmento s_i passa a ser representado, a partir desta etapa, por grupos de intervalos de variação para ϕ e para ψ, $s_i = \{k_i = (\underline{\phi}, \overline{\phi}, \underline{\psi}, \overline{\psi}), \ldots, k_f = (\underline{\phi}, \overline{\phi}, \underline{\psi}, \overline{\psi})\}$, onde $\underline{\phi}$, $\overline{\phi}$, $\underline{\psi}$ e $\overline{\psi}$ representam, respectivamente, o limite inferior e o limite superior para ϕ e ψ e f representa o número de k grupos.

A seguir é apresentado um exemplo de criação destes intervalos:
Seja k_1 um grupo associado a um fragmento s_i, onde para ϕ:
$m(k_1, \phi) = -75.0$,
$\sigma(k_1, \phi) = 15.0$,

e para ψ:
$m(k_1, \psi) = -30.0$,
$\sigma(k_1, \psi) = 20.0$,

então o limite inferior para $\underline{\phi}$ e para $\underline{\psi}$ é dado por:
$\underline{\phi} = m(k_1, \phi) - \sigma(k_1, \phi)$,
$\underline{\phi} = -75.0 - 15.0 = -90.0$,
$\underline{\psi} = m(k_1, \psi) - \sigma(k_1, \psi)$,
$\underline{\psi} = -30.0 - 20.0 = -50.0$,

e o limite superior para $\overline{\phi}$ e para $\overline{\psi}$ é:
$\overline{\phi} = m(k_1, \phi) + \sigma(k_1, \phi)$,
$\overline{\phi} = -75.0 + 15.0 = -60.0$,
$\overline{\psi} = m(k_1, \psi) + \sigma(k_1, \psi)$,
$\overline{\psi} = -30.0 + 20.0 = -10.0$.

O limite inferior e superior do intervalo de cada ângulo diedro (*phi* e *psi*) representam uma área limitada de variação dos ângulos (tuplas-molde) de um grupo k_i no mapa de Ramachandran. O intervalo formado pelo ponto inferior $P_1 = (\underline{\phi}, \underline{\psi})$ e pelo ponto superior $P_2 = (\overline{\phi}, \overline{\psi})$ pode ser representado como uma região delimitada no mapa de Ramachandran (Figura 25 B). Esta região representa a variação das tuplas-molde de um grupo k_i associado a um fragmento s_i (Figura 25 A).

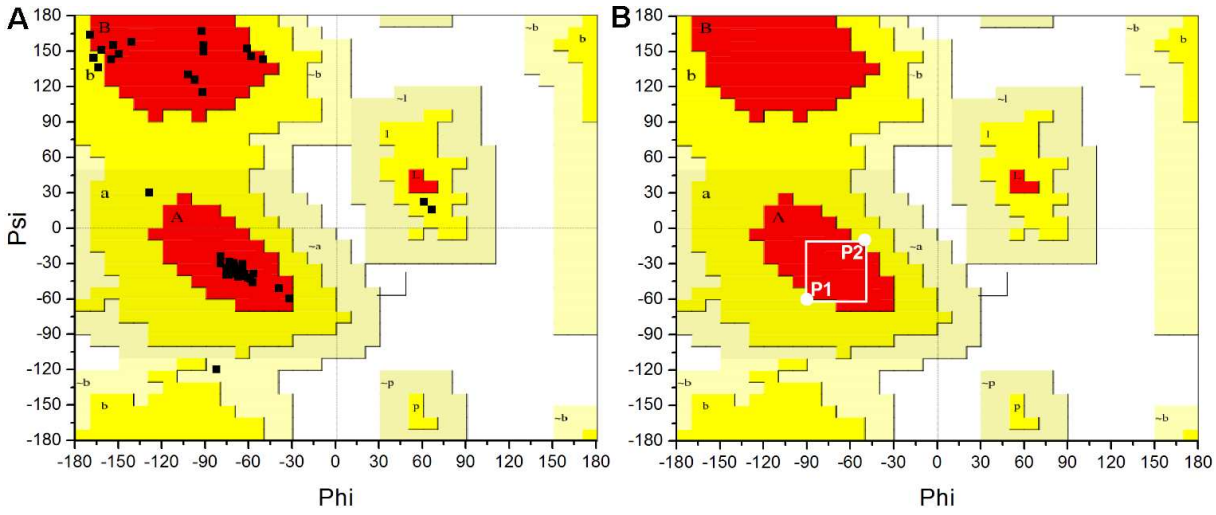

Figura 25 – Mapas de Ramachandran: (A) tuplas ocupando as regiões do mapa de Ramachandran, a região mais favorável é expressa em vermelho, a região permitida é expressa em amarelo, a região ainda aceitável é apresentada em amarelo claro e a região não permitida em branco. A região em vermelho no canto superior esquerdo representa a região de estrutura secundária regular do tipo folhas β paralelas e anti-paralelas. A região em vermelho no centro esquerdo, e no centro direito representam a região de estrutura secundária regular do tipo hélice α a direita e a esquerda respectivamente. (B) representa a delimitação de um intervalo entre os ângulos de torção mínimos e máximos de ϕ e ψ, através dos pontos P_1 e P_2.

A seguir é exemplificado o cálculo do tamanho e do valor central de um intervalo:

Assumindo,
$\underline{\phi} = -90.0$, $\overline{\phi} = -60.0$,
$\underline{\psi} = -50.0$, $\overline{\psi} = -10.0$,

o tamanho do intervalo do ângulo ϕ é:
$w([\phi]) = \overline{\phi} - \underline{\phi}$,
$w([\phi]) = -60 - (-90)$,
$w([\phi]) = 30$,

e o tamanho do ângulo ψ é:
$w([\psi]) = \overline{\psi} - \underline{\psi}$,
$w([\psi]) = -10 - (-50)$,
$w([\psi]) = 40$.

O valor central do intervalo do ângulo ϕ é dado por:
$c([\phi]) = \underline{\phi} + \frac{w([\phi])}{2}$,
$c([\phi]) = -90 + (30/2)$,
$c([\phi]) = -75$,

e valor central do intervalo do ângulo ψ é dado por:
$c([\psi]) = \underline{\psi} + \frac{w([\psi])}{2}$,
$c([\psi]) = -50 + (40/2)$,
$c([\psi]) = -30$.

4.3.6 Etapa 6: classificação dos grupos em regiões ocupadas no mapa de Ramachandran

Após representar cada grupo k_i de s_i na forma de intervalos de variação angular, estes são rotulados. A partir da ponto médio $m(k_i, \phi)$ e do ponto médio $m(k_i, \psi)$ de um grupo k_i, é criado um rótulo de identificação para o mesmo. Este rótulo tem a função de relacionar um grupo k_i com a região que este ocupa no mapa de Ramachandran. Todos os grupos $k_i \in S$ são rotulados.

Para que esta classificação se torne possível, criou-se uma biblioteca mapeando as regiões mais favoráveis estéreoquimicamente identificadas pela combinação de ângulos ϕ e ψ. Esta biblioteca foi criada tendo por base os trabalhos de Thornton e colaboradores [62] [48], nos quais o mapa de Ramachandran é dividido em 11 regiões preferenciais (Figura 25 A). No entanto, por simplificação, reduziu-se o número de regiões possíveis no mapa de Ramachandran para 8 regiões. Estas 8 regiões são identificadas por: A, B, L, a, b, l, p e c, onde c passa a ser chamada de região de volta e representa as regiões ∼a, ∼b, ∼l, ∼p e o restante da área não favorável no mapa de Ramachandran. A Tabela 3 apresenta o código de identificação e a descrição das 11 regiões em que o mapa de Ramachandran é dividido, segundo Thornton e colaboradores [62] [48].

A partir da biblioteca criada, realizando o mapeamento das regiões do mapa de Ramachandran, foi criada uma função de classificação, onde a partir do valor do centro do intervalo de *phi* e do centro do intervalo de *psi* de um grupo k_i é realizada a atribuição de rótulos de identificação para cada k_i representado por um intervalo de variação.

Um grupo rotulado tem a forma $k_i : rot$, onde rot é o rótulo de identificação da região, podendo este assumir uma das oito regiões conformacionais definidas. Cada fragmento s_i de S passa a ser representado como $s_i = \{k_1 : rot, k_2 : rot, k_3 : rot, k_4 : rot\}$.

Após rotulados, os grupos k_i de cada s_i são ordenados pelo número de elementos t_j molde associados a cada grupo. A ordenação é feita do grupo k_i com maior número de elementos t_j para o grupo com o menor número, como $s_i = \{k_i : rot > k_{i+1} : rot >, \ldots, k_f : rot\}$. Esta ordenação (grupos k_i por número de elementos t_j) é importante para a construção da estrutura inicial da proteína-alvo nas etapas seguintes do método.

Tabela 3: Regiões conformacionais do mapa de Ramachandran. Descrição segundo Thornton e colaboradores [62] [48].

Código da região	Ocorrência	Descrição
A	região mais favorável	hélice-α
B	região mais favorável	folha-β
L	região mais favorável	hélice-α à esquerda
a	região favorável	hélice-α
b	região favorável	folha-β
l	região favorável	hélice-α à esquerda
p	região favorável	extensão de hélice-α
$\sim$a	região aceitável	hélice-α
$\sim$b	região aceitável	folha-β
$\sim$l	região aceitável	hélice-α à esquerda
$\sim$p	região aceitável	extensão de hélice-α
restante da área	região não permitida	exceto para a Glicina

Uma concentração maior de elementos t_i de um fragmento s_i em um grupo k_i significa que os dupletos do aminoácido central dos fragmentos-molde pertencentes a s_i ocorrem em maior número no intervalo $k_i = (\underline{\phi}, \overline{\phi}, \underline{\psi}, \overline{\psi})$ e os ângulos diedros do dupleto do aminoácido central de um fragmento-alvo s_i estão mais propícios a estarem neste intervalo. Um grupo k_i com um maior número de elementos molde possuí uma maior probabilidade de ser utilizado na construção da conformação, representada na forma de intervalos, da seqüência-alvo.

4.3.7 Etapa 7: predição da estrutura secundária

Nesta etapa, é realizada a predição da estrutura secundária da seqüência alvo K. Para cada resíduo de aminoácido i, da seqüência-alvo K, é determinada a região em que os ângulos de torção (phi e psi) deste aminoácido possivelmente estarão ocupando no mapa de Ramachadran. Na literatura são encontrados diversos métodos que realizam a predição da estrutura secundária de uma proteína a partir de sua seqüência de aminoácidos. Dentre os mais utilizados é possível citar: DSC [46], PHD [78], PREDATOR [24], GOR [25, 26, 30], SIMPA96 [52], DPM [21], SOPMA [28], SOPM [27], MLRC [33] (combina vários métodos de predição, tais como SIMPA96 [52], GOR4 [25, 26, 30], SOPMA [28]) e SCRATCH [15]. Estes métodos são disponibilizados em servidores de predição e estão disponíveis para uso gratuito. Exemplos destes servidores: o servidor NPS@ (NPS@: *Network Protein Sequence @nalysis*) [16] e o servidor SCRATCH: *protein prediction server* que utiliza o método SCRATCH [15].

Cada método de predição da estrutura secundária de uma proteína faz uso de diferentes

técnicas para realizar as suas predições. Alguns métodos analisam a formação de ligações de hidrogênio, outros incorporam padrões oriundos de bases de dados com informações de proteínas com estrutura determinada experimentalmente, outros utilizam redes neurais artificiais, etc. Devido a estas diferentes metodologias, alguns métodos de predição de estruturas secundárias apresentam melhores resultados quando comparados com outros em suas predições. Alguns, conseguem predizer melhor as estruturas secundárias do tipo folhas β outros de hélices α. Após analisar e avaliar a qualidade das predição realizadas por diferentes métodos, optou-se por obter um consenso entre a predição de estrutura secundária obtida a partir de três ou mais métodos de predição. O consenso, busca mapear regiões (com um ou mais aminoácidos) em que há uma concordância no resultado da predição da estrutura secundária pelos métodos utilizados. A Tabela 4 mostra um exemplo do consenso obtido a partir da predição realizada por 3 diferentes métodos (DSC [46], PHD [78], PREDATOR [24]) para uma seqüência alvo K=FNMQCQRRFYEALHDPNLNEEQRNAKIKSIRDDC composta por 34 aminoácidos.

Tabela 4: Exemplo da predição da estrutura secundária de uma seqüência de 34 resíduos de aminoácidos obtendo o consenso entre os métodos DSC, PHD e PREDATOR.

Método	FNMQCQRRFYEALHDPNLNEEQRNAKIKSIRDDC
DSC	ccchhhhhhhhhhhccccchhhhhhhhhhhccccc
PHD	ccchhhhhhhhhhhcccccchhhhchhhhhhcccc
PREDATOR	ccchhhhhhhhhhhccccccchhhhhhhhhhhccccc
Consenso	ccchhhhhhhhhhhccccchhhhhhhhhhhccccc

Os métodos de predição de estruturas secundárias dividem o mapa de Ramachandran em oito estados conformacionais. Por simplificação, neste trabalho estados foram representadas em somente três: h (hélice α), b (folha β) e c (volta ou alça). Portanto, criou-se uma tabela de associação para converter os oito estados conformacionais para os três estados conformacionais. A Tabela 5 apresenta a codificação (rótulos) adotada (três estados conformacionais) para representação da estrutura secundária de um polipeptídeos e a codificação correspondente adotada pelos métodos de predição de estrutura secundária utilizados (oito estados conformacionais).

Tabela 5: Representação dos estados conformacionais no servidor NPS@, servidor Scratch e a codificação correspondente aos três estados adotado pelo método de predição. h representa o estado conformacional do tipo hélice α, b representa o estado conformacional do tipo folha β e c representa o estado conformacional de volta ou alça.

Servidor NPS@	Servidor SCRATCH	Descrição	Três estados
H, G, I	H, G, I	hélice α	h
B, E	B, E	folha β	b
T, S, C	T , S , C	voltas	c

Esta informação, obtida durante a predição da estrutura secundária, é utilizada para guiar a escolha de intervalos de variação angular para a construção da conformação inicial em etapas seguintes do método.

4.3.8 Etapa 8: construção da conformação inicial

Após identificados e rotulados os grupos k_i e realizada a predição da estrutura secundária da seqüência da proteína alvo, é construída a conformação inicial do polipeptídeo. Os dupletos de ângulos de torção de cada resíduo de aminoácido da seqüência-alvo K do polipeptídeo são representados por intervalos de variação angular. Desta forma, a conformação C de uma proteína passa a ser representada por um vetor $C = \{[X_1], [X_2], \ldots, [X_n]\}$, onde $[X_i]$ é um dupleto de ângulos de torção representado como $X_i = \{\underline{x}, \overline{x}\}$, ou seja, $X_i = \{(\underline{\phi}, \overline{\phi}), (\underline{\psi}, \overline{\psi})\}$. Devido à restrição existente em suas rotações, os ângulos ω (ômega) são fixados em 180° (Seção 2.3.1). A Figura 26 representa um modelo de peptídeo na forma de intervalos de variação angular, exceto os ângulos ω.

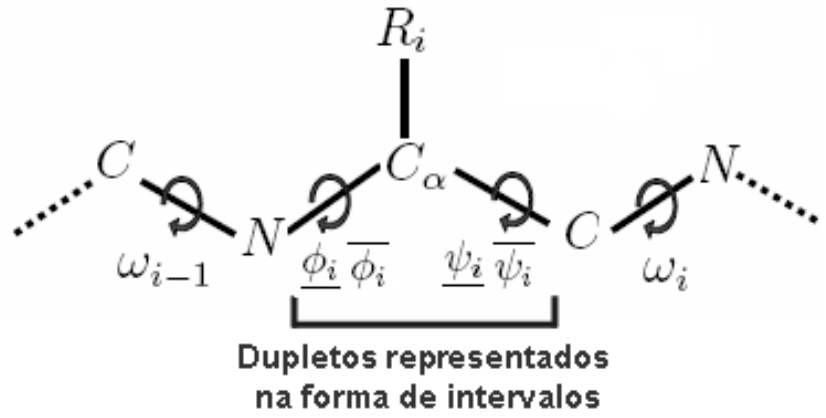

Figura 26 – Representação esquemática de um modelo de peptídeo identificando os dupletos (ϕ, ψ) representados na forma de intervalos de variação angular.

O aminoácido central de cada fragmento consecutivo s_i de K corresponde a um único aminoácido da seqüência-alvo K, isto é, para $l = 5$ o aminoácido central do primeiro fragmento s_1 de uma seqüência K corresponde à posição $i + 3$ em K, o aminoácido central do segundo fragmento s_2 à posição $i + 4$ e assim por diante. Para a construção da conformação inicial, representada na forma de intervalos, são utilizadas as informações dos grupos k_i de cada fragmento s_i. A escolha de um $k_i = (\underline{\phi}, \overline{\phi}, \underline{\psi}, \overline{\psi})$ grupo de s_i para representar o i-ésimo resíduo de aminoácido de uma seqüência K é feita seguindo 2 regras:

Regra 1: encontrar em s_i o(s) grupo(s) k_i em que rot seja igual ao rótulo identificado para o i-ésimo resíduo de aminoácido no consenso da predição da estrutura secundária. E dentre estes, escolher o grupo k_i que possui o maior número de tuplas t_i pertencentes ao mesmo.

Conforme discutido anteriormente, durante a predição da estrutura secundária é utilizado

um modelo de três estados conformacionais (h, b, c) para representar a estrutura secundária da proteína e, durante a fase de atribuição de rótulos as grupos k_i, são utilizados oito estados conformacionais (A, B, L, a, b, l, p e c). Para satisfazer a Regra 1 é necessário criar uma função que irá, a partir do estado conformacional de um aminoácido, determinado durante a predição da estrutura secundária, selecionar de maneira adequada o grupo k_i que melhor descreva o seu estado conformacional. Criou-se um conjunto de regras para escolha destes grupos, as mesmas são descritas a seguir:

- Para aminoácidos, identificados na fase de predição da estrutura secundária, com estrutura secundária em estado de hélice-α (h): são escolhidos primeiramente os grupos rotulados por "A"(região mais favorável), caso não exista nenhum grupo, procede-se escolhendo um grupo rotulado por "a"(região favorável);

- Para aminoácidos, identificados na fase de predição da estrutura secundária, com estrutura secundária em estado de folhas-β (b): são escolhidos primeiramente os grupos rotulados por "B"(região mais favorável), em seguida um grupo rotulado por "b"(região favorável);

- Para aminoácidos com estrutura secundária em estado de volta ou alça (c): são escolhidas primeiramente as regiões rotuladas por "c"($\sim$a, $\sim$b, $\sim$l, $\sim$p e restante da área), em seguida por "L"(região mais favorável), "l"(região favorável), "p"(região favorável), "a"(região favorável), "b"(região favorável), "B"(região mais favorável), "A"(região mais favorável). As regiões pertences a estruturas regulares do tipo hélices α e folhas β (A, a, B, b) são utilizadas para representar as regiões de volta. A inclusão destas regiões, pertencentes a estruturas regulares, se deve ao fato que estruturas irregulares podem ocorrer em qualquer região do mapa de Ramachandran [96].

Regra 2: se durante a predição da estrutura secundária for predito para o *i*-ésimo resíduo de aminoácido da seqüência-alvo K um estado conformacional pertencente a uma região de estrutura regular (h ou b), e não existir um grupo k_i com rótulo rot igual ao identificado para este resíduo, procede-se calculando o valor médio entre o ângulos presentes no resíduo de aminoácido $i-1$ e $i+1$ e o mesmo. O valor obtido é substituído no *i*-ésimo resíduo de aminoácido da seqüência-alvo K. Esta estratégia é justificada pelo fato de que nas estruturas secundárias regulares existe um padrão entre os valores dos ângulos diedros dos aminoácidos que a compõe.

Os intervalos de ϕ e ψ do grupo $k_i = (\underline{\phi}, \overline{\phi}, \underline{\psi}, \overline{\psi})$ de s_i selecionados passam a representar o dupleto de ângulos de torção do resíduo de aminoácido correspondente na seqüência-alvo K. A Figura 27 ilustra o processo de construção da conformação representada na forma de intervalos.

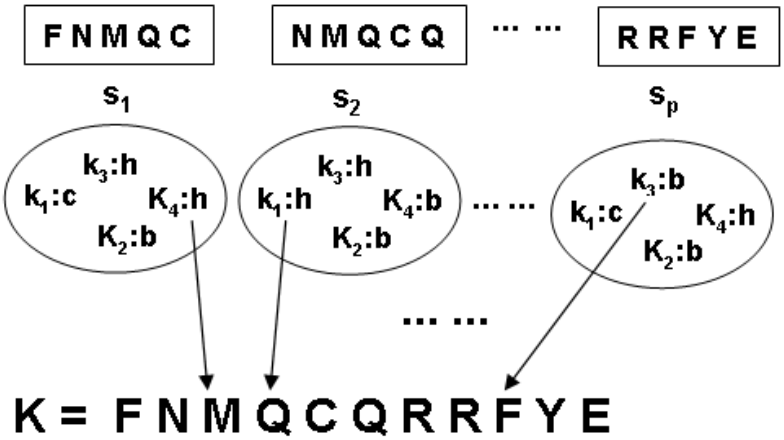

Figura 27 – Representação esquemática do processo de escolha dos k_i grupos para representar os ângulos de torção dos resíduos de aminoácidos da seqüência alvo K.

4.3.9 Etapa 9: otimização das regiões de volta

Após construída a conformação inicial (representada na forma de intervalos), aplica-se uma estratégia para redução do intervalo com o objetivo de encontrar a conformação que apresente a menor energia potencial. Nesta etapa, as regiões de voltas [6] (Figura 29) identificadas na predição da estrutura secundária têm um tratamento especial, pois são elas que irão, principalmente, determinar a forma de enovelamento da proteína, ou seja, a forma como as estruturas secundárias regulares (folhas β e hélices α) estarão organizadas em nível terciário [23]. Por representarem regiões pequenas (em média 3-7 resíduos de aminoácidos) de segmento na cadeia polipeptídica, estas regiões provocam distorções no enovelamento da proteína, as quais, são difíceis de serem avaliadas a *priori* [23]. Devido a esta particularidade, e ao fato das estruturas secundárias regulares apresentarem um certo padrão no valor dos seus ângulos de torção, optou-se por realizar a otimização e redução do intervalo da conformação inicial somente nas regiões identificadas como regiões de volta.

As regiões do polipeptídeo alvo identificadas durante a predição da estrutura secundária como folhas β ou hélices α não têm seu intervalo reduzido. Assume-se para estas regiões o centro do intervalo (do ângulo de torção), conforme foi descrito na Equação 4.10.

Buscando a redução do intervalo dos ângulos de torção dos resíduos de aminoácidos presentes em todas as regiões de volta, identificadas mediante a análise da predição da estrutura secundária, desenvolveu-se um algoritmo composto por 6 passos:

1. Inicialmente são identificadas as regiões de volta na seqüência alvo (predição da estrutura secundária);

[6] Região de volta: identificada por um único, por dois ou mais resíduos de aminoácidos consecutivos identificados por "c"(volta) durante a predição da estrutura secundária.

2. São escolhidos, aleatóriamente, os ângulos de torção pertencentes a uma determinada região de volta;

3. A partir destes ângulos aleatórios e do ponto médio do intervalo das demais regiões de volta e das regiões identificadas como pertencentes a estruturas regulares é construída a conformação do polipeptídeo-alvo;

4. As cadeias laterais são adicionadas de forma otimizada na conformação construída;

5. A energia potencial da conformação é calculada;

6. Os passos 2-4 são repetidos até um número ι de conformações seja atingido. Posteriormente, são analisadas as τ (%) melhores conformações construídas que apresentam a menor energia potencial. Em seguida, é calculado o valor médio de variação para cada ângulo de torção de cada resíduo de aminoácido pertencente à região de volta analisada. A partir deste valor médio, é feita a redução do intervalo de cada ângulo de torção de cada resíduo de aminoácido da região de volta que esta sendo processada. Um nova região de volta, se existente, é escolhida. O algoritmo repete os passos 2-6 até que todas as regiões identificadas como volta sejam processadas e até que seja atingido um critério de parada. Como critério de parada, estabeleceu-se o tamanho do intervalo de cada ângulo diedro dos resíduos de aminoácidos das regiões de volta. O algoritmo pára quando atinge um limite λ, não sendo mais possível reduzir o intervalo. Após processar todas as regiões de voltas, retorna-se à primeira região de volta processada, e novamente o algoritmo é executado buscando, caso possível, reduzir o intervalo.

O fluxograma da Figura 28 esquematiza a estrutura do método de redução do intervalo das regiões de volta. A seguir são discutidos em maiores detalhes cada um dos 6 passos que compõem o algoritmo para otimização das regiões de volta e redução de intervalos.

Passo 1 - identificação de regiões de volta: a partir da predição da estrutura secundária da seqüência alvo de aminoácidos são identificadas as regiões de volta. Uma região é formada por resíduos de aminoácidos consecutivos que possuem o mesmo estado conformacional. Uma região também pode ser formada por um único resíduo de aminoácido.

A Figura 29 ilustra a identificação de regiões a partir da predição da estrutura secundária da seqüência-alvo. Uma única região de volta é analisada a cada passo de execução do algoritmo, as demais regiões de voltas permanecem "bloqueadas", sendo utilizados o valor central de seus intervalos. Somente os resíduos de aminoácidos pertencentes a regiões de volta (Figura 29) tem o intervalo de seus ângulos diedros reduzido.

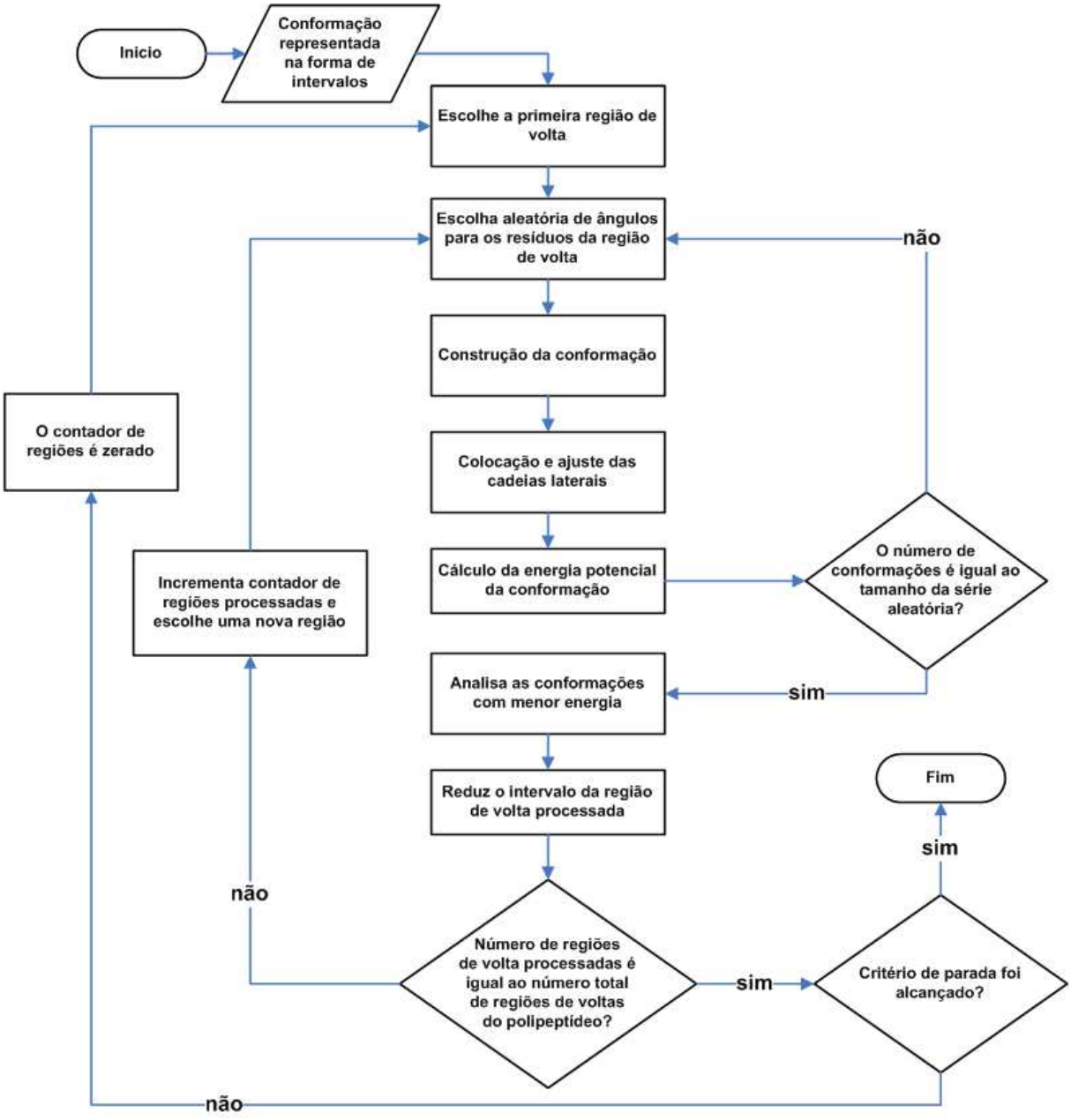

Figura 28 – Fluxograma esquematizando o método desenvolvido para a redução do intervalo nas regiões de volta do polipeptídeo representado na forma de intervalos de variação angular.

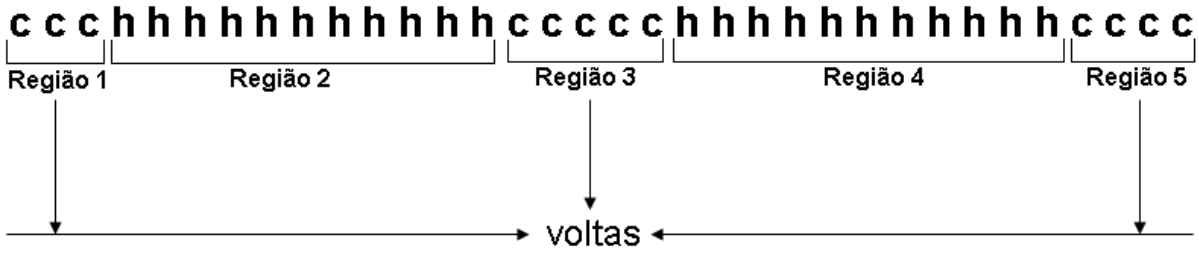

Figura 29 – Representação esquemática do processo de identificação das regiões de volta, a partir do resultado da predição da estrutura secundária de uma seqüência alvo K.

Passo 2 - escolha aleatória de ângulos de torção em um intervalo: devido à complexidade existente para analisar todas as possíveis combinações de ângulos de torção [65] entre os intervalos de todos os dupletos de cada resíduo de aminoácido da região de volta, optou-se pela geração de uma série pseudo-aleatória [7] [31] para guiar a escolha "aleatória"de ângulos em um intervalo. Optou-se pela geração de uma seqüência pseudo-aleatória e não de uma seqüência puramente aleatória, devido ao fato de que o uso de números realmente aleatórios torna impossível a repetição de uma dada seqüência de números e desta forma não é possível verificar a acurácia da simulação ou mesmo repetir o experimento uma ou mais vezes buscando corrigir algum erro no programa desenvolvido [40].

Implementou-se um método linear congruente [77] para a geração de variáveis pseudo-aleatórias, criando números uniformemente distribuídos entre 0 e 1. O método é descrito pela Equação 4.11 [77], onde a é um número inteiro escolhido entre 1 e M e M é um número primo. A Equação 4.11 procede realizando a multiplicação de a e Z e dividindo o resultado por M e fazendo Z_k o resto da divisão.

$$Z_k = a.Z_{k-1} \tag{4.11}$$

O algoritmo 1 gera uma seqüência (*resu*) pseudo-aleatória com tamanho de período fixo (*tamanhoPeriodo*). Assume-se um valor para $M = 2147483647$, $a = 25717$ e um valor para semente igual a 1 ($sem = 1$).

Algoritmo 1 Geração de uma série aleatória

```
numerosAleatorios(sem,tamanhoPeriodo)
resultado = []
z = 0, Za = sem, M = 2147483647, a = 25717
for (i=0 ate tamanhoPeriodo) do
   z = (a*Za)% M
   resu.add(Za/(M-1))
   Za = Z
end for
return resu
```

Com base na série pseudo-aleatória gerada são escolhidos ângulos junto ao intervalo da conformação. A Equação 4.12 obtém, com base em um valor da série periódica, um ângulo pertencente a um intervalo $\phi = (\underline{\phi}, \overline{\phi})$ e a um intervalo $\psi = (\underline{\psi}, \overline{\psi})$ de um resíduo de aminoácido.

$$\eta(\theta) = \underline{\theta} + (resultado[i].w([\theta])), \tag{4.12}$$

[7] Gerador de números pseudo-aletórios: é um algoritmo que usa da aritmética para gerar uma seqüência de números com propriedades de números aleatórios.

onde, θ é um ângulo de torção (ϕ ou ψ) e i é o indice do vetor $resultado$ que armazena a série pseudo-aleatória.

A seguir é exemplificada a escolha de um ângulo de torção em um intervalo:

Seja
$[\phi] = (\underline{\phi}, \overline{\phi}) = (-90.0, -50.0)$
$[\psi] = (\underline{\psi}, \overline{\psi}) = (-60.0, -10.0)$
e
$resultado[1] = 0.5$, *índice* $i = 1$

então o ângulo aleatório para o intervalo $[\phi]$ e $[\psi]$ de um resíduo de aminoácido em região de volta de uma seqüência K é:
$\eta(\phi) = -90.0 + (0.5 * 40.0) = -70.0$
$\eta(\psi) = -60.0 + (0.5 * 50.0) = -35.0$

Para todos os resíduos de aminoácidos correspondentes à região de volta identificada no consenso da predição da estrutura secundária são escolhidos, dentre os respectivos intervalos, ângulos baseados na distribuição pseudo-aleatória obtida através do Algoritmo 1.

Passo 3 - construção da conformação: após escolhidos os ângulos de torção nos intervalos associados a cada resíduo de aminoácido de uma determinada região de volta é construída a conformação. Para os dupletos de ângulos de torção (ϕ, ψ) de resíduos de aminoácidos classificados em regiões conformacionais de hélices α e folhas β durante a predição da estrutura secundária e de outras regiões de voltas, que no momento não estão sendo analisadas, é considerado o valor central $c([\theta])$ do intervalo na conformação inicial. Todos os ângulos ω são fixados em 180°. As informações sobres os ângulos do polipeptídeo são processadas pelo módulo *teLeap* do pacote para modelagem molecular AMBER [13], gerando a conformação.

Passo 4 - colocação e ajuste das cadeias laterais: após obtida uma conformação, são adicionadas de forma otimizada as cadeias laterais. A atribuição e ajuste das cadeias laterais é feita utilizando um banco de dados constituído por rotâmeros de cadeias laterais (ângulos χ) de proteínas do PDB. No método desenvolvido utilizou-se o programa SCAP [39,95] para automatizar a colocação e otimização das cadeias laterais nas conformações construídas. A partir de uma biblioteca de rotâmeros (bilioteca de Dunbrack [22]) o programa SCAP estima a probabilidade de cada cadeia lateral assumir uma determinada conformação em função dos ângulos torsionais da cadeia principal (ϕ e ψ) e do tipo de resíduo de aminoácido. O software SCAP foi utilizado devido a qualidade de suas predições e por considerar fatores como acessibilidade do solvente e hidrofobicidade durante a colocação das cadeias laterais.

Passo 5 - cálculo da energia potencial: após adicionadas as cadeias laterais é calculada a energia potencial da conformação. Esta energia é obtida através de rotinas do pacote para modelagem molecular TINKER [73] e o campo de força CHARMM (versão 27) [11, 56]. O valor da energia potencial é a base para proceder a redução do intervalo.

Passo 6 - redução do intervalo: para cada ciclo de execução do algoritmo é gerado um número ι de conformações igual ao tamanho da série periódica (*tamanhoPeriodo*) obtida pelo Algoritmo 1. A partir das melhores energias obtidas nas conformações geradas para uma dada região de volta, é feita a redução do intervalo para esta região. A partir de uma porcentagem δ das ι (Algoritmo 1, *tamanhoPeriodo*) conformações geradas são obtidos parâmetros que irão conduzir a redução do intervalo de cada ângulo do dupleto de cada resíduo de aminoácido da região de volta que esta sendo analisada.

Dentre todas as δ conformações selecionadas é calculada a média aritmética para cada ângulo de torção de cada resíduo de aminoácido pertencente a uma região de volta. A Equação 4.13 calcula a média aritmética dos ângulos θ de um resíduo de aminoácido i das δ estruturas com menor energia potencial encontradas no passo de simulação corrente.

$$m(\theta) = \frac{1}{\delta} \sum_{i=1}^{\delta} \theta_i \tag{4.13}$$

Para o intervalo de um ângulo diedro θ (ϕ ou ψ) de um resíduo de aminoácido i na seqüência alvo K presente em uma região de volta é realizada a redução do intervalo deste ângulo conforme os passos descritos a seguir:

- Calcula-se o centro do intervalo $c([\theta])$;
- Calcula-se a média aritmética deste ângulo nas δ estruturas geradas que apresentam a menor energia potencial;
- A partir de $c([\theta])$, incluindo um limiar para mais ou para menos (definiu-se empiricamente 10% para mais e para menos a partir do centro do intervalo) é verificado se o valor da média $m(\theta)$ das estruturas com menor energia está concentrada próximo ao limite inferior ou próximo ao limite superior do intervalo. Caso esteja concentrada próxima ao limite inferior, procede-se reduzindo um percentual ν do tamanho do intervalo a partir do limite superior. Caso esteja concentrada em uma região próxima ao limite superior, procede-se reduzindo um percentual ν do tamanho do intervalo a parir do limite inferior. Caso a média dos ângulos das δ estruturas com menor energia potencial esteja localizada na região do limiar, procede-se a redução do intervalo com $\frac{\nu}{2}$ do tamanho do intervalo para o limite superior e o limite inferior do intervalo. A Figura 30 representa um intervalo e identifica a região de limiar próxima a região central a partir da qual será analisada e identificada a melhor forma para proceder a redução do intervalo.

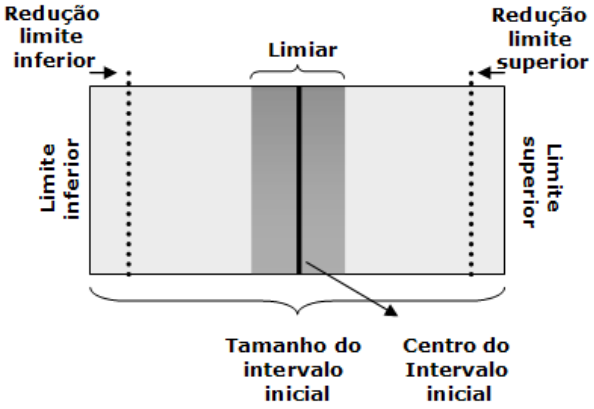

Figura 30 – Representação esquemática do processo de redução de um intervalo, identificando o limite superior e inferior do intervalo e o limiar a partir do centro do intervalo.

A cada passo de iteração do método de otimização da(s) região(ões) de volta, o intervalo de cada ângulo diedro, dos aminoácidos da volta processada, é reduzido. Um passo de iteração do algoritmo significa que as ι conformacões foram geradas. Calcula-se o tamanho de cada intervalo e o algoritmo de otimização e redução pára quanto o tamanho ($w([\theta])$) de todos os intervalos forem menores ou iguais a um limite ϵ.

4.3.10 Implementação

Nesta seção é brevemente descrito cada módulo implementado para automatizar o método desenvolvido. Todos os módulos foram implementados na linguagem Python.

1. **Módulo de fragmentação:** realiza a fragmentação da seqüência-alvo em subseqüentes fragmentos.

2. **Módulo para busca de fragmentos moldes no PDB:** realiza a busca por proteínas moldes para todos os subseqüentes fragmentos utilizando o pacote BioPython e o programa BLASTp e cria uma lista de pdbs a serem obtidos do PDB.

3. **Módulo para obter os arquivos pdbs do PDB:** obtêm os arquivos pdbs do PDB.

4. **Módulo para calcular os ângulos de torção:** calcula os ângulos de torção de cada arquivo pdb obtido do PDB e cria o arquivo de ângulos referente a este pdb.

5. **Módulo para localizar os fragmentos molde:** analisa cada arquivo de ângulo criado e localiza neste arquivo, o fragmento molde refere ao fragmento alvo s_i associado a este pdb.

6. **Módulo para gerar os arquivo de variação angular:** a partir de cada fragmento molde é obtidos os valores de torção do resíduo de aminoácido deste fragmento. Todos os valores

de torção associados a um fragmento s_i são armazenados em um arquivo de variação angular.

7. **Módulo para gerar arquivos *Weka*:** a partir do arquivo de variação angular, associado a cada fragmento s_i, é criado o arquivo de entrada para o programa de mineração de dados - *Weka*.

8. **Módulo para execução do *Weka*:** o algoritmo de agrupamento é executado tendo como entrada cada arquivo de entrado gerado pelo módulo para gerar arquivos weka. O resultado do agrupamento é armazenado em um arquivo.

9. **Módulo para geração de intervalos de variação:** a partir dos resultados do algoritmo de agrupamento, para cada um dos grupos associados a um fragmento s_i são gerados intervalos de variação.

10. **Biblioteca de estados conformacionais:** biblioteca representando os estados conformacionais do mapa de Ramachandran.

11. **Função de mapeamento de regiões conformacionais:** função que utiliza a biblioteca de estados conformacionais para classificar a partir dos ângulos *phi* e *psi* a região ocupada no mapa de Ramachandran.

12. **Módulo de rotulamento:** utiliza a biblioteca de regiões conformacionais e a função de mapeamento para atribuir rótulos à todos os grupos k_i de todos os s_i fragmentos.

13. **Módulo para predição da estrutura secundária:** realiza a predição da estrutura secundária para seqüência alvo através de um servidor de predição.

14. **Módulo para construção da conformação inicial:** utilizando os grupos rotulados de cada s_i fragmento e a informação da estrutura secundária, é construída a conformação representada na forma de intervalos de variação.

15. **Módulo para otimização das regiões de volta:** identifica as regiões de volta do polipetídeo-alvo e realiza a redução do intervalos das regiões de volta, buscando encontrar o menor intervalo fechado que represente a conformação de menor energia potencial.

16. **Módulo de análise:** ferramentas de suporte para análise dos resultados obtidos na predição da estrutura 3D de polipeptídeos-alvo.

4.3.11 Resumo do capítulo

Neste capítulo foi apresentada um novo método para a predição *in sílico* da estrutura 3D de polipeptídeos. O método proposto utiliza técnicas de agrupamento aplicadas a dados de es-

truturas determinadas experimentalmente. A partir do agrupamento, são criados intervalos de variação angular que passam a representar a conformação de um polipeptídeo-alvo. Os ângulos de torção são obtidos de fragmentos moldes de estruturas 3D experimentais armazenadas no PDB. Os ângulos de torção dos aminoácidos nas regiões de volta têm seu intervalos de variação angular reduzido buscando, desta forma, encontrar a conformação com a menor energia pontencial. A construção de uma conformação, representada na forma de intervalos de variação, através de informações obtidas de estruturas experimentais, diminui o espaço de busca conformacional.

O método desenvolvido é capaz de predizer novas formas de enovelamento. Isto, se deve à forma em que as conformações são construídas. O método desenvolvido não está limitado à informação de proteínas-molde. Nele, os ângulos diedros do resíduo de aminoácido central, obtido de proteínas-molde, fornecem apenas a informação a respeito dos possíveis valores que o correspondente resíduo de aminoácido na proteína-alvo pode adotar. O agrupamento destes ângulos, permite identificar os estados conformacionais mais prováveis que este resíduo de aminoácido possa estar assumindo.

Quando comparado à outros métodos *de novo*, como o ROSETTA, o método de predição proposto se diferencia, principalmente, pela forma de obtenção e utilização das informações de proteínas-molde. No método ROSETTA é utilizada a informação de todos os resíduos de aminoácidos de um fragmento obtido de uma proteína-molde. No entanto, no método de predição desenvolvido é utilizada somente a informação do resíduo de aminoácido central de um fragmento obtido de uma proteína-molde. As técnicas de agrupamento unidas à forma de utilização da informações obtidas após a sua execução, dispensam a combinação de fragmentos. A construção de intervalos de variação angular para cada ângulo diedro da seqüência de resíduos de aminoácidos da proteína-alvo, permite que sejam realizadas alterações conformacionais que possam conduzir o método de predição a encontrar uma conformação com a menor energia potencial.

No próximo capítulo são apresentados os experimentos realizados com o método de predição desenvolvido. É realizada a predição da estrutura 3D de proteínas-alvo pertencentes a diferentes classes estruturais.

5 Experimentos

5.1 Introdução

Nesta seção, são apresentados os resultados obtidos com a utilização do método de predição desenvolvido. É predita a estrutura 3D aproximada de seis proteínas. Estas proteínas têm a sua estrutura 3D conhecida experimentalmente e armazenada no PDB [7, 8]. Foram escolhidos os polipeptídeos com os seguintes códigos PDB: 1ZDD [64, 88], 1K43 [69], 1ROP [5], 1UTG [61], 1GAB [55] e 1GB1 [32]. Esta escolha foi realizada com o objetivo de testar o método desenvolvido em diferentes classes de proteínas. A seguir são listadas as proteínas testadas e a sua classificação segundo o SCOP [64]:

- **Cadeia A da proteína A estabilizada por ponte de sulfeto:** código no PDB: 1ZDD [64,88]; classe: proteína projetada (*Designed protein*), enovelamento: grampo α - *Protein A Ig(Fc)-binding domain mimics* ;
- **Cadeia A da proteína MBH12:** código no PDB: 1K43 [69]; classe: proteína projetada (*Designed protein*), enovelamento: grampo beta projetado (*beta-hairpin design*);
- **Cadeia A da protéina ROP:** código PDB: 1ROP [5]; classe: hélice α (*All alpha protein*), enovelamento: grampo α - *ROP-like*;
- **Domínio B1 da proteína G do streptococcal:** código no PDB: 1GB1 [32]; classe: α + β (*Alpha and beta protein (a+b)*), enovelamento: mistura α e β - *beta-Grasp (ubiquitin-like)*.
- **Cadeia A da proteína PAB:** código no PDB: 1GAB [55]; classe: hélice α (*All alpha protein*), enovelamento: pacote de 3 hélices - *immunoglobulin/albumin-binding domain-like*;
- **Cadeia A da Uteroglobina:** código no PDB: 1UTG [61]; classe: hélice α (*All alpha protein*), enovelamento: multi hélices - *Uteroglobin-like* (*multihelical*);

5.2 Materiais e métodos

Utilizou-se o algoritmo desenvolvido, descrito no Capítulo 6, para a predição da estrutura 3D aproximada das proteínas-alvo. Para a fase de fragmentação utilizou-se um tamanho padrão para o número de resíduos de aminoácidos em cada fragmentos (l=5 resíduos de aminoácidos, pentapeptídeo). Em todos os estudos de caso foram eliminadas as proteínas-molde que possuem alguma relação evolucionária com a seqüência da proteína-alvo. Para esta seleção, eliminou-se todos as proteínas-moldes que apresentam a sua seqüência de aminoácidos com $\geq$ 50% de identidade em relação à seqüência da proteína-alvo.

A qualidade estéreoquímica das estruturas 3D preditas é analisada através do programa PROCHECK[1] [48]. As estruturas secundárias são analisadas utilizando os programas DSSP[2] [43] e PROMOTIF[3] [38]. Todas as representações gráficas das estruturas 3D são preparadas com o software PYMOL [20]. Todos os cálculos de RMSD foram realizados com o programa PROFIT (Grupo Dr. Andrew C. R. Martim) e são obtidos a partir da sobreposição do C_α da estrutura predita e do C_α da estrutura nativa da proteína-alvo. Em todos os cálculos de RMSD são desconsiderados os 2 resíduos de aminoácidos iniciais (região N-terminal) e os 2 resíduos de aminoácidos finais (região C-terminal) das estruturas 3D sobrepostas. Esta decisão foi tomada porque, conforme descrito no Capítulo 6, os ângulos de torção destes aminoácidos são fixados em 180°[4].

Adotou-se uma configuração única na parametrização do algoritmo de otimização das regiões de volta. Os parâmetros do algoritmo se referem ao número de conformações geradas, ao número de conformações utilizadas para realizar a redução do intervalo, ao tamanho mínimo do intervalo e ao tamanho do limiar próximo à região central do intervalo. Estes parâmetros são detalhados a seguir:

- Número de conformações iniciais (ι): são geradas inicialmente ι = 1.000 conformações para cada região de volta (primeiro passo de execução do algoritmo[5]);
- O intervalo é reduzido em cada região;
- Número de conformações após o primeiro passo de execução (ι): são geradas ι = 100 conformações para cada região de volta em cada passo de execução do algoritmo;

[1]PROCHECK: programa que checa a qualidade estereoquímica de uma estrutura de proteína gerando análises gráficas sobre a geometria espacial da proteína, resíduo por resíduo. Através de mapas de Ramachandran, os aminoácidos da conformação são analisados em relação às regiões energeticamente favoráveis.

[2]DSSP: o programa calcula a estrutura secundária de uma proteína, as suas coordenadas x, y e z e a acessibilidade do solvente.

[3]PROMOTIF: programa que provê detalhes da localização e dos tipos de motivos estruturais em estruturas 3D.

[4]Devido à forma de fragmentação, não é possível obter informações de proteínas molde para modelar os 2 resíduos de aminoácidos iniciais e os dois resíduos de aminoácidos finais.

[5]Passo de execução do algoritmo: um passo de execução do algoritmo compreende gerar todas as conformações necessárias de todas as regiões de volta.

- Porcentagem de conformações analisadas para determinar a forma de redução do intervalo (δ): adotou-se um valor de δ = 10 % para determinar o número das ι conformações que serão utilizadas para determinar a forma de redução do intervalo de uma determinada região de volta.
- Limiar do intervalo: adotou-se um limiar de 10% para mais e para menos a partir do centro do intervalo para a escolha da forma de redução do intervalo;
- Tamanho mínimo de um intervalo: o tamanho limite para redução do intervalo é de $w(\theta) = 10$.

Os gráficos de análise dos resultados são construídos utilizando o programa OriginLab (*Scientific Graphing an Analysis Software*). Todos os softwares de análise foram implementados usando a linguagem de programação Python e C++. A predição da estrutura secundária é realizada pelo servidor NPS@ [16] e SCRATCH [15]. Os testes foram executados numa máquina PC Intel Core 2 Duo E6400 2.4GHZ 2MB Cache e 2GB de RAM HD, 250MB com sistema operacional Linux.

5.3 Estudo de caso 1: 1ZDD

No estudo de caso 1, realizou-se a predição da estrutura 3D aproximada da mini proteína cujo código PDB é 1ZDD [88], composta por 34 resíduos de aminoácidos e conhecida pelo arranjo de duas estruturas secundárias em forma de hélices α conectadas por uma volta (código PDB: 1ZDD - Figura 31A) [64].

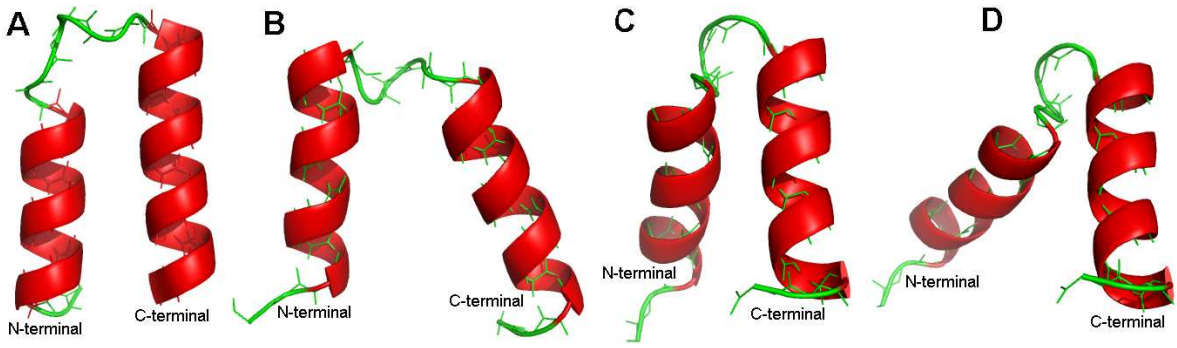

Figura 31 – Representação do tipo *Ribbon* da estrutura 3D experimental e das estruturas 3D preditas da proteína cujo código PDB é 1ZDD. (A) estrutura 3D experimental da proteína cujo código PDB é 1ZDD; (B) estrutura 3D predita, obtida a partir do centro do intervalo da conformação inicial; (C) estrutura 3D predita com menor RMSD em relação a estrutura experimental, encontrada ao longo da execução do método de predição; (D) estrutura 3D predita obtida após a otimização da região de volta. As pontes de sulfeto e as cadeias laterais foram removidas para facilitar a visualização.

A seqüência alvo K=FNMQCQRRFYEALHDPNLNEEQRNAKIKSIRDDC da proteína foi fragmentada em 30 fragmentos-alvo s_i com tamanho l=5 resíduos de aminoácidos (Tabela 6, coluna 1). Para cada fragmento-alvo s_i é realizada a busca por fragmentos-molde no PDB. Eliminou-se as proteínas cujas seqüências são idênticas ou muito similares à seqüência alvo K da proteína com código PDB igual a 1ZDD, identificadas por: 1ZDC, 1ZDD, 1L6X, 1OQO, 1OQX, 1ZDA, 1ZDB, 2SPZ, 1LP1, 1Q2N, 1FC2, 1BDC, 1BDD, 1SS1, 1DEE, 1EDK, 1EDJ, 1EDI, 1EDL. Após obtidos os arquivos *pdbs* do PDB, são calculados os ângulos de torção do aminoácido central de cada fragmento-molde (Tabela 6, coluna 2). A Tabela 6, coluna 3, apresenta o número de tuplas-molde retornadas para cada fragmento s_i.

Tabela 6: Classificação dos dupletos-molde de cada fragmento alvo da proteína de código PDB igual a 1ZDD nos três estados conformacionais (h, b e c).

Fragmento	Resíduo central	N° de moldes	Hélice α(%)	Folha β(%)	Volta(%)
FNMQC	M	34	82.35	17.65	0.00
NMQCQ	Q	40	80.00	17.50	2.50
MQCQR	C	38	86.84	13.16	0.00
QCQRR	Q	28	78.57	3.57	17.86
CQRRF	R	47	80.85	17.02	2.13
QRRFY	R	30	86.67	3.33	10.00
RRFYE	F	52	96.15	3.85	0.00
RFYEA	Y	30	70.00	30.00	0.00
FYEAL	E	42	97.62	2.38	0.00
YEALH	A	37	78.38	18.92	2.70
EALHD	L	45	80.00	17.78	2.22
ALHDP	H	49	69.39	30.61	0.00
LHDPN	D	53	3.77	86.79	9.43
HDPNL	P	22	63.64	22.73	13.64
DPNLN	N	50	98.00	2.00	0.00
PNLNE	L	24	16.67	83.33	0.00
NLNEE	N	46	21.74	69.57	8.70
LNEEQ	E	34	91.18	2.94	5.88
NEEQR	E	37	97.30	0.00	2.70
EEQRN	Q	35	82.86	17.14	0.00
EQRNA	R	57	71.93	22.81	5.26
QRNAK	N	14	92.86	7.14	0.00
RNAKI	A	29	24.14	75.86	0.00
NAKIK	K	20	90.00	10.00	0.00
AKIKS	I	28	53.57	46.43	0.00

continua na próxima página

Fragmento	Resíduo central	N° de moldes	Hélice α(%)	Folha β(%)	Volta(%)
KIKSI	K	40	90.00	10.00	0.00
IKSIR	S	74	32.43	41.89	25.68
KSIRD	I	22	90.91	9.09	0.00
SIRDD	R	16	87.50	6.25	6.25
IRDDC	D	35	57.14	40.00	2.86

A Figura 32 e a Figura 33 apresentam o mapa de Ramachandran das tuplas-molde de cada fragmento-alvo s_i. A partir de sua análise é possível identificar as regiões no mapa de Ramachandran onde se encontra o maior número de tuplas-molde.

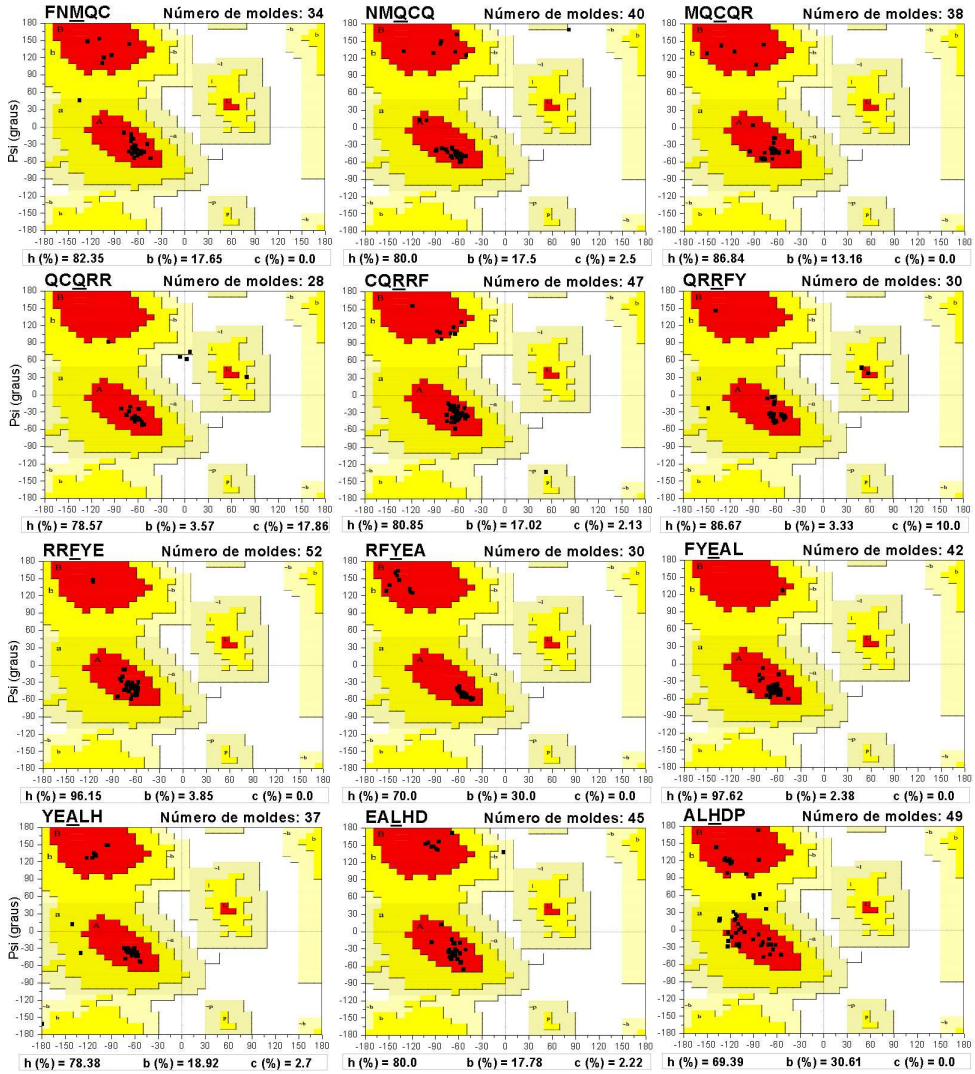

Figura 32 – Mapa de Ramachandran das t_i tuplas de cada fragmento s_i da proteína cujo código PDB é 1ZDD (parte 1).

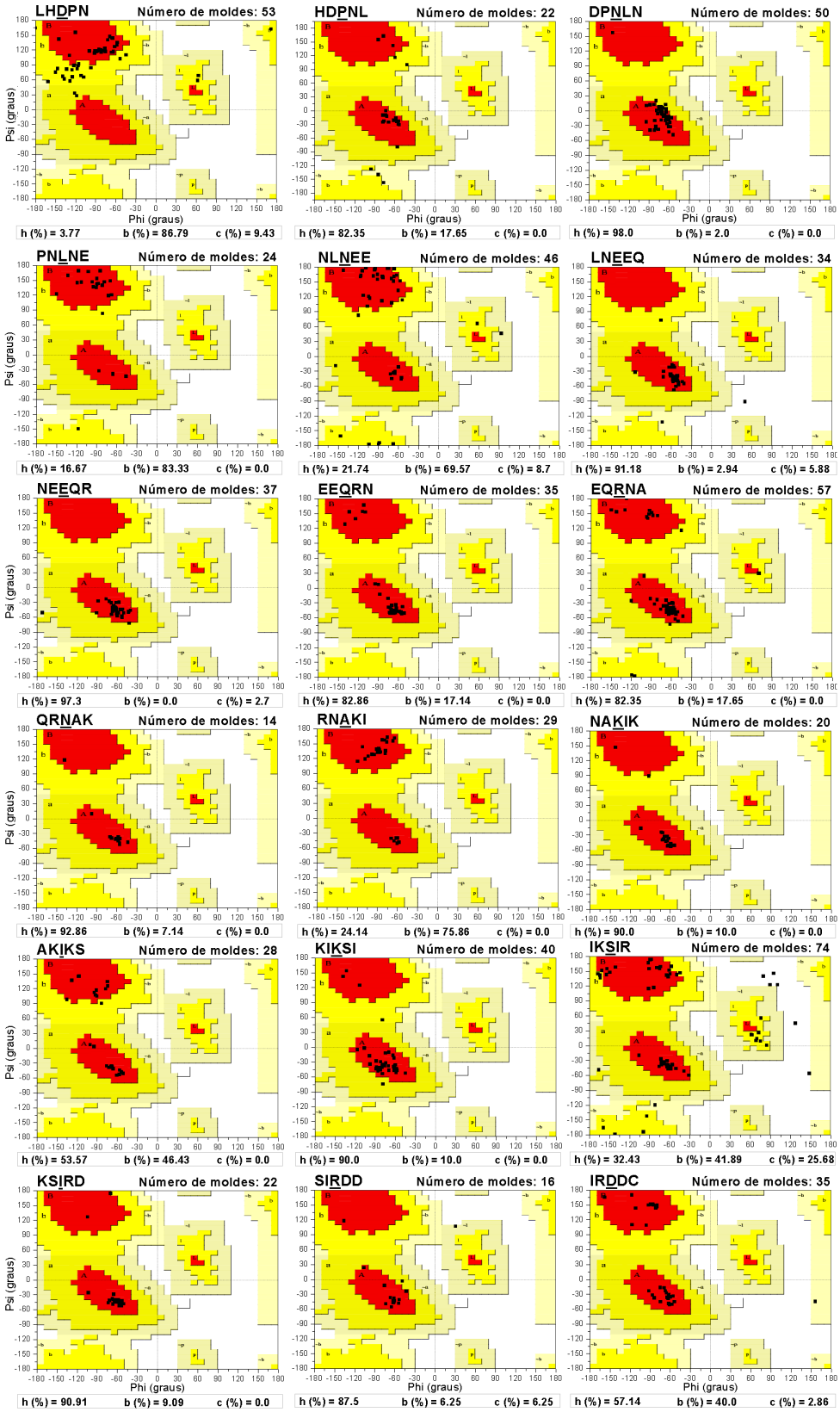

Figura 33 – Mapa de Ramachandran das t_i tuplas de cada fragmento s_i da proteína cujo código PDB é 1ZDD (parte 2).

A Tabela 6, coluna 4, 5 e 6 apresenta a porcentagem das tuplas-molde associadas a cada um dos três estados conformacionais (h, b e c). Nesta classificação, o estado conformacional de hélice α (h) compreende os resíduos de aminoácidos nos estados "A", "a", "L", "l"e "p", o estado de folha β (b) compreende os resíduos de aminoácidos em estados "B"e "b"e as regiões de volta (c) compreendem os resíduos de aminoácidos em estado "c"(seguindo o modelo de 8 estados descrito na seção 4.3.6 e baseando-se no modelo para escolha dos grupos apresentados na seção 4.3.8).

Cada fragmento s_i têm as suas tuplas-molde agrupadas em 4 grupos. Para cada grupo de s_i é calculada a média e o desvio padrão estimado. A Tabela 7 apresenta o resultado obtido com o agrupamento das tuplas de cada fragmento s_i.

Tabela 7: Agrupamento das tuplas-molde associadas a um fragmento alvo s_i da proteína cujo código PDB é 1ZDD: (*m*) é o valor médio e (σ) é o desvio padrão estimado de cada grupo k_i.

		Grupo 01		Grupo 02		Grupo 03		Grupo 04	
Frag		phi	psi	phi	psi	phi	psi	phi	psi
FNMQC	*m*	-71.19	143.62	-112.19	116.92	-58.29	-43.59	-67.29	-31.27
FNMQC	σ	21.07	67.65	13.76	35.08	6.97	6.21	4.01	12.13
NMQCQ	*m*	-63.36	-46.89	-92.62	-22.07	-80.8	138.91	81.31	169.34
NMQCQ	σ	6.38	6.46	10.08	24.14	24.35	12.14	29.3	76.51
MQCQR	*m*	-132.39	133.88	-66.66	-37.61	-82.38	126.06	-69.57	-40.95
MQCQR	σ	14.43	5.85	1.62	1.43	4.73	17.36	11.48	14.63
QCQRR	*m*	-71.17	-27.31	79.54	30.64	-23.71	73.14	-61.06	-43.38
QCQRR	σ	7.24	5.02	0.36	0.27	43.22	11.42	3.99	4.47
CQRRF	*m*	52.91	-133.26	-90.38	118.06	-62.76	-32.01	-63.96	114.25
CQRRF	σ	20.68	60.25	16.65	21.31	6.76	9.74	5.33	8.51
QRRFY	*m*	-147.42	-24.05	-60.61	-33.16	52.31	43.63	-137.61	145.55
QRRFY	σ	42.27	40.81	6.57	13.20	3.98	4.49	0.41	0.89
RRFYE	*m*	-115.34	145.51	-60.6	-40.44	-77.39	-27.44	-67.04	-40.95
RRFYE	σ	0.17	1.51	1.65	4.86	3.65	15.44	4.44	8.88
RFYEA	*m*	-121.32	127.63	-142.52	148.54	-46.19	-57.19	-57.16	-46.99
RFYEA	σ	1.55	2.97	5.99	12.21	3.55	2.85	2.95	4.24
FYEAL	*m*	-59.85	-44.42	-53.22	127.40	-67.18	-45.65	-83.41	-26.78
FYEAL	σ	4.59	9.22	8.88	28.33	3.25	5.72	6.22	13.88
YEALH	*m*	-70.86	-32.23	-108.58	135.34	-61.45	-40.5	-149.87	-62.71
YEALH	σ	2.16	2.76	9.12	8.70	5.09	6.36	21.37	72.98
EALHD	*m*	-88.38	-3.28	-2.55	137.63	-64.52	-39.08	-89.65	151.56
EALHD	σ	6.25	15.37	15.88	76.78	5.41	10.41	9.78	8.75
ALHDP	*m*	-85.93	76.44	-71.79	-28.82	-124.16	119.91	-114.17	0.03

continua na próxima página

		Grupo 01		Grupo 02		Grupo 03		Grupo 04	
Frag		phi	psi	phi	psi	phi	psi	phi	psi
ALHDP	σ	7.23	51.28	8.75	10.65	5.62	10.71	9.06	18.89
LHDPN	m	-83.45	124.87	-118.23	85.35	98.73	96.45	-130.05	62.93
LHDPN	σ	27.86	14.8	9.18	3.90	51.66	46.69	15.66	17.02
HDPNL	m	-65.11	134.98	-85.59	-141.81	-56.6	-32.11	-74.2	-15.16
HDPNL	σ	14.81	23.57	7.81	12.02	2.63	19.90	2.95	8.79
DPNLN	m	-75.55	3.99	-144.42	157.21	-81.98	-32.95	-62.59	-19.22
DPNLN	σ	7.50	7.91	14.30	29.35	5.96	7.35	3.96	11.48
PNLNE	m	-84.14	138.23	-117.49	-149.25	-65.93	-38.54	-122.48	149.63
PNLNE	σ	9.79	20.68	23.81	88.36	14.21	4.00	16.08	15.97
NLNEE	m	-99.17	-173.48	75.39	55.88	-72.89	-34.3	-89.25	148.99
NLNEE	σ	26.17	6.46	17.85	9.88	27.32	8.55	22.02	26.98
LNEEQ	m	-75.18	72.6	-98.2	-63.98	49.64	-91.61	-56.01	-41.89
LNEEQ	σ	24.44	29.39	20.39	46.59	24.44	29.39	6.48	12.37
NEEQR	m	-70.86	-28.57	-171.62	-51.22	-60.12	-43.46	-56.74	-53.8
NEEQR	σ	10.17	2.61	21.22	10.30	5.77	3.76	11.85	6.55
EEQRN	m	-68.55	-44.77	-90.67	0.77	-121.81	149.16	-54.96	-48.05
EEQRN	σ	4.07	7.49	2.92	9.53	11.42	12.33	2.18	4.89
EQRNA	m	-64.51	-42.05	-102.11	149.49	70.28	29.12	-116.72	-176.78
EQRNA	σ	13.92	15.08	28.1	14.09	0.56	0.02	2.12	1.31
QRNAK	m	-137.76	117.90	-44.73	-48.13	-97.2	9.52	-60.08	-42.14
QRNAK	σ	23.6	44.34	3.83	0.75	23.6	44.34	4.59	4.66
RNAKI	m	-62.2	-44.11	-86.49	134.61	-107.07	126.13	-73.8	157.81
RNAKI	σ	3.48	4.11	4.9	3.71	8.13	9.81	6.52	3.16
NAKIK	m	-68.95	-34.84	-60.52	-48.93	-141.48	146.49	-98.06	36.09
NAKIK	σ	3.37	6.14	3.29	4.22	20.02	50.50	5.98	52.55
AKIKS	m	-84.76	113.8	-63.16	-43.46	-97.52	5.15	-122.85	133.66
AKIKS	σ	7.91	12.62	7.71	6.62	2.81	1.98	7.36	18.04
KIKSI	m	-60.22	-42.96	-128.02	140.33	-80.09	-21.43	-81.62	-40.62
KIKSI	σ	1.80	4.80	10.62	11.98	18.85	29.46	5.73	6.00
IKSIR	m	-67.55	-40.23	-97.00	148.73	86.51	48.39	-107.37	-150.98
IKSIR	σ	23.93	8.09	40.83	13.40	23.93	58.70	25.60	18.73
KSIRD	m	-103.63	50.18	-69.82	174.34	-66.59	-41.57	-55.37	-47.16
KSIRD	σ	0.65	76.34	13.98	57.9	3.53	5.36	2.5	3.85
SIRDD	m	30.54	106.98	-135.71	117.80	-60.57	-37.61	-106.37	23.48
SIRDD	σ	33.65	53.85	2.97	5.38	8.67	14.68	0.00	53.85
IRDDC	m	-90.04	141.60	-67.02	-35.74	-138.19	167.76	156.10	-44.80

continua na próxima página

		Grupo 01		Grupo 02		Grupo 03		Grupo 04	
Frag		phi	psi	phi	psi	phi	psi	phi	psi
IRDDC	σ	12.17	14.59	10.11	10.32	20.29	2.41	45.24	91.13

A partir do valor médio e do desvio padrão estimado de cada grupo k_i de s_i (Tabela 7) é criado o intervalo de variação de cada grupo. Em seguida, a partir do ângulo central do intervalo, cada grupo é rotulado em uma das 8 regiões conformacionais empregadas no método. Dando seqüência, é realizada a predição da estrutura secundária para a seqüência-alvo K (Figura 34). Com base no consenso entre os resultados obtidos pelos métodos de predição da estrutura secundária da seqüência K é construída a conformação inicial representada na forma de intervalos.

```
                    10        20        30
                    |         |         |
          FNMQCQRRFYEALHDPNLNEEQRNAKIKSIRDDC
DSC       ccchhhhhhhhhhccccchhhhhhhhhhhccccc
PHD       ccchhhhhhhhhhccccchhhhchhhhhhccccc
Predator  ccchhhhhhhhhhcccccchhhhhhhhhhhccccc
Consenso  ccchhhhhhhhhhcccccchhhhhhhhhhhccccc
```

Figura 34 – Predição da estrutura secundária da seqüência-alvo K da proteína cujo código PDB é 1ZDD. O consenso representa a estrutura secundária obtida pela análise simultânea da predição realizada pelo método DSC, PHD e Predator.

A Figura 31B apresenta a estrutura 3D predita da proteína cujo código PDB é 1ZDD. Esta estrutura é obtida a partir do centro do intervalo da conformação inicial representada por intervalos de variação angular. Após construída a conformação inicial, representada por intervalos, é realizada a otimização das regiões de volta. Os segmentos de resíduos de aminoácidos, identificados na predição da estrutura secundária como aminoácidos de regiões conformacionais de volta, têm o intervalo de seus ângulos diedros reduzido objetivando encontrar a estrutura com menor energia potencial. A Figura 31C e 31D ilustra, respectivamente, a estrutura 3D com menor RMSD encontrada durante a otimização da região de volta e a estrutura 3D predita obtida como o resultado final do método de predição. A conformação final (31D) obtida pelo método de predição é a conformação de menor EP encontrada no último passo de execução do algoritmo.

A Tabela 8 mostra o valor de RMSD e de energia potencial das estruturas 3D preditas. A estrutura 3D predita com menor RMSD em relação a estrutura experimental, encontrada ao longo de todo processo de otimização da região de volta, apresenta RMSD = 4.42Å(Figura 31C). A estrutura 3D de menor energia potencial encontrada no último passo de execução do algoritmo

apresenta um valor de RMSD = 5.00Å(Figura 31D). Esta estrutura representa a estrutura 3D final predita pelo algoritmo.

Através da análise das conformações preditas é possível verificar a discordância existente entre os valores de energia e RMSD (Tabela 8). Estruturas com alta energia (C) possuem um RMSD menor do que as estruturas com menor energia (D). O valor elevado da energia é ocasionado por choques estereoquímicos entre átomos da cadeia principal e da cadeia lateral do polipeptídeo. A colocação incorreta das cadeias laterais provoca esse aumento na energia. Isto, afeta a forma que o algoritmo procede a redução do intervalo das regiões de volta. Sendo EP o critério para a escolha das conformações que são utilizadas para decidir a forma de redução do intervalo, então as conformações com baixa energia potencial e com alto RMSD escolhidas, podem provocar uma redução incorreta do intervalo de variação angular.

Tabela 8: Valor de energia potencial ($Kcal.mol^{-1}$) estruturas 3D preditas e o valor de RMSD (Å) do C_α em relação à estrutura 3D experimental da proteína cujo código PDB é 1ZDD.

Estrutura predita	RMSD(Å)	EP($Kcal.mol^{-1}$)
B	5.51	815113064.11
C	4.42	107049.31
D	5.00	4652.85

No entanto, ao analisar a qualidade da estrutura secundária, é possível verificar que a formação da estrutura secundária das estruturas preditas é semelhante à da estrutura 3D experimental da 1ZDD (Tabela 9).

Tabela 9: Estrutura secundária das estruturas 3D preditas da proteína cujo código PDB é 1ZDD. (A) estrutura 3D experimental da proteína cujo código PDB é 1ZDD; (B) estrutura 3D predita, obtida a partir do centro do intervalo da conformação inicial representada por intervalos de variação angular; (C) estrutura 3D predita de menor RMSD encontrada ao longo da execução do método de otimização da região de volta; (D) estrutura 3D predita como resultado final do método de predição.

Estrutura	Folha β	Hélice α	Hélice α 3^{10}	Outras	Total resíduos
A	0 (0.0%)	25 (73.5%)	0 (0.0%)	9 (26.5%)	34
B	0 (0.0%)	26 (76.5%)	0 (0.0%)	8 (23.5%)	34
C	0 (0.0%)	26 (76.5%)	0 (0.0%)	8 (23.5%)	34
D	0 (0.0%)	26 (76.5%)	0 (0.0%)	8 (23.5%)	34

Os mapas de Ramachandran da Figura 35 demonstram que os resíduos de aminoácidos das estruturas 3D (B, C e D) preditas se encontram em regiões similares às ocupadas na estrutura 3D experimental (A). A porcentagem média de resíduos de aminoácidos das estruturas 3D preditas, que ocupam as regiões mais favoráveis no mapa de Ramachandran, é de aproximadamente 86%. Claramente, este valor, demonstra que a estrutura secundária das estruturas 3D preditas estão bem formadas (Tabela 10).

A Tabela 11, apresenta os valores de RMSD obtidos a partir da sobreposição das regiões de estruturas secundárias regulares da estrutura 3D experimental da proteína cujo código PDB é 1ZDD e da estrutura 3D final predita (Figura 31D) pelo método desenvolvido. Os valores obtidos mostram que as estruturas secundárias regulares estão bem formadas.

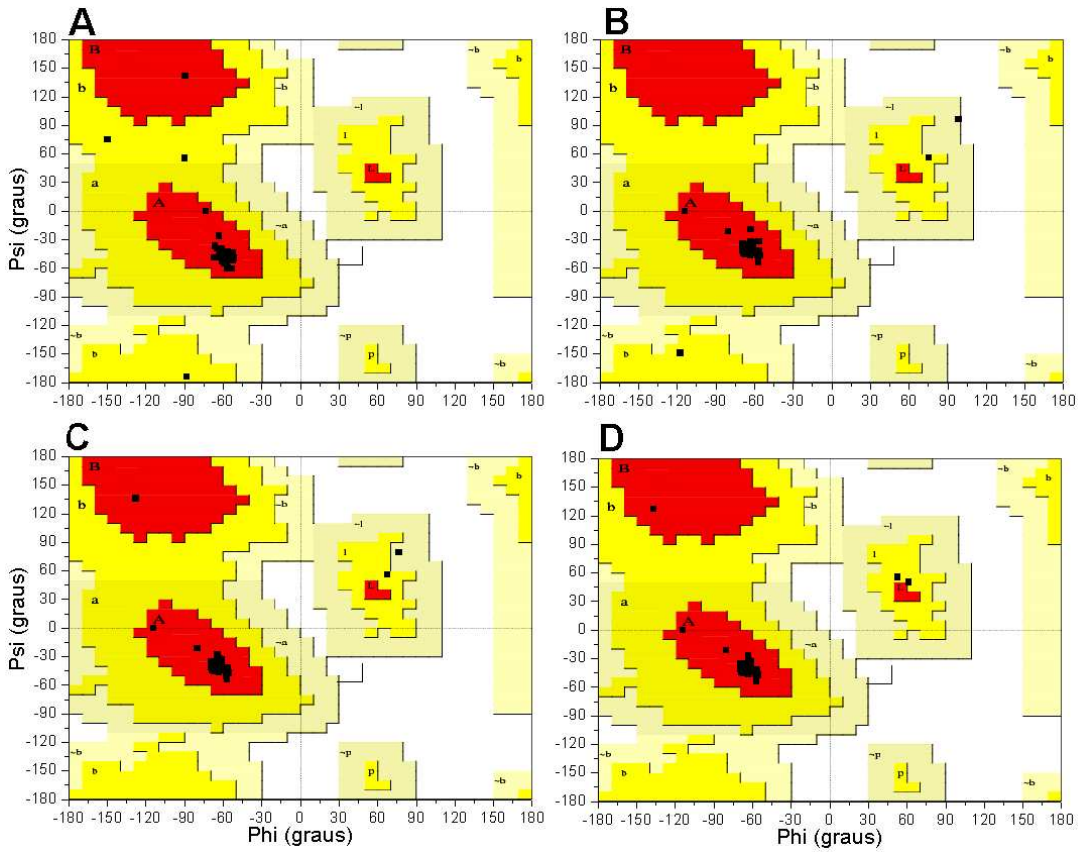

Figura 35 – Mapa de Ramachandran da estrutura 3D experimental e das estruturas 3D preditas da proteína cujo código PDB é 1ZDD. (A) estrutura 3D experimental da proteína cujo código PDB é 1ZDD; (B) estrutura 3D predita, obtida a partir do meio do intervalo; (C) estrutura 3D predita de menor RMSD encontrada durante a otimização da região de volta; (D) estrutura 3D predita representando o a estrutura final obtida pelo método de predição.

Tabela 10: Análise da localização dos resíduos de aminoácidos das estruturas 3D preditas para a proteína cujo código PDB é 1ZDD no mapa de Ramachandran.

Estrutura	Mais favorável (%)	Favorável (%)	Aceitável (%)	Não aceitável (%)
A	87.10	12.90	0.00	0.00
B	83.90	12.90	3.20	0.00
C	87.10	9.70	3.20	0.00
D	87.10	12.90	0.00	0.00

Tabela 11: Valor de RMSD do Cα da estrutura 3D final predita em relação à estrutura 3D experimental da proteína cujo código PDB é 1ZDD nas regiões de estruturas secundárias regulares.

Intervalo de aminoácidos (i-j)	RMSD Cα(Å)
4 - 14	0.60
20 - 30	0.40

A Figura 36 apresenta a relação RMSD *versus* EP das 1.000 conformações geradas durante a execução do algoritmo de predição e que apresentam a menor EP. Através de sua análise, é possível verificar que, embora determinada conformação possua um valor de RMSD baixo, o valor de sua energia potencial pode ser alto. Isto, pode ocasionar uma decisão incorreta durante a redução do intervalo.

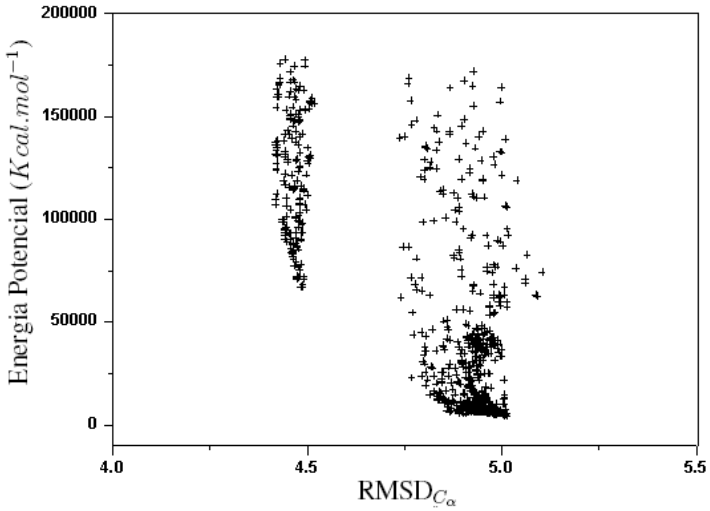

Figura 36 – Gráfico de energia *versus* RMSD. Relação entre a energia potencial e RMSD das 1.000 conformações com menor RMSD geradas pelo algoritmo de predição para a proteína cujo código PDB é 1ZDD.

O algoritmo de otimização das regiões de volta alcançou o seu critério de parada (tamanho do intervalo) após gerar 2.900 conformações. Por meio da análise dos resultados obtidos pela execução do método proposto verifica-se que as estruturas 3D preditas apresentam resultados ótimos quando a formação da estrutura secundária. E satisfatórios no que se refere à organização das estruturas secundárias regulares no espaço 3D.

5.4 Estudo de caso 2: 1K43

No estudo de caso 2, realizou-se a predição da estrutura 3D aproximada da mini proteína cujo código PDB é 1K43 [69], composta por 14 resíduos de aminoácidos e conhecida pelo arranjo de duas estruturas secundárias em forma de folhas β conectadas por uma volta (código PDB: 1K43 - Figura 37A) [64].

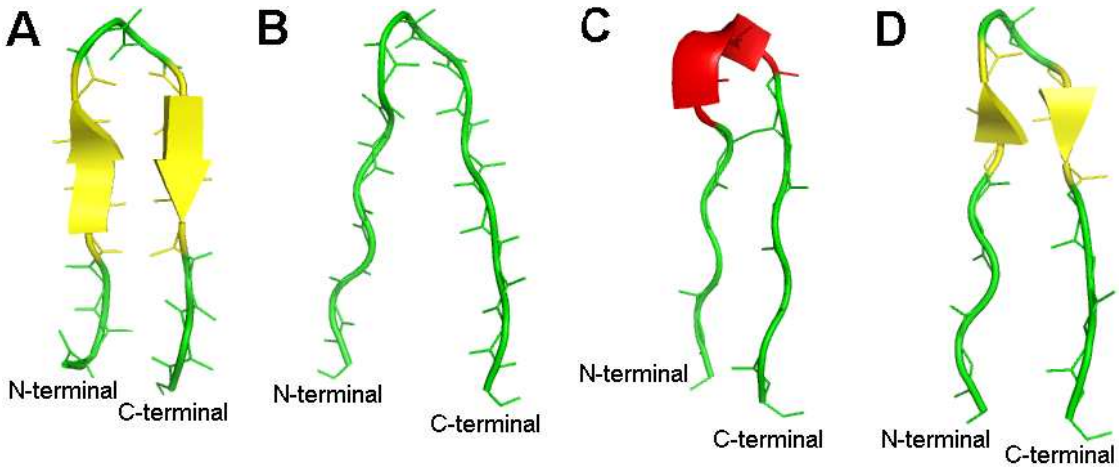

Figura 37 – Representação do tipo *Ribbon* da estrutura 3D experimental e das estruturas 3D preditas da proteína cujo código PDB é 1K43. (A) estrutura 3D experimental da proteína cujo código PDB é 1K43; (B) estrutura 3D predita, obtida a partir do centro do intervalo da conformação inicial; (C) estrutura 3D predita com menor RMSD em relação a estrutura experimental, encontrada ao longo da execução do método de predição; (D) estrutura 3D predita obtida após a otimização da região de volta. As cadeias laterais foram removidas para facilitar a visualização.

A seqüência alvo K=RGKWTYNGITYEGR da proteína foi fragmentada em 10 fragmentos-alvo s_i com tamanho l=5 resíduos de aminoácidos (Apêndice A, Tabela 32, coluna 1). Para cada fragmento alvo s_i, foi realizada a busca por proteínas-molde no PDB (Apêndice A, Tabela 32, coluna 3, apresenta o número de moldes identificados para cada fragmento alvo s_i). Eliminou-se as proteínas cujas seqüências são idênticas ou muito similares à seqüência alvo K da proteína de código PDB igual a 1K43, identificadas por: 1K43, 1J4M. Após obtidos os arquivos pdb das proteínas-molde do PDB, são calculados os ângulos do aminoácido central de cada fragmento-molde (Apêndice A, Tabela 32, coluna 2).

A partir dos estados conformacionais adotados pelos fragmentos moldes (Apêndice A, Tabela 32, coluna 4, 5 e 6), é possível estimar a *priori* a formação de duas folhas β ligadas por uma volta (Apêndice A, dos resíduos de aminoácidos RGKWT até WTYNG e de NGITY até TYEGR). Nesta classificação, o estado conformacional de hélice α (h) compreende os resíduos de aminoácidos nos estados "A", "a", "L", "l"e "p", o estado de folha β (b) compreende os resíduos de aminoácidos em estados "B"e "b"e as regiões de volta (c) compreendem os resíduos de aminoácidos em estado "c"(seguindo o modelo de 8 estados descrito na seção 4.3.6 e baseando-se no modelo para escolha dos grupos apresentados na seção 4.3.8).

Os ângulos de torção de cada fragmento s_i são agrupados (Apêndice F, Tabela 37). Para cada grupo de s_i é calculada a média e o desvio padrão estimado.

A partir do valor médio e do valor de desvio padrão estimado é criado o intervalo de variação de cada grupo. Em seguida, a partir do centro do intervalo, cada grupo é rotulado nos 8 estados conformacionais. Posteriormente, é predita a estrutura secundária (ES) da seqüência K (Figura 38). Com base na ES é construída a conformação inicial, representada na forma de intervalos. A Figura 37B apresenta a estrutura 3D obtida a partir do centro do intervalo da conformação inicial representada por intervalos de variação angular.

```
                    10
                    |
           RGKWTYNGITYEGR
GOR3       cceeeeeceeeeee
PHD        cceeeeccccecec
SCRATCH    ccccecccceeeccc
Sec.Cons.  cceeeecceeeeecc
```

Figura 38 – Predição da estrutura secundária da seqüência alvo K da proteína cujo código PDB é 1K43. O consenso representa a estrutura secundária obtida pela análise simultânea da predição realizada pelo método GOR3, PHD e SCRATCH.

As regiões de volta são otimizadas. A Figura 37C apresenta a estrutura 3D com menor RMSD encontrada durante a otimização das regiões de volta. A Figura 37D apresenta a estrutura 3D final predita obtida como resultado final do método de predição. A conformação final obtida (Figura 37D) pelo método de predição é a conformação de menor EP encontrada no último passo de execução do algoritmo.

A Tabela 12 mostra o valor de RMSD e de EP das conformações preditas. A estrutura 3D predita de menor RMSD em relação à estrutura 3D experimental, encontrada ao longo de todo processo de otimização da região de volta, apresenta um valor de RMSD = 0.95Å(Figura 37C). A estrutura de menor energia potencial, encontrada no último passo de execução do algoritmo, apresenta um valor de RMSD = 1.28Å(Figura 37D). As estruturas 3D preditas da proteína de código PDB igual a 1K43 apresentam valores de RMSD baixos em relação à estrutura 3D

experimental. A organização das estruturas secundárias no espaço 3D e são bastante similares à adotada pela estrutura nativa da proteína.

Tabela 12: Valor de energia potencial ($Kcal.mol^{-1}$) estruturas 3D preditas e o valor de RMSD (Å) do C_α em relação à estrutura 3D experimental da proteína cujo código PDB é 1K43.

Estrutura predita	RMSD(Å)	EP($Kcal.mol^{-1}$)
B	1.37	141.91
C	0.95	1047.67
D	1.28	-40.21

No entanto, conforme pode ser observado na Figura 37 (A, B ,C e D), as regiões de estruturas regulares em forma de folhas β, não estão suficientemente próximas para a formação de ligações de hidrogênio. São estas ligações as responsáveis por formar as folhas β antiparelas presentes na estrutura 3D da proteína.

É possível comprovar a não formação das ligações de hidrogênio a partir da análise das conformações preditas. Apesar, dos resíduos de aminoácidos das estruturas 3D preditas ocuparem as regiões permitidas do mapa de Ramachandran (Figura 39, Tabela 13), as estruturas em folha β não estão totalmente formadas (Tabela 15).

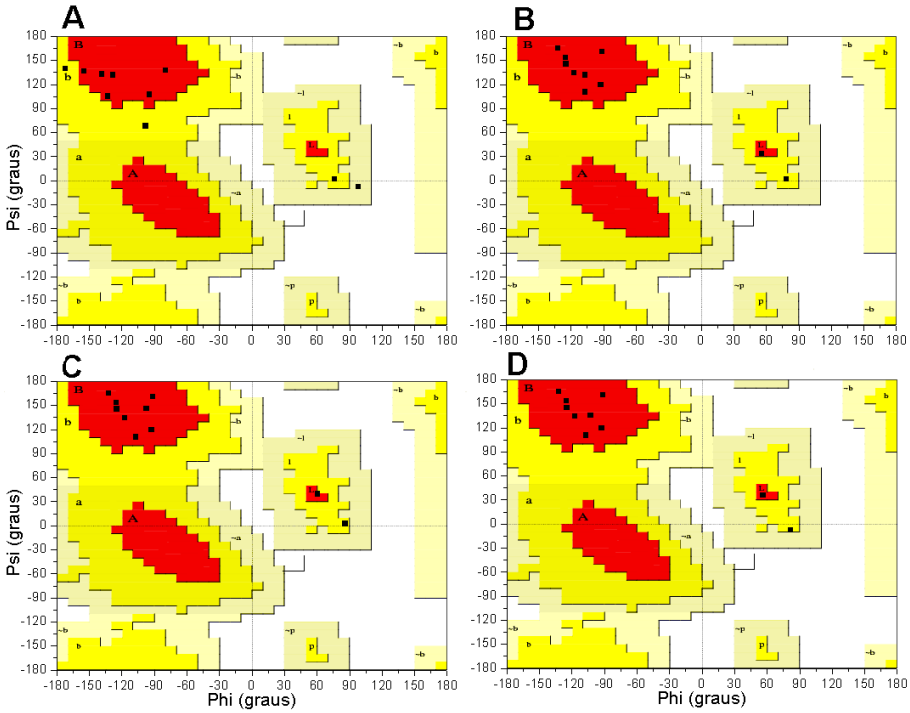

Figura 39 – Mapa de Ramachandran da estrutura 3D experimental e das estruturas 3D preditas da proteína cujo código PDB é 1K43. (A) estrutura 3D experimental da proteína cujo código PDB é 1K43; (B) estrutura 3D predita, obtida a partir do meio do intervalo; (C) estrutura 3D predita de menor RMSD encontrada durante a otimização da região de volta; (D) estrutura 3D predita representando o a estrutura final obtida pelo método de predição.

Tabela 13: Análise da localização dos resíduos de aminoácidos das estruturas 3D das conformações preditas para a proteína cujo código PDB é 1K43 no mapa de Ramachandran.

Estrutura	Mais favorável (%)	Favorável (%)	Aceitável (%)	Não aceitável (%)
A	66.70	33.30	0.00	0.00
B	100.00	0.00	0.00	0.00
C	100.00	0.00	0.00	0.00
D	100.00	0.00	0.00	0.00

A Tabela 14, apresenta os valores de RMSD obtidos a partir da sobreposição das regiões de estruturas secundárias regulares da estrutura 3D experimental e da estrutura 3D final predita (Figura 37D).

Tabela 14: Valor de RMSD do Cα da estrutura 3D final predita em relação à estrutura 3D experimental da proteína cujo código PDB é 1K43 nas regiões de estruturas secundárias regulares.

Intervalo de aminoácidos (i-j)	RMSD Cα(Å)
3 - 6	0.90
9 - 12	1.64

De forma similar ao que foi observado no estudo de caso 1, choques estereoquímicos entre átomos da cadeia lateral e de átomos da cadeia principal causam distorções no valor de EP de cada conformação, isto, vem a ocasionar situações em que conformações preditas apresentam valores de RMSD baixos, porém energia potencial elevada (Figura 40). Desta forma, durante a execução do algoritmo de otimização das regiões de volta, podem estar sendo tomadas decisões incorretas para a redução do intervalo.

Tabela 15: Estrutura secundária das estruturas 3D preditas da proteína cujo código PDB é 1K43. (A) estrutura 3D experimental da proteína cujo código PDB é 1K43; (B) estrutura 3D predita, obtida a partir do centro do intervalo da conformação inicial representada por intervalos de variação angular; (C) estrutura 3D predita de menor RMSD encontrada ao longo da execução do método de otimização da região de volta; (D) estrutura 3D predita como resultado final do método de predição.

Estrutura	Folha β	Hélice α	Hélice α 3^{10}	Outras	Total resíduos
A	6 (42.9%)	0 (0.0%)	0 (0.0%)	8 (57.1%)	14
B	0 (0.0%)	0 (0.0%)	0 (0.0%)	14 (100.0%)	14

continua na próxima página

Estrutura	Folha β	Hélice α	Hélice α 3^{10}	Outras	Total resíduos
C	0 (0.0%)	0 (0.0%)	3 (21.4%)	11 (78.6%)	14
D	4 (28.6%)	0 (0.0%)	0 (0.0%)	10 (71.4%)	14

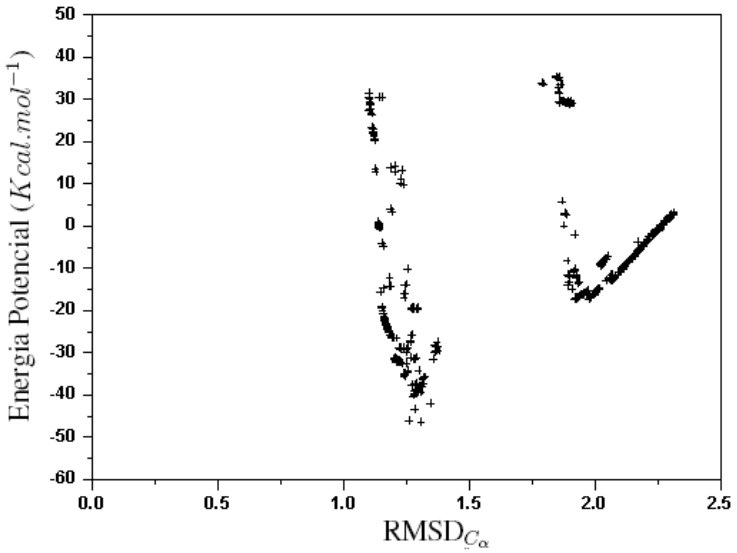

Figura 40 – Gráfico de energia *versus* RMSD. Relação entre a energia potencial e RMSD das 1.000 conformações com menor RMSD geradas pelo algoritmo de predição para a proteína cujo código PDB é 1K43.

O algoritmo de otimização das regiões de volta alcançou o seu critério de parada (tamanho do intervalo) após gerar 2.900 conformações. Por meio da análise dos resultados obtidos pela execução do método proposto verifica-se que as estruturas 3D preditas apresentam resultados satisfatórios quando a formação da estrutura secundária. E resultados ótimos no que se refere à organização das estruturas secundárias regulares no espaço 3D.

5.5 Estudo de caso 3: 1ROP

No estudo de caso 3, testou-se o método desenvolvido na predição da estrutura 3D aproximada da proteína cujo código PDB é 1ROP [5], composta por 54 resíduos de aminoácidos e conhecida pelo arranjo de duas estruturas secundárias em forma de hélice α conectadas por uma volta (código PDB: 1ROP - Figura 41A).

A seqüência alvo K=FMTKQEKTALNMARFIRSQTLTLLEKLNELDADEQADICESLH DHADELYRSCLA é fragmentada em 52 fragmentos-alvo s_i de tamanho l=5 resíduos de aminoácidos (Apêndice B, Tabela 33, coluna 1). Para cada fragmento alvo s_i, foi realizada a busca por proteínas-molde no PDB (Apêndice B, Tabela 33, coluna 3, apresenta o número de moldes identificados para cada fragmento alvo s_i). Eliminou-se as proteínas cujas seqüências são

idênticas ou muito similares à seqüência alvo K da proteína de código PDB igual a 1ROP, identificados por: 1ROP, 1B6Q, 1GMG, 1RPR, 2GHY, 1GTO, 1RPO, 1NKD, 1YO7, 1QX8, 1F4M, 1F4N. Após obtidos os arquivos *pdbs* do PDB, são calculados os ângulos de torção do aminoácido central de cada fragmento-molde (Apêndice B, Tabela 33, coluna 2).

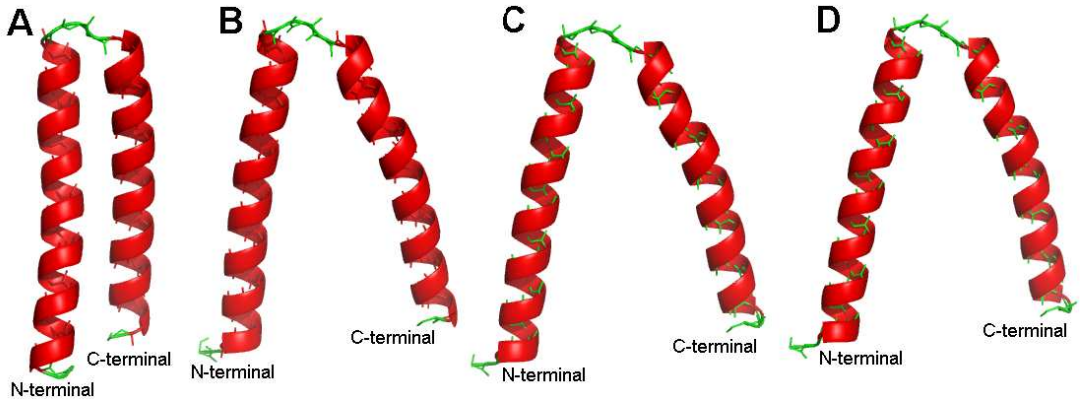

Figura 41 – Representação do tipo *Ribbon* da estrutura 3D experimental e das estruturas 3D preditas da proteína cujo código PDB é 1ROP. (A) estrutura 3D experimental da proteína cujo código PDB é 1ROP; (B) estrutura 3D predita, obtida a partir do centro do intervalo da conformação inicial; (C) estrutura 3D predita com menor RMSD em relação a estrutura experimental, encontrada ao longo da execução do método de predição; (D) estrutura 3D predita obtida após a otimização da região de volta. As cadeias laterais foram removidas para facilitar a visualização.

O Apêndice B, Tabela 33, coluna 4, 5 e 6 apresenta a porcentagem das tuplas-molde associadas a cada um dos três estados conformacionais (h, b ou c). Nesta classificação, o estado conformacional de hélice α (h) compreende os resíduos de aminoácidos nos estados "A", "a", "L", "l"e "p", o estado de folha β (b) compreende os resíduos de aminoácidos em estados "B"e "b"e as regiões de volta (c) compreendem os resíduos de aminoácidos em estado "c"(seguindo o modelo de 8 estados descrito na seção 4.3.6 e baseando-se no modelo para escolha dos grupos apresentados na seção 4.3.8).

Os ângulos de torção de cada fragmento s_i são agrupados em 4 grupos. Para cada grupo de s_i é calculada a média e o desvio padrão estimado. O Apêndice G, Tabela 38 apresenta o resultado obtido com o agrupamento das tuplas-molde de cada fragmento s_i.

A partir do valor médio e do valor de desvio padrão estimado de cada grupo k_i de s_i é criado o intervalo de variação de cada grupo. Em seguida, a partir do centro do intervalo, cada grupo é rotulado nos 8 estados conformacionais. Posteriormente, é predita a estrutura secundária (ES) da seqüência-alvo K (Figura 42). Com base na ES é construída a conformação inicial, representada na forma de intervalos. A Figura 41B apresenta a conformação obtida a partir do centro do intervalo da conformação inicial representada por intervalos de variação angular.

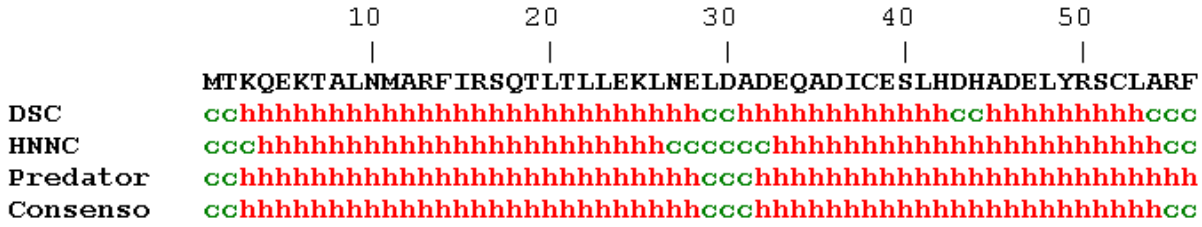

Figura 42 – Predição da estrutura secundária da seqüência-alvo K da proteína cujo código PDB é 1ROP. O consenso representa a estrutura secundária obtida pela análise simultânea da predição realizada pelo método DSC, HNNC e Predator.

As regiões de volta são otimizadas. A Figura 41C apresenta a estrutura 3D com menor RMSD encontrada durante a execução do algoritmo de otmização das regiões de volta. A Figura 41D apresenta a estrutura 3D final predita obtida como resultado final do método de predição.

A Tabela 16 apresenta o valor EP e de RMSD das conformações preditas. A estrutura predita com menor RMSD em relação a estrutura experimental, encontrada ao longo de todo processo de otimização da região de volta, apresenta um valor de RMSD = 6.03Å(Figura 41C). A estrutura de menor energia, encontrada no último passo de execução do algoritmo, apresenta um valor de RMSD = 6.69Å(Figura 41D).

Tabela 16: Valor de energia potencial ($Kcal.mol^{-1}$) estruturas 3D preditas e o valor de RMSD (Å) do C_α em relação à estrutura 3D experimental da proteína cujo código PDB é 1ROP.

Estrutura predita	RMSD(Å)	EP($Kcal.mol^{-1}$)
B	5.15	32983.35
C	6.03	1276.34
D	6.692	1267.78

Ao analisar a qualidade da estrutura secundária, é possível verificar que a formação da estrutura secundária das estruturas 3D preditas é idêntica à da estrutura 3D experimental da proteína-alvo (Tabela 17).

Tabela 17: Estrutura secundária das estruturas 3D preditas da proteína cujo código PDB é 1ROP. (A) estrutura 3D experimental da proteína cujo código PDB é 1ROP; (B) estrutura 3D predita, obtida a partir do centro do intervalo da conformação inicial representada por intervalos de variação angular; (C) estrutura 3D predita de menor RMSD encontrada ao longo da execução do método de otimização da região de volta; (D) estrutura 3D predita como resultado final do método de predição.

Estrutura	Folha β	Hélice α	Hélice α 3^{10}	Outras	Total resíduos
A	0 (0.0%)	50 (89.3%)	0 (0.0%)	6 (10.7%)	56
B	0 (0.0%)	50 (89.3%)	0 (0.0%)	6 (10.7%)	56
C	0 (0.0%)	50 (89.3%)	0 (0.0%)	6 (10.7%)	56
D	0 (0.0%)	50 (89.3%)	0 (0.0%)	6 (10.7%)	56

A Figura 43 apresenta a relação EP *versus* RMSD das 1.000 conformações geradas, durante a execução do algoritmo, que apresentam a menor EP. De forma semelhante, aos resultados obtidos nos estudos de caso anteriores é possível verificar que, embora determinada conformação predita possua um valor de RMSD baixo, o valor de sua energia potencial pode ser alta.

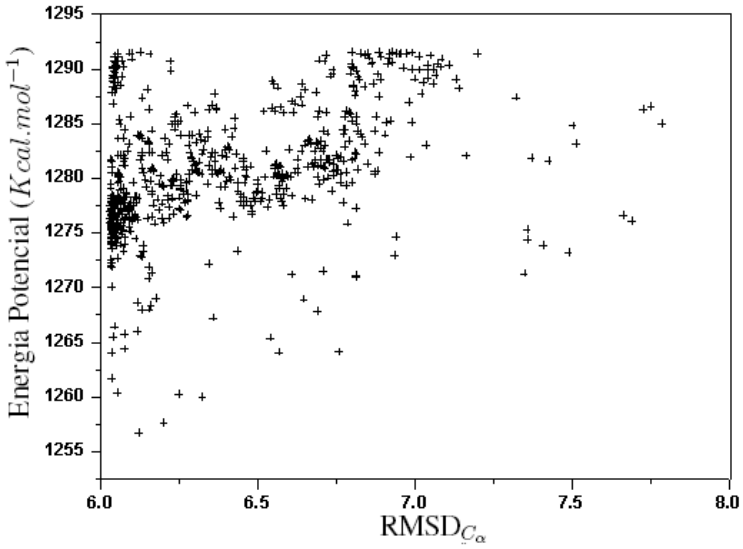

Figura 43 – Gráfico de energia *versus* RMSD. Relação entre a energia potencial e RMSD das 1.000 conformações com menor RMSD geradas pelo algoritmo de predição para a proteína cujo código PDB é 1ROP.

A Tabela 18, apresenta os valores de RMSD obtidos a partir da sobreposição das regiões de estruturas secundárias regulares da estrutura 3D experimental e da estruturas 3D final predita (Figura 41D). Os valores encontrados demonstram que as estruturas secundárias preditas estão bastante similares com às da estrutura 3D experimental da proteína-alvo.

Tabela 18: Valor de RMSD do Cα da estrutura 3D final predita em relação à estrutura 3D experimental da proteína cujo código PDB é 1ROP nas regiões de estruturas secundárias regulares.

Intervalo de aminoácidos (i-j)	RMSD Cα(Å)
3 - 28	0.90
32 - 54	0.80

Os mapas de Ramachandran da Figura 44 (A, B, C e D) demonstram que os resíduos de aminoácidos das estruturas 3D (B, C e D) preditas se encontram em regiões similares às ocupadas na estrutura 3D experimental da proteína-alvo. A porcentagem média de resíduos de aminoácidos das estruturas 3D preditas, que ocupam as regiões mais favoráveis no mapa de Ramachandran, é de aproximadamente 94,40%. Claramente, este valor, demonstra que a estrutura secundária das estruturas 3D preditas estão bem formadas (Tabela 19).

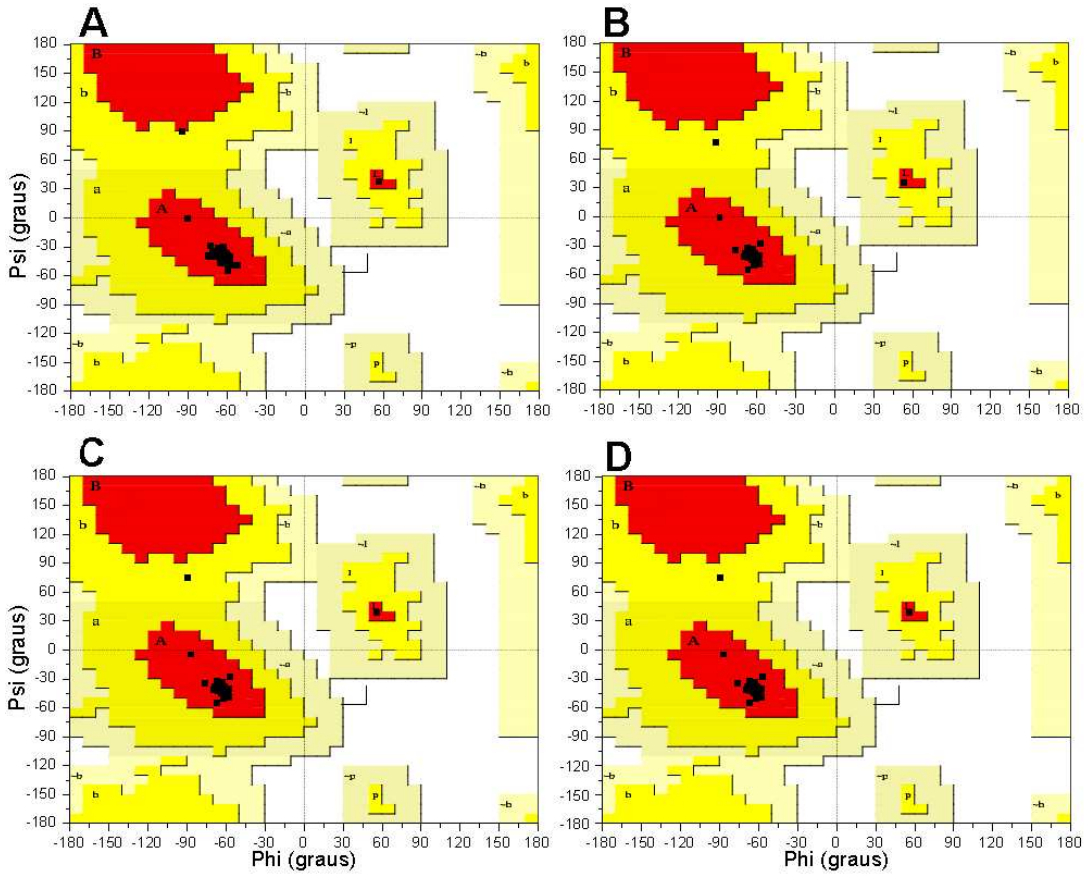

Figura 44 – Mapa de Ramachandran da estrutura 3D experimental e das estruturas 3D preditas da proteína cujo código PDB é 1ROP. (A) estrutura 3D experimental da proteína cujo código PDB é 1ROP; (B) estrutura 3D predita, obtida a partir do meio do intervalo; (C) estrutura 3D predita de menor RMSD encontrada durante a otimização da região de volta; (D) estrutura 3D predita representando o a estrutura final obtida pelo método de predição.

Tabela 19: Análise da localização dos resíduos de aminoácidos das estruturas 3D das conformações preditas para a proteína cujo código PDB é 1ROP no mapa de Ramachandran.

Estrutura	Mais favorável (%)	Favorável (%)	Aceitável (%)	Não aceitável (%)
A	98.10	1.90	0.00	0.00
B	94.40	5.60	0.00	0.00
C	94.40	3.70	1.90	0.00
D	94.40	3.70	1.90	0.00

Ao longo de toda a execução do algoritmo de otimização das regiões de volta foram geradas 2.300 conformações (momento em que atinge o critério de parada). Por meio da análise dos resultados obtidos pela execução do método proposto verifica-se que as estruturas 3D preditas apresentam resultados ótimos quando a organização das estruturas secundárias no espaço 3D e resultados ótimos no que se refere a formação das estruturas secundárias regulares.

5.6 Estudo de caso 4: 1GB1

No estudo de caso 4, realizou-se a predição da estrutura 3D aproximada da proteína cujo código PDB é 1GB1 [32], composta por 56 resíduos de aminoácidos e pertencente à classe das proteínas α/β [64] (código PDB: 1GB1 - Figura 45A).

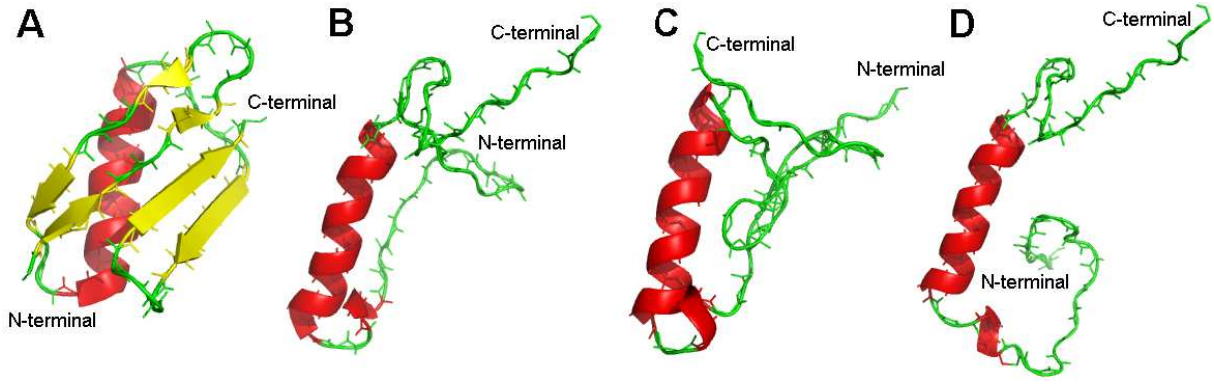

Figura 45 – Representação do tipo *Ribbon* da estrutura 3D experimental e das estruturas 3D preditas da proteína cujo código PDB é 1GB1. (A) estrututura 3D experimental da proteína cujo código PDB é 1GB1; (B) estrutura 3D predita, obtida a partir do centro do intervalo da conformação inicial; (C) estrutura 3D predita com menor RMSD em relação a estrutura experimental, encontrada ao longo da execução do método de predição; (D) estrutura 3D predita obtida após a otimização da região de volta. As cadeias laterais foram removidas para facilitar a visualização.

A seqüência alvo K=MTYKLILNGKTLKGETTTEAVDAATAEKVFKQYANDNGVDG-EWTYDDATKTFTVTE foi fragmentada em 54 fragmentos-alvo s_i de tamanho l=5 resíduos de aminoácidos (Apêndice C, Tabela 34, coluna 1). Para cada fragmentos-alvo s_i, foi realizada a busca por proteínas-molde no PDB (Apêndice C, Tabela 34, coluna 3, apresenta o número de moldes identificados para cada fragmento alvo s_i). Eliminou-se as proteínas cujas seqüências são idênticas ou muito similares à seqüência alvo K da proteína-alvo, identificados por: 1GB1 2GB1, 1PGA, 1PGB, 3GB1, 1GB4, 2IGG, 1QKZ, 1PGX, 1UWX, 2PLP, 1FD6, 1Q10, 1MPE, 1MVK, 1FCC, 1P7E, 1P7F, 2OED, 1FCL, 2IGH, 1IGD, 1IGC, 2IGD, 2NMQ, 1EM7, 1MHX, 1MI0, 2I2Y, 2I38, 1PN5, 1IBX, 2CWB, 2DEN, 2GI9.

Obtidos os arquivos *pdbs* do PDB, são calculados os ângulos de torção do aminoácido central dos fragmentos-molde (Apêndice C, Tabela 34, coluna 2). Os ângulos de torção de cada fragmento s_i são agrupados (Apêndice H, Tabela 39). Os intervalos são construídos. A partir da predição da estrutura secundária (Figura 46) é construída a conformação representada por intervalos de ângulos de torção. A Figura 45B apresenta a estrutura predita a partir do meio do intervalo da conformação inicial representada por intervalos.

```
                  10        20        30        40        50
                   |         |         |         |         |
                  MTYKLILNGKTLKGETTTEAVDAATAEKVFKQYANDNGVDGEWTYDDATKTFTVTE
MLRC              ceeeeeeeccccccccchhhhhhhhhhhhhhhhhcccccccceeeeccccceeeeeec
Consenso          ceeeeeeeccccccccchhhhhhhhhhhhhhhhhcccccccceeeeccccceeeeeec
```

Figura 46 – Predição da estrutura secundária da seqüência-alvo K da proteína cujo código PDB é 1GB1. O método MLRC utiliza os métodos GOR4, SIMPA96 e SOPMA para fazer a predição da estrutura secundária. O consenso representa a estrutura secundária predita pelo método MLRC.

As regiões de volta são otimizadas. A Figura 45C apresenta a estrutura com menor RMSD encontrada durante a execução do algoritmo de otimização de voltas. A Figura 45D apresenta a estrutura final predita pelo método desenvolvido. A Tabela 20 apresenta o valor de RMSD e de energia potencial das conformações preditas. A estrutura predita com menor RMSD em relação a estrutura experimental, encontrada ao longo de todo processo de otimização da região de volta, apresenta RMSD = 9.25 (Figura 45C). A estrutura de menor energia encontrada no último passo de execução apresenta um valor de RMSD = 11.18Å (Figura 45D).

Analisando as conformações preditas, demonstradas na Figura 45, é possível verificar que a estrutura secundária em forma de hélice α está bem formadas. Porém, as estruturas secundárias em forma de folhas β não estão formadas, apesar de os aminoácidos das estruturas preditas estarem em posições semelhantes às da estrutura experimental no mapa de Ramachandran (Figura 47 e Tabela 21).

Tabela 20: Valor de energia potencial ($Kcal.mol^{-1}$) estruturas 3D preditas e o valor de RMSD (Å) do C_α em relação à estrutura 3D experimental da proteína cujo código PDB é 1GB1.

Estrutura predita	RMSD(Å)	EP($Kcal.mol^{-1}$)
B	12.31	13480000000000.00
C	9.25	684400000000.00
D	11.18	201026141.36

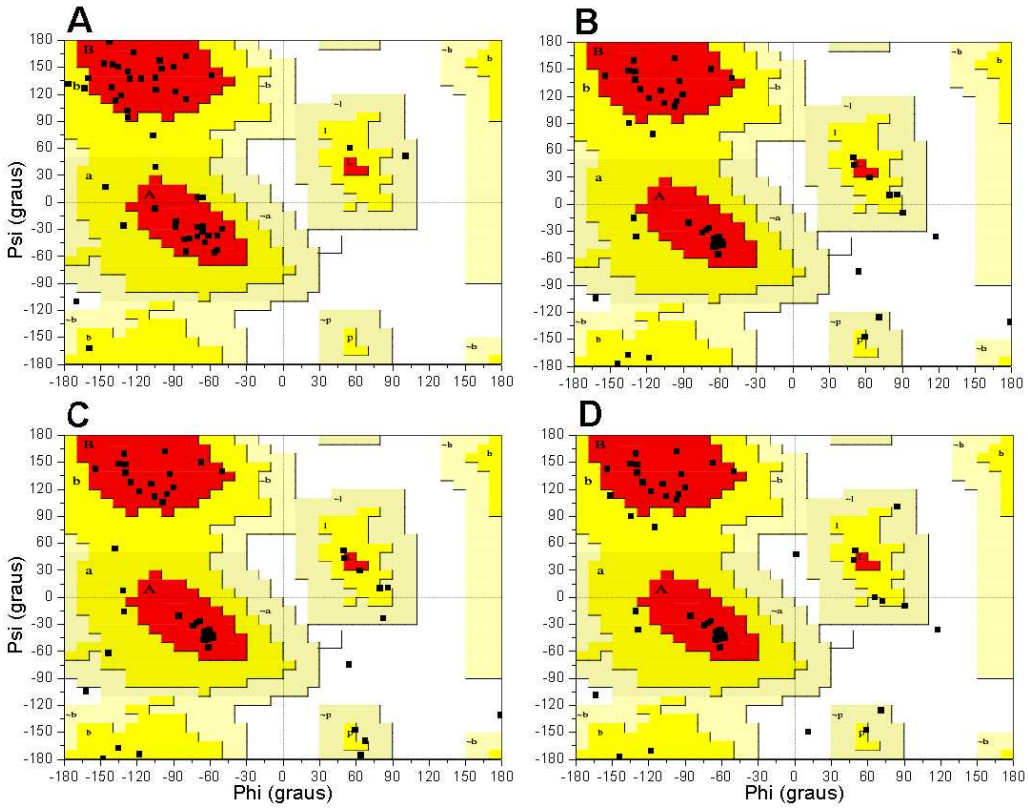

Figura 47 – Mapa de Ramachandran da estrutura 3D experimental e das estruturas 3D preditas da proteína cujo código PDB é 1GB1. (A) estrutura 3D experimental da proteína cujo código PDB é 1GB1; (B) estrutura 3D predita, obtida a partir do meio do intervalo; (C) estrutura 3D predita de menor RMSD encontrada durante a otimização da região de volta; (D) estrutura 3D predita representando o a estrutura final obtida pelo método de predição.

Tabela 21: Análise da localização dos resíduos de aminoácidos das estruturas 3D das conformações preditas para a proteína cujo código PDB é 1GB1 no mapa de Ramachandran.

Estrutura	Mais favorável (%)	Favorável (%)	Aceitável (%)	Não aceitável (%)
A	84.00	16.00	0.00	0.00
B	68.00	24.00	0.00	8.00
C	68.00	24.00	4.00	6.00
D	66.00	24.00	2.00	8.00

A Tabela 22 descreve a formação das estruturas secundárias. Conforme pode ser observado, as estruturas secundárias em forma de folha β não foram obtidas pela não formação de ligações de hidrogênio. No entando a estrutura secundária em forma de hélice α está formada adequadamente.

A Figura 48 apresenta a relação EP *versus* RMSD das 1.000 conformações geradas, durante a execução do algoritmo, que apresentam a menor EP. Novamente, a explosão da EP desencadeou uma série de perturbações na forma de redução dos intervalos de variação angular. Fazendo com que, as regiões de resíduos de aminoácidos referentes à folhas β) não puderam estar próximas o suficiente para formar ligações de hidrogênio. Estas ligações são necessárias para a estabilização e formação das estruturas regulares em forma de folha β.

A Tabela 23, apresenta os valores de RMSD obtidos a partir da sobreposição das regiões de estruturas secundárias regulares da estrutura 3D experimental e da estrutura 3D final predita (Figura 45D). A partir dos resultados obtidos, é possível verificar que as regiões estão bastante similares, exceto na região formada pelo intervalo 18-33 onde se observa um valor de RMSD = 4.77Å.

Tabela 22: Estrutura secundária das estruturas 3D preditas da proteína cujo código PDB é 1GB1. (A) estrutura 3D experimental da proteína cujo código PDB é 1GB1; (B) estrutura 3D predita, obtida a partir do centro do intervalo da conformação inicial representada por intervalos de variação angular; (C) estrutura 3D predita de menor RMSD encontrada ao longo da execução do método de otimização da região de volta; (D) estrutura 3D predita como resultado final do método de predição.

Estrutura	Folha β	Hélice α	Hélice α 3^{10}	Outras	Total resíduos
A	20 (35.7%)	13 (23.2%)	0 (0.0%)	23 (41.1%)	56
B	0 (0.0%)	16 (28.6%)	3 (5.4%)	37 (66.1%)	56
C	0 (0.0%)	17 (30.4%)	3 (5.4%)	36 (64.3.6%)	56
D	0 (0.0%)	16 (28.6%)	3 (54.4%)	37 (66.1%)	56

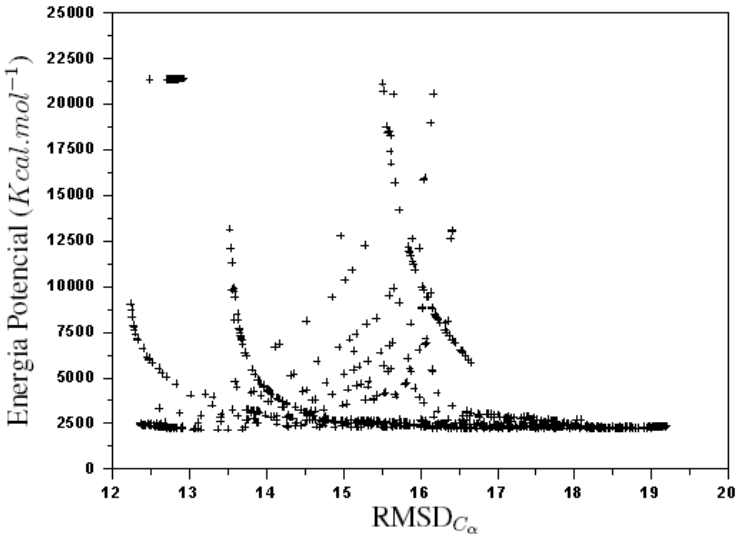

Figura 48 – Gráfico de energia *versus* RMSD. Relação entre a energia potencial e RMSD das 1.000 conformações com menor RMSD geradas pelo algoritmo de predição para a proteína cujo código PDB é 1GB1.

Tabela 23: Valor de RMSD do Cα da estrutura 3D final predita em relação à estrutura 3D experimental da proteína cujo código PDB é 1GB1 nas regiões de estruturas secundárias regulares.

Intervalo de aminoácidos (i-j)	RMSD Cα(Å)
3 - 8	0.22
18 - 33	4.77
42 - 45	0.20
50 - 54	1.06

O algoritmo de otimização das regiões de volta alcançou o seu critério de parada (tamanho do intervalo) após gerar 4.600 conformações. Por meio da análise dos resultados obtidos pela execução do método proposto verifica-se que as estruturas 3D preditas apresentam resultados regulares quanto à organização das estruturas secundárias no espaço 3D e resultados satisfatórios no que se refere a formação das estruturas secundárias do tipo hélice α. As estruturas secundárias em forma de folha β não foram formadas devido a não formação de ligações de hidrogênio entre regiões próximas da cadeia polipeptídica.

5.7 Estudo de caso 5: 1GBA

No estudo de caso 5, realizou-se a predição da estrutura 3D aproximada da proteína cujo código PDB é 1GAB [55], composta por 53 resíduos de aminoácidos (código PDB: 1GAB - Figura 37A).

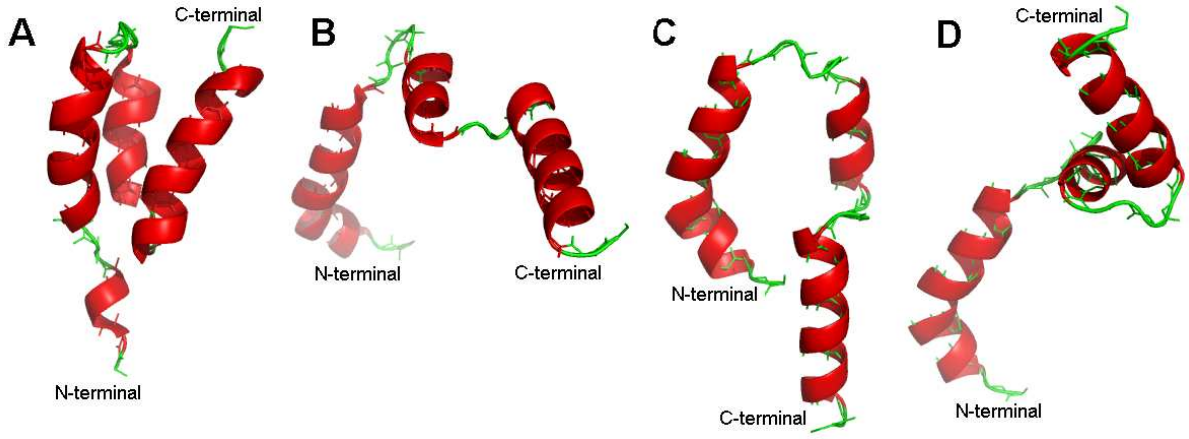

Figura 49 – Representação do tipo *Ribbon* da estrutura 3D experimental e das estruturas 3D preditas da proteína cujo código PDB é 1GAB. (A) estrututura 3D experimental da proteína cujo código PDB é 1GAB; (B) estrutura 3D predita, obtida a partir do centro do intervalo da conformação inicial; (C) estrutura 3D predita com menor RMSD em relação a estrutura experimental, encontrada ao longo da execução do método de predição; (D) estrutura 3D predita obtida após a otimização da região de volta. As cadeias laterais foram removidas para facilitar a visualização.

A seqüência alvo K=TIDQWLLKNAKEDAIAELKKAGITSDFYFNAINKAKTVEEVNA LKNEILKAHA é fragmentada em 49 fragmentos-alvo s_i de tamanho l=5 resíduos de aminoácidos (Apêndice D, Tabela 35, coluna 1). Para cada fragmento alvo s_i, foi realizada a busca por proteínas-molde no PDB. O Apêndice D, Tabela 35, coluna 3 apresenta o número de moldes identificados para cada fragmento molde. Eliminou-se as proteínas cujas seqüências são idênticas ou muito similares à seqüência alvo K, identificadas por: 1GAB, 2J5Y, 1GSJ, 1F6G, 1FFK, 1JQ1, 1LRP, 1MI6, 1ML5, 1PNU, 1POS, 1PQV, 1XI4, 2OCW, 4CRO.

Após obter os arquivos pdbs do PDB são calculados os ângulos de torção do aminoácido central de cada fragmento-molde (Apêndice D, Tabela 35, coluna 3). O Apêndice D, Tabela 35, coluna 4, 5 e 6 apresenta a porcentagem das tuplas-molde associadas a cada um dos três estados conformacionais (h, b ou c). Nesta classificação, o estado conformacional de hélice α (h) compreende os resíduos de aminoácidos nos estados "A", "a", "L", "l"e "p", o estado de folha β (b) compreende os resíduos de aminoácidos em estados "B"e "b"e as regiões de volta (c) compreendem os resíduos de aminoácidos em estado "c"(seguindo o modelo de 8 estados descrito na seção 4.3.6 e baseando-se no modelo para escolha dos grupos apresentados na seção 4.3.8).

Os ângulos de torção de cada fragmento s_i são agrupados (Apêndice I, Tabela 40). A partir do valor médio e do valor de desvio padrão estimado de cada grupo k_i de s_i é criado o intervalo de variação de cada grupo. Em seguida, a partir do centro do intervalo, cada grupo é rotulado em um dos 8 estados conformacionais. A estrutura secundária é predita para a seqüência-alvo K (Figura 50). A partir do resultado da predição da estrutura secundária é construída a conformação inicial representada na forma de intervalos. A Figura 49B apresenta a estrutura 3D obtida a partir do centro do intervalo da conformação inicial representada por intervalos de variação angular.

```
                  10        20        30        40        50
                  |         |         |         |         |
          TIDQWLLKNAKEDAIAELKKAGITSDFYFNAINKAKTVEEVNALKNEILKAHA
MLRC      ccchhhhhhhhhhhhhhhhccccccchhhhhhhccchhhhhhhhhhhhhhhcc
PHD       ccchhhhhhhhhhhhhhhhccccchhhhhhhhhhhhhhhhhhhhhhhhhhcc
Predator  ccchhhhhccchhhhhhhhhccccccchhhhhhccccchhhhhhhhhhhhcccc
Consenso  ccchhhhhhhhhhhhhhhhccccccchhhhhhhccchhhhhhhhhhhhhhcc
```

Figura 50 – Predição da estrutura secundária da seqüência-alvo K da proteína cujo código PDB é 1GAB. O consenso representa a estrutura secundária obtida pela análise simultânea da predição realizada pelo método MLRC, PHD e Predator.

As regiões de volta são otimizadas. A Figura 49C apresenta a estrutura 3D com menor RMSD encontrada durante a execução do algoritmo de otimização das regiões de volta. A Figura 49D apresenta a estrutura 3D final predita pelo método de predição. A Tabela 24 apresenta o valor de EP e de RMSD das conformações preditas. A estrutura 3D com menor RMSD, em relação à estrutura 3D experimental, encontrada ao longo de todo o processo de otimização das regiões de volta possui um valor de RMSD = 9.56Å. A estrutura final predita pelo algoritmo apresenta um valor de RMSD = 11.91 (Figura 49D).

Tabela 24: Valor de energia potencial ($Kcal.mol^{-1}$) estruturas 3D preditas e o valor de RMSD (Å) do C_α em relação à estrutura 3D experimental da proteína cujo código PDB é 1GAB.

Estrutura predita	RMSD(Å)	EP($Kcal.mol^{-1}$)
B	13.57	223170.63
C	9.56	5460000000000.00
D	11.91	154263.98

Os mapas de Ramachandran da Figura 51 (A, B, C e D) demonstram que os resíduos de aminoácidos das estruturas 3D preditas se encontram em regiões similares às ocupadas na estrutura 3D experimental. A porcentagem média de resíduos de aminoácidos das estruturas 3D preditas, que ocupam as regiões mais favoráveis no mapa de Ramachandran, é de aproximadamente 78,53%. Claramente, este valor, demonstra que a estrutura secundária das estruturas 3D preditas estão bem formadas (Tabela 25).

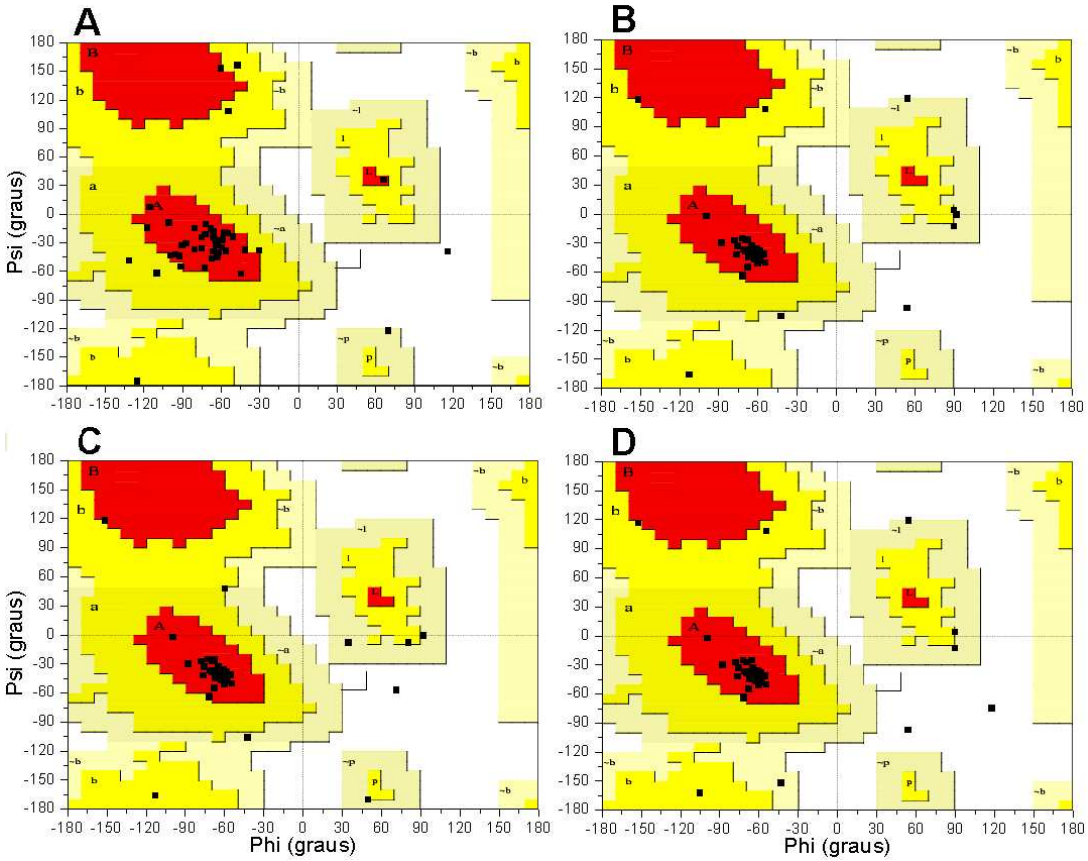

Figura 51 – Mapa de Ramachandran da estrutura 3D experimental e das estruturas 3D preditas da proteína cujo código PDB é 1GAB. (A) estrutura 3D experimental da proteína cujo código PDB é 1GAB; (B) estrutura 3D predita, obtida a partir do meio do intervalo; (C) estrutura 3D predita de menor RMSD encontrada durante a otimização da região de volta; (D) estrutura 3D predita representando o a estrutura final obtida pelo método de predição.

A Tabela 26 descreve a formação das estruturas secundárias. Conforme pode ser observado, as estruturas secundárias em forma de hélice α estão bem formadas.

Tabela 25: Análise da localização dos resíduos de aminoácidos das estruturas 3D das conformações preditas para a proteína 1GAB no mapa de Ramachandran.

Estrutura	Mais favorável (%)	Favorável (%)	Aceitável (%)	Não aceitável (%)
A	80.00	18.00	2.00	0.00
B	78.00	12.00	8.00	2.00
C	78.00	10.00	10.00	2.00
D	79.60	10.20	6.10	4.10

Tabela 26: Estrutura secundária das estruturas 3D preditas da proteína cujo código PDB é 1GAB. (A) estrutura 3D experimental da proteína cujo código PDB é 1GAB; (B) estrutura 3D predita, obtida a partir do centro do intervalo da conformação inicial representada por intervalos de variação angular; (C) estrutura 3D predita de menor RMSD encontrada ao longo da execução do método de otimização da região de volta; (D) estrutura 3D predita como resultado final do método de predição.

Estrutura	Folha β	Hélice α	Hélice α 3^{10}	Outras	Total resíduos
A	0 (0.00%)	35 (66.00%)	4 (7.50%)	14 (26.40%)	53
B	0 (0.00%)	40 (75.50%)	0 (0.00%)	13 (24.50%)	53
C	0 (0.00%)	40 (75.50%)	0 (0.00%)	13 (24.50%)	53
D	0 (0.00%)	39 (73.60%)	0 (0.00%)	14 (26.40%)	53

A Tabela 27, apresenta os valores de RMSD obtidos a partir da sobreposição das regiões de estruturas secundárias regulares da estruturas 3D experimental e da estrutura 3D final predita (Figura 49 D) pelo algoritmo de predição. Os valores obtidos mostram que as estruturas secundárias regulares estão bem formadas, especialmente na região 27 - 33 e na região 37 - 52.

Tabela 27: Valor de RMSD do Cα da estrutura 3D final predita em relação à estrutura 3D experimental da proteína cujo código PDB é 1GAB nas regiões de estruturas secundárias regulares.

Intervalo de aminoácidos (i-j)	RMSD Cα(Å)
4 - 20	3.15
27 - 33	0.36
37 - 52	1.05

A Figura 56 apresenta a relação EP *versus* RMSD das 1.000 conformações geradas, durante a execução do algoritmo, que apresentam a menor EP. Novamente, a explosão da EP pode ter desencadeado uma série de perturbações na forma de redução dos intervalos de variação angular, fazendo com que a redução do intervalo fosse afetada.

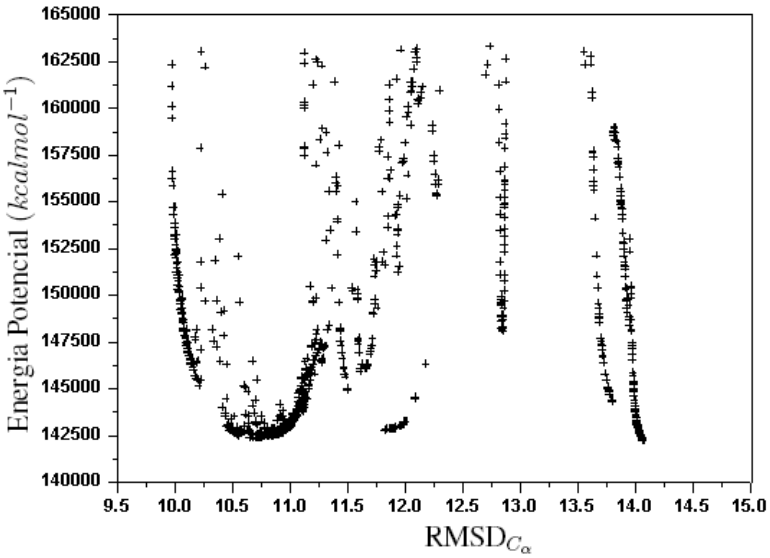

Figura 52 – Gráfico de energia *versus* RMSD. Relação entre a energia potencial e RMSD das 1.000 conformações com menor RMSD geradas pelo algoritmo de predição para a proteína cujo código PDB é 1GAB.

Ao longo da execução do algoritmo de otimização das regiões de volta foram geradas 3.800 conformações (momento em que atinge o critério de parada). Por meio da análise dos resultados obtidos pela execução do método de predição verifica-se que as estruturas 3D preditas apresentam resultados regulares quando a organização das estruturas secundárias no espaço 3D e resultados ótimos no que se refere a formação das estruturas secundárias regulares em forma de folha β.

5.8 Estudo de caso 6: 1UTG

No estudo de caso 6, testou-se o método desenvolvido na predição da estrutura 3D aproximada da proteína 1UTG [61], composta por 70 resíduos de aminoácidos (código PDB: 1UTG - Figura 53 A).

A seqüência alvo K=GICPRFAHVIENLLLGTPSSYETSLKEFEPDDTMKDAGMQMKK VLDSLPQTTRENIMKLTEKIVKSPLCM foi fragmentada em 66 fragmentos s_i (Apêndice E, Tabela 36, coluna 1). Para cada fragmento alvo s_i, foi realizada a busca por proteínas moldes no PDB (Apêndice E, Tabela 36, coluna 3, apresenta o número de moldes identificados para cada fragmento alvo s_i). Eliminou-se os arquivos *pdbs* que possuem seqüências idênticas ou muito similares à seqüência alvo K da proteína 1UTG, identificados por: 1UTG, 2UTG e 1UTR.

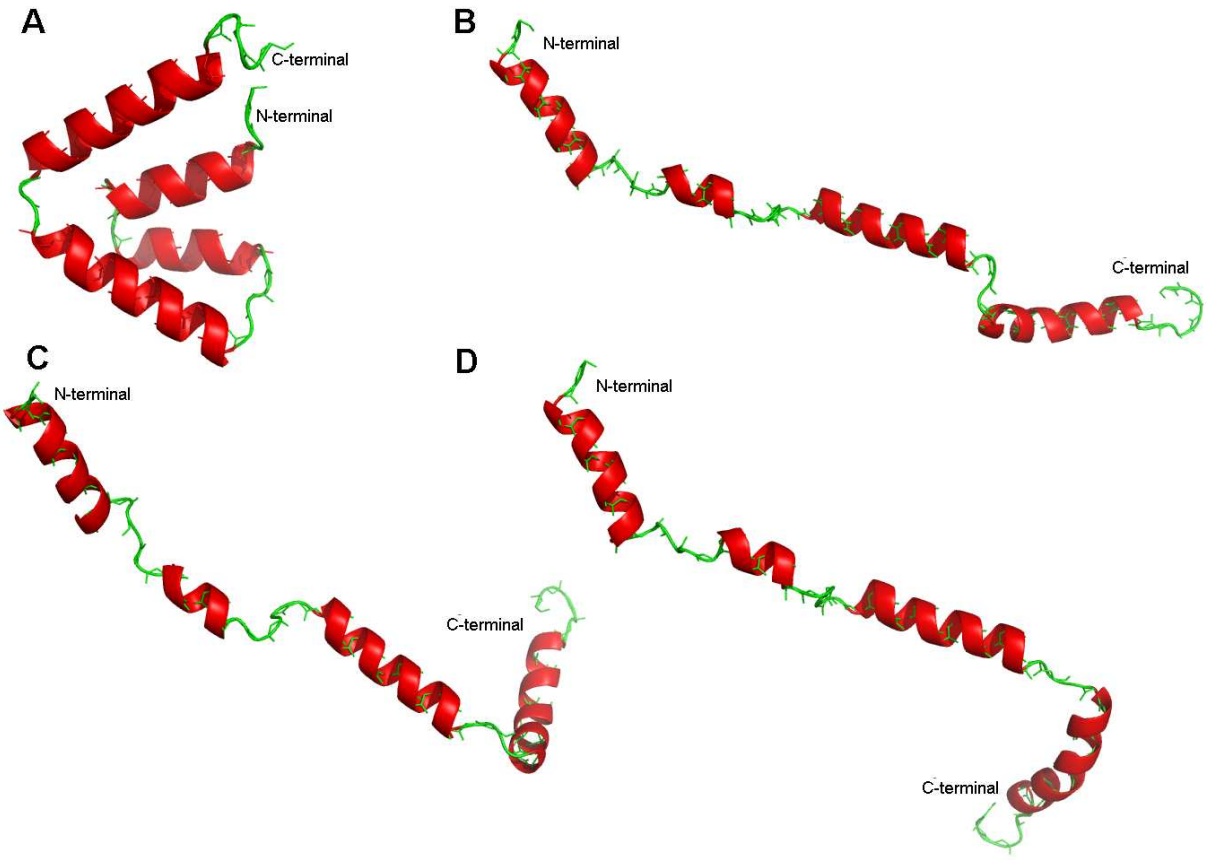

Figura 53 – Representação do tipo *Ribbon* da estrutura 3D experimental e das estruturas 3D preditas da proteína 1UTG. (A) estrutura 3D experimental da proteína cujo código PDB é 1UTG; (B) estrutura 3D predita, obtida a partir do centro do intervalo da conformação inicial; (C) estrutura 3D predita com menor RMSD em relação a estrutura experimental, encontrada ao longo da execução do método de predição; (D) estrutura 3D predita obtida após a otimização da região de volta. As cadeias laterais foram removidas para facilitar a visualização.

Os ângulos de torção das proteínas-molde são calculados (Apêndice E, Tabela 36, coluna 2). O Apêndice E, Tabela 36, coluna 4, 5 e 6 apresenta a porcentagem das tuplas-molde associadas a cada um dos três estados conformacionais (h, b ou c). Nesta classificação, o estado conformacional de hélice α (h) compreende os resíduos de aminoácidos nos estados "A", "a", "L", "l"e "p", o estado de folha β (b) compreende os resíduos de aminoácidos em estados "B"e "b"e as regiões de volta (c) compreendem os resíduos de aminoácidos em estado "c"(seguindo o modelo de 8 estados descrito na seção 4.3.6 e baseando-se no modelo para escolha dos grupos apresentados na seção 4.3.8).

Os ângulos de torção de cada fragmento s_i são agrupados (Apêndice J, Tabela 41). A partir do valor médio e do valor de desvio padrão estimado de cada grupo k_i de s_i é criado o intervalo de variação de cada grupo. Em seguida, a partir do centro do intervalo, cada grupo é rotulado em um dos 8 estados conformacionais.

A partir da predição da estrutura secundária (Figura 54) é construída a conformação repre-

sentada por intervalos de ângulos de torção. A Figura 53B apresenta a estrutura predita a partir do meio do intervalo da conformação inicial representada por intervalos.

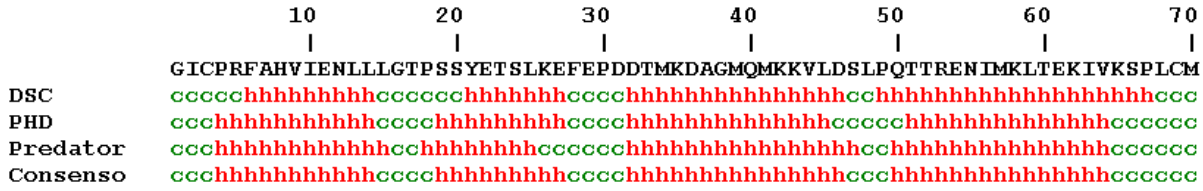

Figura 54 – Predição da estrutura secundária para a seqüência alvo K da proteína 1UTG (PDB ID: 1UTG).

As regiões de volta são otimizadas. A Figura 53C apresenta a estrutura 3D com menor RMSD encontrada durante a execução do algoritmo de otimização das regiões de volta. A Figura 53D apresenta a estrutura 3D final predita pelo método de predição. A Tabela 28 apresenta o valor de EP e de RMSD das conformações preditas.

Tabela 28: Valor de energia potencial ($Kcal.mol^{-1}$) estruturas 3D preditas e o valor de RMSD (Å) do C_α em relação à estrutura 3D experimental da proteína cujo código PDB é 1UTG.

Estrutura predita	RMSD(Å)	EP($Kcal.mol^{-1}$)
B	26.23	4656.80
C	23.53	4469.70
D	23.90	4510.10

A Tabela 29 descreve a formação das estruturas secundárias. Conforme pode ser observado, as estruturas secundárias em forma de hélice α estão bem formadas.

Tabela 29: Estrutura secundária das estruturas 3D preditas da proteína cujo código PDB é 1UTG. (A) estrutura 3D experimental da proteína cujo código PDB é 1UTG; (B) estrutura 3D predita, obtida a partir do centro do intervalo da conformação inicial representada por intervalos de variação angular; (C) estrutura 3D predita de menor RMSD encontrada ao longo da execução do método de otimização da região de volta; (D) estrutura 3D predita como resultado final do método de predição.

Estrutura	Folha β	Hélice α	Hélice α 3^{10}	Outras	Total resíduos
A	0 (0.00%)	50 (71.40%)	3 (4.30%)	17 (24.30%)	70
B	0 (0.00%)	41 (58.60%)	5 (7.10%)	24 (34.30%)	70
C	0 (0.00%)	41 (58.60%)	5 (7.10%)	24 (34.30%)	70
D	0 (0.00%)	41 (58.60%)	5 (7.10%)	24 (34.30%)	70

A Tabela 30, apresenta os valores de RMSD obtidos a partir da sobreposição das regiões de estruturas secundárias regulares da estruturas 3D experimental e da estrutura 3D predita (Figura 53 D). Os valores obtidos mostram que as estruturas secundárias regulares estão bem formadas.

Tabela 30: Valor de RMSD do Cα da estrutura 3D final predita em relação à estrutura 3D experimental da proteína cujo código PDB é 1UTG nas regiões de estruturas secundárias regulares.

Intervalo de aminoácidos (i-j)	RMSD Cα(Å)
4 - 14	0.70
19 - 27	1.19
32 - 46	0.55
50 - 64	1.42

Os mapas de Ramachandran da Figura 55 demonstram que os resíduos de aminoácidos das estruturas 3D (B, C e D) preditas se encontram em regiões similares às ocupadas na estrutura 3D experimental (A). A porcentagem média de resíduos de aminoácidos das estruturas 3D preditas, que ocupam as regiões mais favoráveis no mapa de Ramachandran, é de aproximadamente 81,66%. Claramente, este valor, demonstra que a estrutura secundária das estruturas 3D preditas estão bem formadas (Tabela 31).

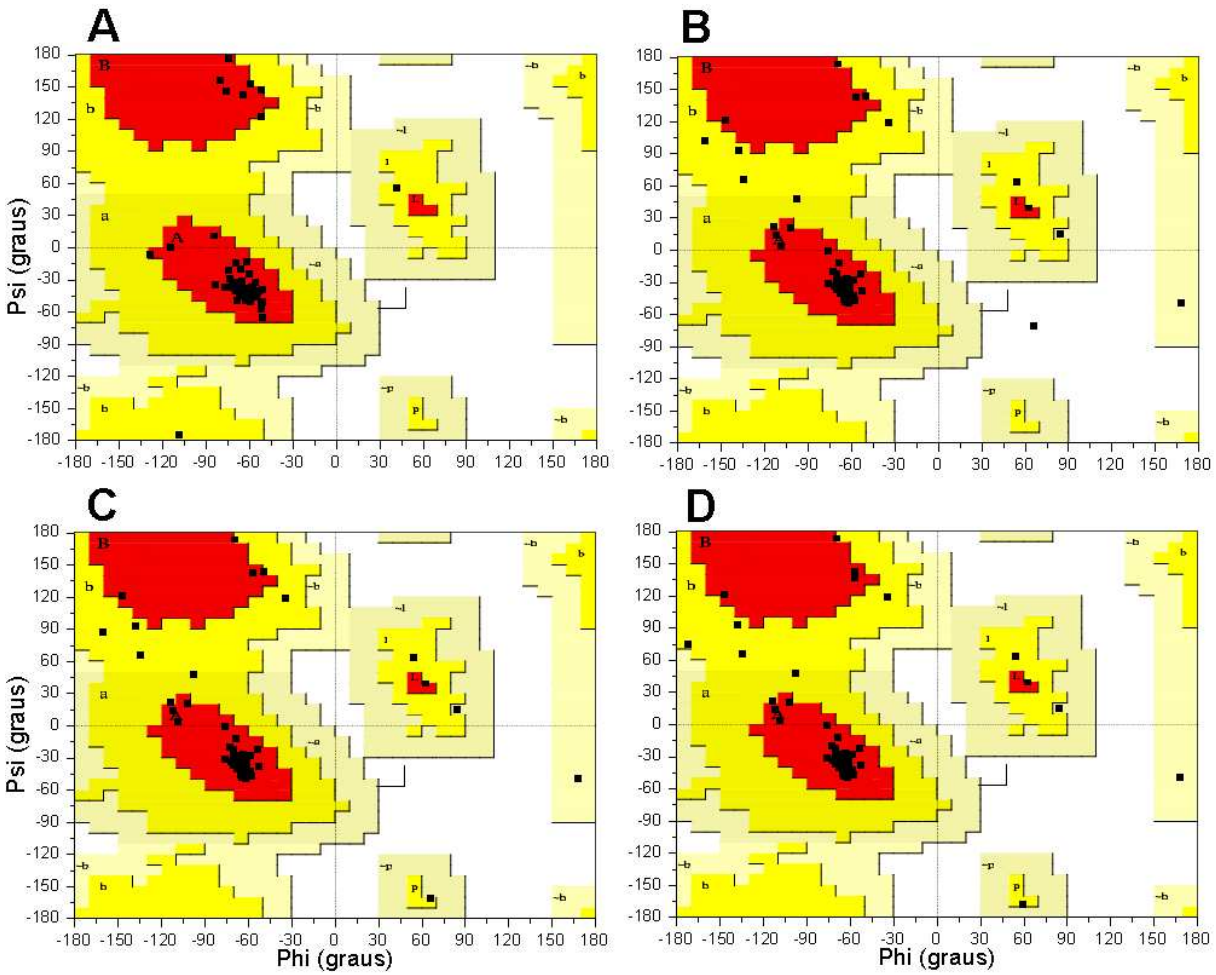

Figura 55 – Mapas de Ramachandran da estrutura experimental e das estruturas preditas da proteína 1UTG. (A) 1UTG experimental; (B) 1UTG predita obtida a partir do meio do intervalo.

Tabela 31: Análise da localização dos resíduos de aminoácidos das estruturas 3D das conformações preditas para a proteína cujo código PDB é 1UTG no mapa de Ramachandran.

Estrutura	Mais favorável (%)	Favorável (%)	Aceitável (%)	Não aceitável (%)
A	96.70	3.30	0.00	0.00
B	81.00	13.80	3.40	1.70
C	82.00	16.40	1.60	0.00
D	82.00	14.80	3.30	0.00

A Figura 56 apresenta a relação EP *versus* RMSD das 1.000 conformações geradas, durante a execução do algoritmo, que apresentam a menor EP. Novamente, a explosão da EP pode

ter desencadeado uma série de perturbações na forma de redução dos intervalos de variação angular, fazendo com que a redução do intervalo fosse afetada.

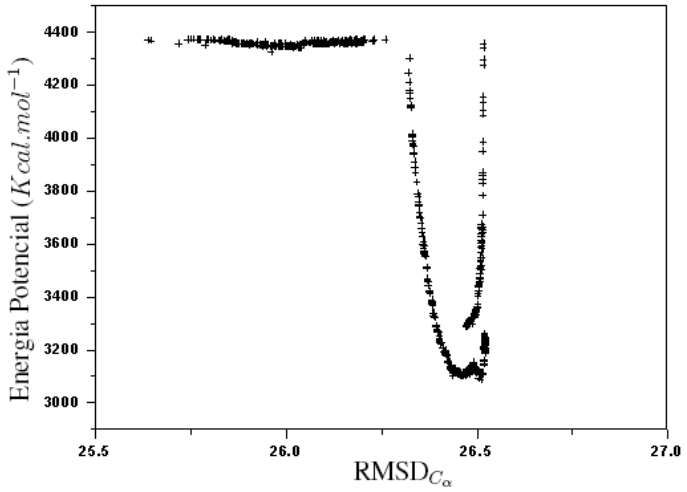

Figura 56 – Gráfico de energia *versus* RMSD. Relação entre a energia potencial e RMSD das 1.000 conformações com menor RMSD geradas pelo algoritmo de predição para a proteína cujo código PDB é 1UTG.

Nete estudo de caso foram geradas 4.000 conformações. Por meio da análise dos resultados obtidos pela execução do método proposto verifica-se que as estruturas 3D preditas apresentam resultados não satisfatórios quanto à organização das estruturas secundárias no espaço 3D e resultados satisfatórios no que se refere a formação das estruturas secundárias do tipo hélice α.

5.9 Tempo de processamento

Após obter as proteínas-modelo do PDB, as conformações iniciais representadas por intervalos de variação angular de cada proteína-alvo foram construídas em poucos minutos (menos de 5 minutos). A etapa de redução do intervalo buscando a otimização das regiões de volta foram as que demandaram maior rempo de processamento. Paras as proteínas com uma única estrutura irregular (Código PDB: 1ZDD, 1ROP, 1K43) foram necessárias de 24 a 48 horas para que o algoritmo de predição atingisse o critério de parada. Para as proteínas com mais regiões de volta esse tempo de processamento aumenta, para as proteínas cujo código PDB é 1GB1, 1GAB e 1UTG foram necessárias de 48 a 96 horas para que o algoritmo atingisse o critério de parada.

5.10 Resumo do capítulo

Neste capítulo foram apresentados os testes realizados com o método de predição desenvolvido. Foi também realizada a análise da qualidade das conformações que foram preditas. No próximo capítulo são feitas as considerações finais do trabalho realizado juntamente com a avaliação geral dos resultados obtidos com o método de predição.

6 Considerações finais

A predição da estrutura 3D de uma proteína é atualmente um dos maiores problemas da Bioinformática Estrutural. O principal desafio é compreender como a informação codificada na seqüência linear de aminoácidos (estrutura primária) traduz-se na estrutura 3D (estrutura terciária) de uma proteína, e a partir deste conhecimento, desenvolver metodologias computacionais que possam predizer, de forma correta, a estrutura nativa e funcional da proteína. Ao longo dos anos, muitas metodologias e algoritmos foram propostos buscando resolver este problema complexo. Métodos baseados em conhecimento e métodos *ab initio* foram propostos. No entanto, limitações impostas pelo tamanho do espaço de busca conformacional (*ab initio* e *de novo*) e pela impossibilidade de obtenção de novas formas de enovelamento (Modelagem comparativa por homologia e via alinhavamento) ainda impedem que o problema da predição tenha uma solução.

Nesta dissertação foi apresentado um novo método para a predição computacional da estrutura 3D aproximada de polipeptídeos. O método proposto agrega princípios dos métodos de predição *de novo* e de modelagem comparativa por homologia, beneficiando-se da capacidade de predição de novas formas de enovelamento e da alta acurácia nas predições, respectivamente. A construção de intervalos de variação angular para representar as torções dos ângulos diedros (ϕ e ψ) de todos os resíduos de aminoácidos da cadeia principal do polipeptídeo-alvo, reduz drasticamente o espaço de busca conformacional. Este problema esta presente nos métodos puramente *ab initio* e em grande parte dos métodos *de novo*. Ao criar um intervalo fechado de variação para cada ângulo diedro de um resíduo de aminoácido de um polipeptídeo-alvo limita-se os possíveis estados conformacionais que este resíduo de aminoácido pode assumir. O agrupamento das tuplas de ângulos de torção de proteínas-molde permite identificar a região no mapa de Ramachandran onde estas estão mais concentradas. A partir desta informação, é possível estimar a conformação que o polipeptídeo-alvo estará assumindo. A predição da estrutura secundária permite a escolha correta dos intervalos de ângulos de torção que representarão as torções de cada resíduo de aminoácido. Com isto, é possível construir a conformação representada por intervalos de variação angular.

Nos métodos *de novo*, um dos principais problemas é atuar na combinação adequada dos fragmentos. No método desenvolvido este problema não existe devido a forma em que a informação das proteínas-molde é obtida e à forma que a conformação da proteína-alvo é construída. Esta forma de construção é que possibilita que novas formas de enovelamento sejam preditas.

Ao analisar os resultados obtidos, com o método de predição desenvolvido, é possível verificar que em termos de formação das estruturas secundárias regulares em forma de hélice α, o

método apresenta resultados bastante satisfatórios. A formação de folhas β, em alguns casos, não ocorreram pela não formação de ligações de hidrogênio. No entanto, ao comparar a localização dos resíduos de aminoácidos das estruturas 3D preditas e da estrutura 3D experimental no mapa de Ramachandran é possível verificar que as estruturas obtidas ocupam regiões muito próximas às ocupadas pela estrutura 3D experimental. Isto mostra, que as estruturas na forma de folhas β somente não foram formadas, em alguns casos, devido à sua organização inadequada no espaço 3D. Esta desorganização ocorreu por problemas na modelagem das regiões de volta. Por serem estruturas irregulares, as regiões de voltas são as mais difíceis de serem preditas. Mesmo seguindo o conjunto de regras criado, o algoritmo de predição, para alguns casos, não obteve resultados satisfatórios na modelagem das regiões de volta.

Um dos possíveis problemas enfrentados durante a redução do intervalo nas regiões de volta são os choques estérioqupimicos entre átomos da cadeia lateral do polipeptídeo. Conforme observado durante a execução do algoritmo de redução do intervalo das regiões de volta, este problema veio a provocar tomadas de decisões incorretas para determinar a forma de redução do intervalo. Este problema apresentou-se como um empecilho para que qualidade das predições realizadas pelo método não fossem melhores. Entretanto ao comparar os resultados obtidos pelo algoritmo de predição com os resultados do CASP VI verifica-se que os melhores métodos de predição *de novo* apresentam valores de RMSD variando entre 4.00Å e 6.00Å na predição de proteínas com até 200 resíduos de aminoácidos [63], o que é um valor compatível com as predições realizadas pelo método desenvolvido. Na versão do CASP VII (2006), o método ROSETTA, para proteínas de até 200 resíduos de aminoácidos, obteve-se resultados de RMSD entre o C_α da estrutura predita e da estrutura experimental, na faixa de 1.40 Åà 5.10 Å(Figura 5 - *Examples of successful free modelling predictions* encontrada em [19]). Experimentos realizados por Cutello *et. al* [18] na predição da estrutura 3D da proteína 1ZDD e 1ROP obtiveram estruturas com variação de 2.00Å à 6.00Å, o que são valores de RMSD compatíveis com os valores encontrados nas predições realizadas pelo método desenvolvido.

Um outro ponto a ser destacado é o tempo de processamento demandado pelo método de predição desenvolvido e o tempo gasto pelos métodos de predição existentes. No CASP VII, por exemplo, o método ROSETTA, contava com uma infraestrutura computacional de 140.000 computadores, com aproximadamente 65.000 computadores para serem utilizados em um mesmo tempo. Esta estrutura, representa uma capacidade de processamento de 37 TFlops, o que possibilitou que as predições de estruturas 3D de cada proteína testada fossem feitas em aproximadamente 500.000 horas de processamento [19]. O algoritmo de predição desenvolvido por sua vez, obteve resultados satisfatórios em um espaço de tempo muito inferior aos descritos acima para proteínas de similar número de resíduos de aminoácidos. A necessidade de uma plataforma computacional de baixo custo é outro ponto positivo no método de predição desenvolvido. Enquanto os demais métodos de predição utilizam plataformas de alto desempenho e de elevado custo, os testes desenvolvidos com o método de predição desenvolvido foram realizados com sucesso utilizando apenas um computador do tipo PC.

Apesar dos choques estereoquímicos entre os átomos afetarem a qualidade das estruturas 3D preditas, estas ainda se apresentam como um bom ponto de partida para refinamentos por métodos de mecânica molecular. Uma estratégia de refinamento das estruturas 3D aproximadas preditas pelo método desenvolvido poderia em pouco espaço de tempo, aumentar a qualidade das predições. Além disto, estas estruturas servindo como ponto de partida para métodos puramente *ab initio*, podem inferir numa redução drástica do espaço de busca conformacional, o qual, é um grande problema nestes métodos de predição.

O método de predição desenvolvido consegue gerar em um curto espaço de tempo estruturas 3D aproximadas da estrutura 3D experimental. Atualmente, não existe nenhuma metodologia computacional que consiga predizer estruturas 3D exatas às estruturas 3D experimentais. Desta forma, considera-se que método desenvolvido tem uma boa acurácia em comparação com os atuais métodos, sendo ainda uma das caracteristicas positivas o fato de consumir pouco tempo para realizar as predições.

6.1 Principais contribuições

As principais contribuições do presente trabalho foram:

- O desenvolvimento de um novo método para predição *in sílico* da estrutura 3D aproximada de polipeptídeos;
- A identificação de formas para obtenção e utilização de informações de estruturais 3D experimentais de proteínas;
- A identificação de estratégias para redução do espaço conformacionail, através da construção de conformações aproximadas de polipeptídeos;
- O desenvolvimento de formas de representação da estrutura 3D de um polipeptídeo na forma de intervalos de variação angular;
- O desenvolvimento de uma estratégia para redução dos intervalos de variação angular objetivando a otimização das regiões de volta da estrutura inicial predita.
- A publicação de dois artigos científicos em eventos:
  - **Artigo 1:**
    Título: CReF: A central-residue-fragment-based method for predicting approximate 3-D polypeptides structures.
    Publicação em: Proceedings of the 23th ACM SAC 2008, Annual ACM Symposium on Applied Computing.
    Ano: 2007 (aceitação), 2008 (publicação).

Tipo: artigo completo.
Qualificação CAPES: Congresso Internacional A (CIA) para a área da computação.
Citação: DORN, M. ; SOUZA, O. Norberto . CReF: A central-residue-fragment-based method for predicting approximate 3-D polypeptides structures. In: Annual ACM Symposium on Applied Computing, 2007, Fortaleza. Proceedings of the 23th ACM SAC 2008, 2008.

– **Artigo 2:**
Título: A fragment-based clustering method for predicting approximate 3-D polypeptide structures. *Publicação em:* Proceedings of the 3rd Conference of the Brazilian Association for Bioinformatics and Computational Biology (X meeting 2007).
Ano: 2007 (aceitação), 2007 (publicação).
Tipo: resumo extendido.
Qualificação CAPES: não consta no Qualis da CAPES.
Citação: DORN, M. ; SOUZA, O. Norberto . A fragment-based clustering method for predicting approximate 3-D polypeptide structures. In: Proceedings of the 3rd Conference of the Brazilian Association for Bioinformatics and Computational Biology (X meeting), 2007, São Paulo. , 2007.

6.2 Trabalhos futuros

Os trabalhos futuros decorrentes desta dissertação estão focados:

- Na análise de estratégias para amenizar o problema do aumento da energia potencial, decorrente do choque entre átomos, buscando aumentar a acurácia das predições. A resolução deste problema pode fazer com que o algoritmo de otimização das regiões de volta consiga proceder de forma correta a redução dos intervalos, e com isto possa melhorar a qualidade das predições;

- No desenvolvimento de novas estratégias para modelar as regiões de estruturas irregulares. Novos critérios para escolha dos grupos que representam os resíduos de aminoácidos das regiões de estruturas irregulares considerando a acessibilidade de solvente e o empacotamento das estruturas regulares podem ser considerados. Com isto, poderia haver uma melhora na modelagem das regiões de volta;

- No estudo do desenvolvimento de novos algoritmos de clusterização e formas de tratamento das informações obtidas de proteínas-molde;

- No desenvolvimento de técnicas para refinamento da estrutura 3D final predita pelo método. Como por exemplo, o desenvolvimento de protocolos de refinamento para métodos de dinâmica molecular.

Referências

[1] G. Alefeld e D. Claudio. The basic properties of interval arithmetic, its software realizations and some applications. *Computer and Structures*, 67(1):3–8, 1998.

[2] F. Allen e etal. Blue Gene: a vision for protein science using a petaflop supercomputer. *IBM Systems Journal*, 40(2):310–327, 2001.

[3] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, e D. J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, 1997.

[4] C. B. Anfinsen, E. Haber, M. Sela, e Jr. F. H. White. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proceedings of the National Academy of Sciences of the United States of America*, 47:1309–1314, 1961.

[5] D. W. Banner, M. Kokkinidis, e D. Tsernoglou. Structure of the ColE1 rop protein at 1.7 A resolution. *Journal of Molecular Biology*, 196(3):657–675, 1987.

[6] A. D. Baxevanis e B. F. F. Ouellette. *Bioinformatics: a practical guide to the analysis of genes and proteins*. John Wiley and Sons, Inc., New Jersey, EUA, 3° edição, 2005.

[7] H. Berman, K. Henrick, H. Nakamura, e J. L. Markley. The worldwide protein data bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Research*, 35(Database Issue):301–303, 2006.

[8] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bath, H. Weissig, I. N. Shindyalov, e P. E. Bourne. The protein data bank. *Nucliec Acids Research*, 28(1):235–242, 2000.

[9] E. Blanc, V. Fremont, P. Sizun, S. Meunier, J. Van Rietschoten, A. Thevand, J.M. Bernassau, e H. Darbon. Solution structure of P01, a natural scorpion peptide structurally analogous to scorpion toxins specific for apamin-sensitive potassium channel. *Proteins*, 24(3):359–369, 1996.

[10] C. Branden e J. Tooze. *Introduction to protein structure* . Garlang Publishing Inc., New York, EUA, 2° edição, 1998.

[11] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, e M. Karplus. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry*, 4(2):187–217, 1983.

[12] J. M. Bujnicki. Protein structure prediction by recombination of fragments. *Chembiochem: a European Journal of Chemical Biology*, 7(1):19–27, 2006.

[13] D. A. Case, T. E. Cheatham, T. Darden, H. Gohlke, R. Luo, K. M. Merz, e A. Onufriev. The *AMBER* biomolecular simulation program. *Journal of Computational Chemistry*, 26(16):1668–1688, 2005.

[14] B. Chapman e J. Chang. Biopython: python tools for computational biology. *ACM SIGBIO Newsletter*, 20(2):15–19, 2000.

[15] J. Cheng, A. Randall, e P. Sweredoski, M. Baldi. SCRATCH: a protein structure and structural feature prediction server. *Nucleic Acids Research*, 33:72–76, 2005.

[16] C. Combet, C. Blanchet, C. Geourjoun, e G. Deleage. NPS@: Network protein sequence analysis. *Trends in Biochemical Sciences*, 25(3):147–150, 2000.

[17] T. E. Creighton. Protein folding. *Biochemical Journal*, 270:1–16, 1990.

[18] V. Cutello, G. Narzisi, e G. Nicosia. A multi-objective evolutionary approach to the protein structure prediction problem. *Journal of the Royal Society Interface*, 3(6):139–151, 2006.

[19] R. Das, B. Qian, S. Raman, R. Vernon, J. Thompson, P. Bradley, S. Khare, M. D. D. Tyka, D. Bhat, D. Chivian, D. E. E. Kim, W. H. H. Sheffler, L. Malmström, A. M. M. Wollacott, C. Wang, I. Andre, e D. Baker. Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home. *Proteins*, 69(S8):118–128, 2007.

[20] W. L. Delano. The PyMOL molecular graphics system. *Delano Scientific, San Carlos, CA, USA*, 2002.

[21] G. Deleage e B. Roux. An algorithm for protein secondary structure prediction based on class prediction. *Protein Engineering*, 1(4):289–294, 1987.

[22] R. L. Dunbrack Jr. e M. Karplus. Backbone-dependent rotamer library for proteins: application to side-chain prediction. *Journal of Molecular Biology*, 230(2):543–574, 1993.

[23] A. Fiser, R. K. Do, e A. Sali. Modeling of loops in protein structures. *Protein Science*, 9(9):1753–1773, 2000.

[24] D. Frishman e P. Frishman. Incorporation of non-local interactions in protein secondary structure prediction from the amino acid sequence. *Protein Engineering*, 9(2):133–142, 1996.

[25] J. Garnier, J-F. Gibrat, e B. Robson. GOR secondary structure prediction method version IV. *Methods in Enzymology*, 266:540–553, 1996.

[26] J. Garnier, D. J. Osguthorpe, e B. Robson. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *Journal of Molecular Biology*, 120(1):97–120, 1978.

[27] C. Geourjon e G. Deleage. SOPM: a self-optimized method for protein secondary structure prediction. *Protein Engineering*, 7(2):157–164, 1994.

[28] C. Geourjon e G. Deleage. SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. *Computers Applied Biosciences*, 11(6):681–684, 1995.

[29] G. Gibas e P. Jambeck. *Desenvolvendo bioinformática*. Editora Campus- O'Reilly, Rio de Janeiro, BR, 1° edição, 2001.

[30] J. F. Gibrat, J. Garnier, e B. Robson. Further developments of protein secondary structure prediction using information theory. New parameters and consideration of residue pairs. *Journal of Molecular Biology*, 198(3):425–443, 1987.

[31] W. J. Graybeal e U. W. Pooch. *Simulation: principles and methods*. Cambridge: Winthrop Publishers, Inc., Cambridge, UK, 1980.

[32] A. M. Gronenborn, D. R. Filpula, N. Z. Essig, A. Achari, M. Whitlow, P. T. Wingfield, e G. M. Clore. A novel, highly stable fold of the immunoglobulin binding domain of streptococcal protein G. *Science*, 253(5020):657–661, 1991.

[33] Y. Guermeur, C. Geourjon, P. Gallinari, e G. Deleage. Improved performance in protein secondary structure prediction by inhomogeneous score combination. *Bioinformatics*, 15(5):413–421, 1999.

[34] S. Henikoff e J. G. Henikoff. Performance evaluation of amino acid substitution matrices. *Proteins*, 17(1):49–61, 1993.

[35] D. Higgins, J. Thompson, T. Gibson, J.D. Thompson, e T.J. Higgins, D.G. an Gibson. CLUSTAL W: improving the sensitivity of progressivemultiple sequence alignment through sequence weighting,position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22:4673–4680, 1994.

[36] L. Holm, C. Ouzounis, C. Sander, G. Tuparev, e G. Vriend. A database of protein structure families with common folding motifs. *Protein Science*, 1(12):1691–1698, 1992.

[37] T. Z. Hovmöller e T. Ohlson. Conformation of amino acids in protein. *Acta Crystallogr. D. Biol. Crystallogr.*, 58(Pt 5):768–776, 2002.

[38] E. G. Hutchinson e J. M. Thornton. PROMOTIF–A program to identify and analyze structural motifs in proteins. *Protein Science*, 5(2):212–220, 1996.

[39] M. P. Jacobson, R. A. Friesner, Z. Xiang, e B. Honig. On the role of the crystal environment in determining protein side chain conformations. *Journal of Molecular Biology*, 320(3):597–608, 2002.

[40] E. T. Jaynes e G. L. Bretthorst. *Probability theory: the logic of science*. Cambridge University Press, Cambridge, UK, 1° edição, 2003.

[41] D. T. Jones. Predicting novel protein folds by using FRAGFOLD. *Proteins*, Suppl 5:127–132, 2001.

[42] D. T. Jones, W. R. Taylort, e J. M. Thornton. A new approach to protein fold recognition. *Nature*, 358(6381):86–89, 1992.

[43] W. Kabasch e C. Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, 1983.

[44] M. Karplus. The Levinthal paradox: yesterday and today. *Folding and Design*, 2(1):S69–S75, 1997.

[45] J. S. Kavanaugh, J. A. Weydert, P. H. Rogers, e A. Arnone. High-resolution crystal structures of human hemoglobin with mutations at tryptophan 37beta: structural basis for a high-affinity T-state. *Biochemistry*, 37(13):4358–4373, 1998.

[46] R. D. King e M. J. Sternberg. Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. *Protein Science*, 5(11):2298–2310, 1996.

[47] A. Kolinski. Protein modelling and structure prediction with a reduced representation. *Acta Biochimica Polonica*, 51:349–371, 2004.

[48] R. A. Laskowski, M. W. MacArthur, D. S. Moss, e J. M. Thornton. PROCHECK: a program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography*, 26(2):283–291, 1993.

[49] Albert L. Lehninger, D. L. Nelson, e M. M. Cox. *Princípios de bioquímica*. Editora Sarvier, São Paulo, BR, 3° edição, 2002.

[50] A. M. Lesk. *Introduction to protein architecture: the structural biology of proteins*. Oxford University Press, Cambridge, UK, 1° edição, 2000.

[51] A. M. Lesk. *Introdução à bioinformática*. Artmed, São Paulo, Brasil, 2° edição, 2007.

[52] J. M. Levin, B. Robson, e J. Garnier. An algorithm for secondary structure determination in proteins based on sequence similarity. *FEBS Letters*, 205(2):303–308, 1986.

[53] C. Levinthal. Are the pathways for protein folding? *Journal of Chemical Physics*, 65:44–45, 1997.

[54] H. Lodish, A. Berk, S. L. Zipursky, P. Matsudaira, D. Baltimore, e J. E. Darnell. *Biologia celular e molecular*. Artmed, Porto Alegre, BR, 5° edição, 2005.

[55] J. E. Mace e D. A. Agard. Kinetic and structural characterization of mutations of glycine 216 in alpha-lytic protease: a new target for engineering substrate specificity. *Journal of Molecular Biology*, 254(4):720–736, 1995.

[56] Jr. A. D. MacKerell, D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, e etal. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *Journal of Physical Chemistry*, 102(18):3586–3616, 1998.

[57] J. MacQueen. Some methods for classification and analysis of multivariate observations. In L. M. LeCam e J. Neyman, editores, *Proceedings Fifth Berkeley Symposium on Mathematics Statistics and Probability*, paginas 281–297. University of California Press, 1967.

[58] M. A. Martim-Remon, A. Stuart, A. Fiser, R. Sánchez, F. Mello, e A. Sali. Comparative protein structure modelling of genes and genomes. *Annual Review of Biophysics and Biomolecular Structure*, 29(16):291–235, 2000.

[59] A. D. McLachlan. Rapid comparison of protein structures. *Acta Crystallographic*, A38:871–873, 1982.

[60] J. Moreira e etal. Designing a highly-scalable operating system: the Blue Gene/L story. pagina 118, New York, NY, USA, 2006. ACM.

[61] I. Morize, E. Surcouf, M. C. Vaney, Y. Epelboin, M. Buehner, F. Fridlansky, E. Milgrom, e J. P. Mornon. Refinement of the C222(1) crystal form of oxidized uteroglobin at 1.34 Å resolution. *Journal of Molecular Biology*, 194(4):725–739, 1987.

[62] A. L. Morris, M. W. MacArthur, E. G. Hutchinson, e J. M. Thornton. Stereochemical quality of protein structure coordinates. *Proteins*, 12(4):345–364, 1992.

[63] J. A. Mout. Decade of CASP: progress, bottlenecks an prognosis in protein structure prediction. *Current Opinion in Structural Biology*, 15:285–289, 2005.

[64] A. G. Murzin, S. E. Brenner, T. Hubbard, e C. Chothia. SCOP: a strutural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247(1):536–540, 1995.

[65] J. T. Ngo, J. Marks, e M. Karplus. Computational complexity, protein structure prediction and the Levinthal Paradox. In K. Merz Jr. e S. Le Grand, editores, *The Protein Folding Problem and Tertiary Structure Prediction*, paginas 435–508, Capítulo 14, Birkhäuser, Boston, MA, EUA, 1994.

[66] C. Notredame e Heringa J. Higgins, D. T-Coffee: A novel method for multiple sequence alignments. *Journal of Molecular Biology*, 302:205–217, 2000.

[67] C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, e J. M. Thornton. CATH- A hierarchic classification of protein domain structures. *Structure*, 5(8):1093–1108, 1997.

[68] D. J. Osguthorpe. Ab *initio* protein folding. *Current Opinion in Structural Biology*, 10(2):146–152, 2000.

[69] M. T. Pastor, M. Lopez de la Paz, E. Lacroix, L. Serrano, e E. Perez-Paya. Combinatorial approaches: a new tool to search for highly structured beta-hairpin peptides. *Proceedings of the National Academy of Sciences of the United States of America*, 99(2):614–619, 2002.

[70] L. Pauling e R. B. Corey. The pleated sheet, a new layer configuration of polypeptide chains. *Proceedings of the National Academy of Sciences of the United States of America*, 37(5):251–256, 1951.

[71] L. Pauling, R. B. Corey, e H. R. Branson. The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. *Proceedings of the National Academy of Sciences of the United States of America*, 37(4):205–211, 1951.

[72] W.R. Pearson e D.J. Lipman. Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences*, 85:2444–2448, 1988.

[73] J. W. Ponder. TINKER - Software tools for molecular design, version 4.2. http://dasher.wustl.edu/tinker/, Junho 2007.

[74] G. N. Ramachandran e V. Sasisekharan. Conformation of polypeptides and proteins. *Advances in Protein Chemistry*, 23:238–437, 1968.

[75] C. A. Rohl, C. E. Strauss, K. M. Misura, e D. Baker. Protein structure prediction using rosetta. *Methods in Enzymology*, 383:66–93, 2004.

[76] C. A. Rohl, C. E. Strauss, K. M. S. Misura, e D. Baker. Protein structure prediction using Rosetta. methods in enzymology. *Methods in Enzymology*, 383:66–93, 2004.

[77] F. H. F. P. da Rosa e V. A. P. Junior. Gerando números aleatórios. http://www.feferraz.net/files/lista/random-numbers.pdf, Junho 2002.

[78] B. Rost e C. Sander. Prediction of protein secondary structure at better than 70 percent accuracy. *Journal of Molecular Biology*, 232(2):584–599, 1993.

[79] R. B. Russell e G. J. Barton. Structural features can be unconserved in proteins with similar folds. An analysis of side-chain to side-chain contacts secondary structure and accessibility. *Journal of Molecular Biology*, 244(3):332–350, 1994.

[80] A. Sali, E. Shakhnovich, e M. Karplus. How does a protein fold? *Nature*, 369:248–251, 1994.

[81] F. X. Schmid. Kinetics of unfolding and refolding of single-domain protein in protein folding. In T.E. Creighton, editor, *Protein Folding*, paginas 197–241, New York, USA, 1992.

[82] W. R. P Scott, P. H. Huenenberger, I. G. Tironi, A. E. Mark, S. R. Billeter, J. Fennen, A. E. Torda, T. Huber, P. Krueger, e V. Gunsteren. The GROMOS biomolecular simulation program. *Journal of Physical Chemistry*, 103(19):3596–3607, 1999.

[83] R. A. Serway e J. W. Jewett. *Princípios de Física: Mecânica Clássica*. Thomson, São Paulo, 2004.

[84] K. T. Simons, R. Bonneau, I. Ruczinski, e D. Baker. Ab initio protein structure prediction of CASP III targets using Rosetta. *Proteins*, Suppl 3(6):171–176, 1999.

[85] R. Srinivasan, P. J. Fleming, e G. D. Rose. Ab initio protein folding using LINUS. *Methods in Enzymology*, 383:48–66, 2004.

[86] R. Srinivasan e G. D. Rose. LINUS - A hierarchic procedure to predict the fold of a protein. *Proteins*, 22:81–99, 1995.

[87] R. Srinivasan e G. D. Rose. Ab initio prediction of protein structure using LINUS. *Proteins*, 47:489–495, 2002.

[88] M. A. Starovasnick, A. C. Brasisted, e J. A. Wells. Structural mimicry of a native protein by a minimized binding domain. *Proceedings of the National Academy of Sciences of the United States of America*, 94:10080–10085, 1997.

[89] M. Sternberg. *Protein structure prediction: a practical approach*. Oxford University Press, Inc., New York, NY, USA, 1997.

[90] M. M. Teeter. Water structure of a hydrophobic protein at atomic resolution: Pentagon rings of water molecules in crystals of crambin. *Proceedings of the National Academy of Sciences of the United States of America*, 81(19):6014–6018, 1984.

[91] A. Tramontano e A. M. Lesk. *Protein structure prediction*. John Wiley and Sons, Inc., Weinheim, Germany, 1° edição, 2006.

[92] R. Unger e J. Moult. Finding the lowest free energy conformation of a protein is an np-hard problem: prof and applications. *Bulletin of Mathematical Biology*, 55:1183–1198, 1993.

[93] S. Velankar, P. McNeil, V. Mittard-Runte, A. Suarez, D. Barrell, R. Apweiler, e K. Henrick. E-MSD: an integrated data resource for bioinformatics. *Nucleic Acids Research*, 33(Database Issue), 2005.

[94] I. H. Witten e E. Frank. *Data mining: practical machine learning tools and techniques*. Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, Oxford, UK, 2° edição, 2005.

[95] Z. Xiang e B. Honig. Extending the accuracy limits of prediction for side chain conformations. *Journal of Molecular Biology*, 311(2):421–430, 2001.

[96] S. Zhirong e B. Jiang. Patters and conformations of commonly occurring superecondary structures (basic motifs) in protein data bank. *Journal of Protein Chemistry*, 15(7):675–690, 1996.

APÊNDICE A – Dupletos moldes da proteína 1K43

Tabela 32: Classificação dos dupletos-molde de cada fragmento alvo da proteína de código PDB igual a 1K43 nos três estados conformacionais (h, b ou c).

Fragmento	Resíduo central	N° de moldes	Hélice α(%)	Folha β(%)	Volta(%)
RGKWT	K	5	40.00	60.00	0.00
GKWTY	W	10	20.00	80.00	0.00
KWTYN	T	16	87.50	12.50	0.00
WTYNG	Y	3	33.33	66.67	0.00
TYNGI	N	9	11.11	55.56	33.33
YNGIT	G	27	33.33	11.11	55.56
NGITY	I	57	5.26	94.74	0.00
GITYE	T	28	21.43	78.57	0.00
ITYEG	Y	21	19.05	80.95	0.00
TYEGR	E	3	33.33	66.67	0.00

APÊNDICE B – Dupletos moldes da proteína 1ROP

Tabela 33: Classificação dos dupletos-molde de cada fragmento alvo da proteína de código PDB igual a 1ROP nos três estados conformacionais (h, b ou c).

Fragmento	Resíduo central	N° de moldes	Hélice α(%)	Folha β(%)	Volta(%)
MTKQE	K	58	50.00	44.83	5.17
TKQEK	Q	37	94.59	5.41	0.00
KQEKT	E	31	58.06	35.48	6.45
QEKTA	K	33	93.94	3.03	3.03
EKTAL	T	54	77.78	18.52	3.70
KTALN	A	51	92.16	3.92	3.92
TALNM	L	43	69.77	25.58	4.65
ALNMA	N	55	80.00	20.00	0.00
LNMAR	M	51	70.59	29.41	0.00
NMARF	A	44	54.55	45.45	0.00
MARFI	R	31	96.77	3.23	0.00
ARFIR	F	38	81.58	15.79	2.63
RFIRS	I	20	100.00	0.00	0.00
FIRSQ	R	33	93.94	6.06	0.00
IRSQT	S	40	65.00	32.50	2.50
RSQTL	Q	46	91.30	6.52	2.17
SQTLT	T	29	89.66	10.34	0.00
QTLTL	L	61	37.70	57.38	4.92
TLTLL	T	29	75.86	24.14	0.00
LTLLE	L	54	98.15	1.85	0.00
TLLEK	L	62	75.81	22.58	1.61
LLEKL	E	56	91.07	7.14	1.79
LEKLN	K	48	95.83	4.17	0.00
EKLNE	L	55	83.64	10.91	5.45
KLNEL	N	43	93.02	6.98	0.00
LNELD	E	57	89.47	10.53	0.00
NELDA	L	44	90.91	9.09	0.00
ELDAD	D	57	29.82	49.12	21.05
LDADE	A	29	34.48	65.52	0.00
DADEQ	D	26	50.00	50.00	0.00
ADEQA	E	54	100.00	0.00	0.00
DEQAD	Q	29	96.55	3.45	0.00
EQADI	A	57	45.61	54.39	0.00
QADIC	D	65	92.31	6.15	1.54
ADICE	I	35	100.00	0.00	0.00
DICES	C	28	78.57	21.43	0.00
ICESL	E	48	95.83	4.17	0.00
CESLH	S	17	100.00	0.00	0.00
ESLHD	L	38	100.00	0.00	0.00

continua na próxima página

Fragmento	Resíduo central	N° de moldes	Hélice α(%)	Folha β(%)	Volta(%)
SLHDH	H	20	100.00	0.00	0.00
LHDHA	D	59	88.14	10.17	1.69
HDHAD	H	46	95.65	4.35	0.00
DHADE	A	46	69.57	10.87	19.57
HADEL	D	21	95.24	4.76	0.00
ADELY	E	32	100.00	0.00	0.00
DELYR	L	40	100.00	0.00	0.00
ELYRS	Y	17	94.12	5.88	0.00
LYRSC	R	45	57.78	42.22	0.00
YRSCL	S	46	82.61	13.04	4.35
RSCLA	C	43	74.42	25.58	0.00
SCLAR	L	23	69.57	30.43	0.00
CLARF	A	25	84.00	16.00	0.00

APÊNDICE C – Dupletos moldes da proteína 1GB1

Tabela 34: Classificação dos dupletos-molde de cada fragmento alvo da proteína de código PDB igual a 1GB1 nos três estados conformacionais (h, b ou c).

Fragmento	Resíduo central	N° de moldes	Hélice α(%)	Folha β(%)	Volta(%)
MTYKL	Y	16	50.00	31.25	18.75
TYKLI	K	37	0.00	100.00	0.00
YKLIL	L	18	22.22	77.78	0.00
KLILN	I	12	33.33	66.67	0.00
LILNG	L	19	63.16	36.84	0.00
ILNGK	N	34	50.00	23.53	26.47
LNGKT	G	10	0.00	30.00	70.00
NGKTL	K	35	65.71	31.43	2.86
GKTLK	T	15	6.67	93.33	0.00
KTLKG	L	15	40.00	60.00	0.00
TLKGE	K	15	73.33	26.67	0.00
LKGET	G	12	8.33	33.33	58.33
KGETT	E	8	50.00	50.00	0.00
GETTT	T	19	15.79	84.21	0.00
ETTTE	T	21	33.33	61.9	4.76
TTTEA	T	3	100.00	0.00	0.00
TTEAV	E	13	76.92	23.08	0.00
TEAVD	A	18	83.33	11.11	5.56
EAVDA	V	42	90.48	9.52	0.00
AVDAA	D	18	66.67	33.33	0.00
VDAAT	A	33	75.76	24.24	0.00
DAATA	A	15	93.33	6.67	0.00
AATAE	T	14	78.57	21.43	0.00
ATAEK	A	14	57.14	35.71	7.14
TAEKV	E	21	80.95	9.52	9.52
AEKVF	K	19	100.00	0.00	0.00
EKVFK	V	16	56.25	43.75	0.00
KVFKQ	F	20	70.00	30.00	0.00
VFKQY	K	15	53.33	46.67	0.00
FKQYA	Q	10	60.00	40.00	0.00
KQYAN	Y	58	100.00	0.0)	0.00
QYAND	A	52	90.38	1.92	7.69
YANDN	N	15	80.00	20.00	0.00
ANDNG	D	5	60.00	20.00	20.00
NDNGV	N	18	55.56	0.00	44.44
DNGVD	G	29	3.45	24.14	72.41
NGVDG	V	13	15.38	76.92	7.69
GVDGE	D	7	14.29	71.43	14.29
VDGEW	G	21	14.29	4.76	80.95

continua na próxima página

Fragmento	Resíduo central	N° de moldes	Hélice α(%)	Folha β(%)	Volta(%)
DGEWT	E	16	6.25	93.75	0.00
GEWTY	W	49	10.20	83.67	6.12
EWTYD	T	17	82.35	17.65	0.00
WTYDD	Y	12	66.67	33.33	0.00
TYDDA	D	11	27.27	63.64	9.09
YDDAT	D	35	31.43	60.00	8.57
DDATK	A	15	80.00	20.00	0.00
DATKT	T	11	100.00	0.00	0.00
ATKTF	K	16	62.50	25.00	12.50
TKTFT	T	5	40.00	60.00	0.00
KTFTV	F	17	47.06	52.94	0.00
TFTVT	T	15	0.00	100.00	0.00
FTVTE	V	18	33.33	66.67	0.00

APÊNDICE D – Dupletos moldes da proteína 1GAB

Tabela 35: Classificação dos dupletos-molde de cada fragmento alvo da proteína de código PDB igual a 1GAB nos três estados conformacionais (h, b ou c).

Fragmento	Resíduo central	N° de moldes	Hélice α(%)	Folha β(%)	Volta(%)
TIDQW	D	61	75.41	24.59	0.00
IDQWL	Q	21	95.24	4.76	0.00
DQWLL	W	31	51.61	45.16	3.23
QWLLK	L	11	90.91	0.00	9.09
WLLKN	L	53	11.32	84.91	3.77
LLKNA	K	33	87.88	6.06	6.06
LKNAK	N	41	63.41	21.95	14.63
KNAKE	A	26	53.85	38.46	7.69
NAKED	K	16	43.75	43.75	12.50
AKEDA	E	20	65.00	15.00	20.00
KEDAI	D	58	89.66	5.17	5.17
EDAIA	A	51	98.04	1.96	0.00
DAIAE	I	53	92.45	7.55	0.00
AIAEL	A	45	97.78	2.22	0.00
IAELK	E	51	100.00	0.00	0.00
AELKK	L	36	91.67	8.33	0.00
ELKKA	K	50	98.00	2.00	0.00
LKKAG	K	63	92.06	7.94	0.00
KKAGI	A	54	62.96	25.93	11.11
KAGIT	G	17	23.53	5.88	70.59
AGITS	I	41	82.93	14.63	2.44
GITSD	T	40	25.00	72.5	2.50
ITSDF	S	70	51.43	31.43	17.14
TSDFY	D	13	38.46	61.54	0.00
SDFYF	F	6	16.67	83.33	0.00
DFYFN	Y	34	26.47	73.53	0.00
FYFNA	F	11	9.09	72.73	18.18
YFNAI	N	16	93.75	6.25	0.00
FNAIN	A	31	64.52	3.23	32.26
NAINK	I	37	100.00	0.00	0.00
AINKA	N	33	90.91	9.09	0.00
INKAK	K	25	72.00	28.00	0.00
NKAKT	A	25	76.00	12.00	12.00
KAKTV	K	31	67.74	29.03	3.23
AKTVE	T	17	47.06	52.94	0.00
KTVEE	V	43	95.35	2.33	2.33
TVEEV	E	50	92.00	8.00	0.00
VEEVN	E	42	45.24	52.38	2.38
EEVNA	V	36	44.44	55.56	0.00

continua na próxima página

Fragmento	Resíduo central	N° de moldes	Hélice α(%)	Folha β(%)	Volta(%)
EVNAL	N	47	78.72	21.28	0.00
VNALK	A	50	100.00	0.00	0.00
NALKN	L	62	100.00	0.00	0.00
ALKNE	K	23	60.87	34.78	4.35
LKNEI	N	36	61.11	33.33	5.56
KNEIL	E	11	54.55	45.45	0.00
NEILK	I	41	97.56	2.44	0.00
EILKA	L	46	100.00	0.00	0.00
ILKAH	K	35	94.29	0.00	5.71
LKAHA	A	13	76.92	23.08	0.00

APÊNDICE E – Dupletos moldes da proteína 1UTG

Tabela 36: Classificação dos dupletos-molde de cada fragmento alvo da proteína de código PDB igual a 1UTG nos três estados conformacionais (h, b ou c).

Fragmento	Resíduo central	N° de moldes	Hélice α(%)	Folha β(%)	Volta(%)
GICPR	C	49	44.90	53.06	2.04
ICPRF	P	34	23.53	73.53	2.94
CPRFA	R	23	13.04	86.96	0.00
PRFAH	F	19	57.89	31.58	10.53
RFAHV	A	38	52.63	47.37	0.00
FAHVI	H	55	36.36	63.64	0.00
AHVIE	V	67	88.06	11.94	0.00
HVIEN	I	44	63.64	34.09	2.27
VIENL	E	58	63.79	32.76	3.45
IENLL	N	28	96.43	3.57	0.00
ENLLL	L	62	24.19	75.81	0.00
NLLLG	L	22	36.36	63.64	0.00
LLLGT	L	27	48.15	51.85	0.00
LLGTP	G	27	14.81	29.63	55.56
LGTPS	T	10	0.00	100.00	0.00
GTPSS	P	8	25.00	75.00	0.00
TPSSY	S	26	38.46	38.46	23.08
PSSYE	S	22	68.18	13.64	18.18
SSYET	Y	17	94.12	5.88	0.00
SYETS	E	14	35.71	57.14	7.14
YETSL	T	59	91.53	6.78	1.69
ETSLK	S	37	91.89	8.11	0.00
TSLKE	L	24	87.50	12.50	0.00
SLKEF	K	25	92.00	8.00	0.00
LKEFE	E	49	69.39	30.61	0.00
KEFEP	F	67	17.91	70.15	11.94
EFEPD	E	39	0.00	69.23	30.77
FEPDD	P	65	72.31	26.15	1.54
EPDDT	D	11	81.82	18.18	0.00
PDDTM	D	43	83.72	0.00	16.28
DDTMK	T	72	98.61	0.00	1.39
DTMKD	M	46	97.83	2.17	0.00
TMKDA	K	16	81.25	12.50	6.25
MKDAG	D	43	90.70	6.98	2.33
KDAGM	A	35	94.29	2.86	2.86
DAGMQ	G	39	71.79	5.13	23.08
AGMQM	M	66	59.09	39.39	1.52
GMQMK	Q	23	52.17	47.83	0.00
MQMKK	M	27	66.67	33.33	0.00

continua na próxima página

Fragmento	Resíduo central	N° de moldes	Hélice α(%)	Folha β(%)	Volta(%)
QMKKV	K	28	64.29	32.14	3.57
MKKVL	K	72	73.61	25.00	1.39
KKVLD	V	86	97.67	2.33	0.00
KVLDS	L	34	26.47	73.53	0.00
VLDSL	D	53	75.47	24.53	0.00
LDSLP	S	47	78.72	14.89	6.38
DSLPQ	L	33	6.06	90.91	3.03
SLPQT	P	2	0.00	0.00	100.00
LPQTT	Q	47	68.09	27.66	4.26
PQTTR	T	28	39.29	57.14	3.57
QTTRE	T	48	83.33	16.67	0.00
TTREN	R	7	71.43	28.57	0.00
TRENI	E	42	90.48	7.14	2.38
RENIM	N	36	55.56	30.56	13.89
ENIMK	I	17	100.00	0.00	0.00
NIMKL	M	32	62.50	34.38	3.13
IMKLT	K	28	64.29	35.71	0.00
MKLTE	L	60	86.67	13.33	0.00
KLTEK	T	39	35.90	64.10	0.00
LTEKI	E	45	55.56	44.44	0.00
TEKIV	K	13	76.92	23.08	0.00
EKIVK	I	40	92.50	7.50	0.00
KIVKS	V	5	80.00	20.00	0.00
IVKSP	K	8	25.00	75.00	0.00
VKSPL	S	25	40.00	52.00	8.00
KSPLC	P	10	100.00	0.00	0.00
SPLCM	L	17	58.82	35.29	5.88

APÊNDICE F – Agrupamento das tuplas molde da proteína 1K43

Tabela 37: Agrupamento das tuplas-molde associadas a um fragmento alvo s_i da proteína cujo código PDB é 1K43: (m) é o valor médio e (σ) é o desvio padrão estimado de cada grupo.

		Grupo 01		Grupo 02		Grupo 03		Grupo 04	
Frag		phi	psi	phi	psi	phi	psi	phi	psi
RGKWT	m	-53.39	-34.94	-91.33	160.66	-95.96	161.01	-134.12	137.54
RGKWT	σ	2.04	3.23	11.30	0.85	10.35	0.78	34.48	103.37
GKWTY	m	-132.12	164.98	-74.63	122.75	-63.77	-40.29	-95.33	120.02
GKWTY	σ	8.76	6.61	0.50	1.70	1.56	1.54	7.50	16.77
KWTYN	m	-107.17	110.41	-96.36	19.23	-105.17	-53.54	-60.26	-40.11
KWTYN	σ	12.58	4.83	6.76	12.52	3.88	4.01	10.08	10.29
WTYNG	m	-92.35	119.68	-64.47	-29.23	-111.86	133.26	-	-
WTYNG	σ	9.90	6.89	28.91	92.01	15.27	10.63	-	-
TYNGI	m	61.50	-142.74	-115.96	141.81	56.16	33.94	-64.24	-34.55
TYNGI	σ	88.06	102.97	31.99	8.24	4.26	2.71	88.06	102.97
YNGIT	m	78.81	1.97	-74.24	-26.32	116.38	-177.25	-125.58	143.09
YNGIT	σ	7.37	14.27	14.86	5.66	3.23	0.61	39.47	18.14
NGITY	m	-107.10	131.68	-74.36	129.93	-133.06	123.29	-68.74	-40.72
NGITY	σ	10.13	5.58	7.81	7.46	7.72	22.50	15.70	6.83
GITYE	m	-117.31	134.29	-86.24	131.89	-123.86	25.02	-53.25	-50.02
GITYE	σ	7.14	21.68	9.47	28.72	9.55	22.08	20.96	58.63
ITYEG	m	-109.34	96.84	-91.27	145.98	-124.70	145.31	-54.21	-26.89
ITYEG	σ	10.55	7.68	0.89	10.82	14.52	16.35	5.03	5.34
TYEGR	m	-124.85	153.28	-56.52	122.02	-87.38	1.40	-	-
TYEGR	σ	34.21	80.20	34.21	80.20	34.21	80.20	-	-

APÊNDICE G – Agrupamento das tuplas molde da proteína 1ROP

Tabela 38: Agrupamento das tuplas-molde associadas a um fragmento alvo s_i da proteína cujo código PDB é 1ROP: (m) é o valor médio e (σ) é o desvio padrão estimado de cada grupo.

		Grupo 01		Grupo 02		Grupo 03		Grupo 04	
Frag		phi	psi	phi	psi	phi	psi	phi	psi
MTKQE	m	-140.42	152.81	-58.10	-40.82	-119.05	-104.10	47.91	39.92
MTKQE	σ	20.65	9.54	5.87	4.37	29.97	61.83	27.88	85.46
TKQEK	m	-61.93	-51.06	-83.18	-8.69	-103.38	95.05	-64.67	-39.81
TKQEK	σ	6.54	14.76	3.04	17.08	25.13	3.37	3.17	1.74
KQEKT	m	-66.05	136.53	-162.63	157.77	93.35	-9.31	-66.28	-55.26
KQEKT	σ	38.88	15.73	1.85	1.24	53.18	101.15	12.68	35.71
QEKTA	m	-63.92	-42.12	-96.89	151.15	52.24	28.61	-81.63	-18.95
QEKTA	σ	5.09	8.47	24.55	36.27	24.55	36.27	9.31	7.91
EKTAL	m	-57.53	-49.93	63.79	86.22	-112.05	154.66	-97.51	-6.58
EKTAL	σ	6.76	15.37	33.20	81.10	17.06	11.68	32.67	19.88
KTALN	m	-61.21	-42.91	-54.02	-52.77	-133.59	159.38	74.19	22.90
KTALN	σ	9.18	4.69	2.15	4.71	0.57	0.13	16.30	11.30
TALNM	m	-108.90	140.04	-64.73	-42.27	53.23	-126.34	-51.29	-52.59
TALNM	σ	9.29	7.19	4.64	5.20	1.01	0.30	4.92	1.79
ALNMA	m	-72.68	-29.91	-52.23	-44.85	-99.87	136.86	-63.99	-42.51
ALNMA	σ	4.18	6.75	2.19	9.31	16.65	5.26	3.54	5.20
LNMAR	m	-87.00	150.01	-92.79	-6.31	-65.85	-41.05	-70.26	-166.21
LNMAR	σ	8.23	13.65	21.44	17.36	4.19	3.79	7.24	9.93
NMARF	m	-119.10	147.49	-60.90	-41.70	-59.75	142.64	-112.09	-26.16
NMARF	σ	17.48	5.09	7.25	5.05	5.41	7.63	8.85	6.02
MARFI	m	-53.58	-40.37	-71.92	-37.24	-63.04	-41.61	-120.63	137.51
MARFI	σ	4.52	2.44	3.08	3.73	2.01	3.72	11.98	32.21
ARFIR	m	-127.23	126.54	-61.45	-47.64	-81.94	-31.78	26.43	-51.18
ARFIR	σ	9.28	23.70	3.85	3.74	16.89	7.06	30.19	64.55
RFIRS	m	-62.76	-38.63	-59.43	-44.41	-66.30	-34.33	-76.09	-55.02
RFIRS	σ	1.21	1.52	0.36	0.28	1.41	2.11	1.74	3.03
FIRSQ	m	-61.58	-42.74	-115.90	73.64	-56.82	-46.76	-68.68	135.85
FIRSQ	σ	2.96	4.83	9.86	57.04	3.81	5.51	14.28	44.08
IRSQT	m	-122.32	116.00	-65.88	-43.07	-111.25	-142.54	-55.42	-19.03
IRSQT	σ	8.19	4.92	7.53	4.85	6.01	4.76	9.10	4.21
RSQTL	m	-62.04	-37.70	-105.26	9.07	-130.90	35.54	-105.25	145.11
RSQTL	σ	7.32	6.49	6.20	58.37	4.46	3.88	47.02	9.22
SQTLT	m	-113.14	-25.78	-66.88	-32.37	-114.86	130.29	-89.78	-28.36
SQTLT	σ	6.64	5.86	3.51	4.94	3.67	3.57	4.45	16.38
QTLTL	m	-147.32	136.54	-67.93	-38.47	74.18	-59.07	-105.73	129.59
QTLTL	σ	13.54	9.84	6.80	11.29	8.36	3.31	5.20	11.35
TLTLL	m	-100.73	133.77	-125.41	131.80	-57.54	-42.54	-64.22	-42.54
TLTLL	σ	3.54	9.36	7.94	10.51	4.26	13.59	3.74	2.56

continua na próxima página

		Grupo 01		Grupo 02		Grupo 03		Grupo 04	
Frag		phi	psi	phi	psi	phi	psi	phi	psi
LTLLE	*m*	-71.31	-37.61	-117.56	136.67	-52.80	-39.03	-63.38	-44.27
LTLLE	σ	2.65	7.18	9.01	25.20	5.87	10.85	2.64	4.44
TLLEK	*m*	-102.24	128.89	-60.08	-45.58	-59.78	136.81	-109.38	-23.67
TLLEK	σ	13.96	78.06	3.86	3.56	5.20	8.02	8.10	3.65
LLEKL	*m*	-55.47	147.31	-160.63	141.67	-60.73	-42.56	-95.11	-33.04
LLEKL	σ	4.27	1.46	5.91	2.95	6.05	7.52	15.90	10.80
LEKLN	*m*	-60.36	-43.36	-69.96	-36.09	-112.83	142.14	-61.98	17.36
LEKLN	σ	4.35	6.26	7.45	7.18	29.44	14.09	2.03	27.68
EKLNE	*m*	-100.49	-21.41	-69.07	132.10	64.30	-126.98	-64.53	-42.06
EKLNE	σ	29.10	58.60	10.36	26.15	3.30	43.12	7.95	6.85
KLNEL	*m*	-99.14	139.49	-66.27	-37.59	-116.70	-8.37	-58.37	-44.62
KLNEL	σ	40.09	11.47	3.97	13.32	5.91	8.47	4.67	6.00
LNELD	*m*	-73.46	-14.15	-63.75	-37.60	-101.34	40.17	-154.09	141.46
LNELD	σ	11.27	8.24	8.23	8.40	22.72	40.26	9.08	12.27
NELDA	*m*	-87.84	-1.56	-63.13	-38.40	-20.69	-49.14	-76.03	124.04
NELDA	σ	6.09	11.11	7.00	10.38	12.17	9.55	6.63	3.99
ELDAD	*m*	53.98	34.85	-109.89	-171.49	-77.66	-10.71	-88.33	134.6
ELDAD	σ	7.42	5.76	34.13	5.62	21.76	28.83	17.24	24.00
LDADE	*m*	-55.32	-38.71	-80.04	146.20	-140.81	146.73	-91.02	76.71
LDADE	σ	8.77	12.36	11.31	10.41	21.47	72.46	7.35	11.55
DADEQ	*m*	-71.69	144.95	-124.98	-176.52	-132.65	138.99	-65.20	-34.43
DADEQ	σ	12.18	30.14	32.36	104.66	14.23	47.41	15.12	23.37
ADEQA	*m*	-91.69	-28.79	-55.98	-52.96	-63.47	-45.04	-69.89	1.95
ADEQA	σ	6.41	24.70	2.51	10.25	2.49	6.41	3.32	2.49
DEQAD	*m*	-69.14	-42.30	-115.73	126.75	-58.66	-46.37	-82.00	-34.24
DEQAD	σ	2.17	2.21	12.49	32.63	2.78	10.15	3.24	5.73
EQADI	*m*	-69.28	154.94	-129.96	132.83	-82.49	12.79	-62.06	-33.84
EQADI	σ	9.44	5.15	12.20	10.91	4.71	2.70	3.88	15.81
QADIC	*m*	-17.84	51.55	-111.68	-37.80	-98.67	130.23	-56.69	-28.27
QADIC	σ	45.51	0.37	16.73	8.22	1.85	1.16	9.19	15.33
ADICE	*m*	-60.65	-16.54	-63.05	-43.87	-107.44	14.14	-110.03	-22.50
ADICE	σ	1.28	3.09	5.10	4.78	0.31	0.55	3.09	1.52
DICES	*m*	-63.86	-34.28	-80.11	-16.25	-173.40	163.04	-53.85	132.62
DICES	σ	4.26	5.66	2.63	3.09	22.71	72.05	4.22	1.39
ICESL	*m*	-143.42	-56.91	-157.68	149.22	-63.60	-42.29	-66.79	-30.94
ICESL	σ	10.38	5.72	1.29	10.38	3.60	3.95	15.81	8.55
CESLH	*m*	-58.91	-43.19	-66.10	-37.23	-49.26	-68.85	-62.76	-42.22
CESLH	σ	1.87	3.77	1.31	2.98	4.65	8.01	0.82	0.92
ESLHD	*m*	-75.51	-34.65	-105.48	0.38	-161.50	-50.06	-57.46	-47.6
ESLHD	σ	3.73	3.36	15.48	1.55	20.96	12.24	9.32	5.47
SLHDH	*m*	-166.95	-62.63	-93.43	6.70	-61.51	-50.96	-61.77	-41.76
SLHDH	σ	25.64	20.07	1.01	0.92	3.30	1.42	5.54	2.06
LHDHA	*m*	-95.29	-1.92	-57.28	-40.71	-145.81	142.35	77.52	-7.57
LHDHA	σ	8.64	7.25	7.55	5.38	5.54	3.30	35.27	51.51
HDHAD	*m*	-53.42	-52.72	-70.03	-23.92	-66.53	-38.13	-126.81	127.04

continua na próxima página

		Grupo 01		Grupo 02		Grupo 03		Grupo 04	
Frag		phi	psi	phi	psi	phi	psi	phi	psi
HDHAD	σ	5.87	3.35	5.45	1.58	6.00	2.86	0.82	1.59
DHADE	*m*	77.42	-0.88	-77.41	-20.64	-60.79	-40.28	-81.12	152.41
DHADE	σ	6.14	10.52	6.06	1.56	5.49	6.72	13.62	6.37
HADEL	*m*	-58.54	-45.47	-71.48	-38.53	-164.87	-169.32	-60.68	-14.18
HADEL	σ	6.70	3.84	5.34	3.75	23.55	29.18	1.38	3.91
ADELY	*m*	-129.06	27.09	-60.07	-44.59	-44.14	-54.03	-59.68	-18.44
ADELY	σ	2.54	4.26	4.71	5.07	3.42	19.51	2.30	2.24
DELYR	*m*	-61.65	-19.56	-88.74	-76.43	-67.53	-38.45	-58.98	-48.47
DELYR	σ	4.41	4.17	7.39	10.85	5.89	1.80	4.48	3.11
ELYRS	*m*	-143.67	170.51	-49.88	-45.10	-62.42	-45.19	-57.13	-48.90
ELYRS	σ	21.15	53.03	0.70	5.28	2.00	5.23	1.19	5.00
LYRSC	*m*	-133.69	144.78	-60.40	-37.93	-79.21	140.65	-85.84	-11.84
LYRSC	σ	6.79	4.83	6.50	5.02	35.53	89.60	7.60	8.23
YRSCL	*m*	152.22	-167.72	-63.95	-39.93	-100.26	-1.01	-93.37	128.96
YRSCL	σ	7.23	1.62	8.52	8.70	17.45	11.67	12.35	23.20
RSCLA	*m*	-119.59	121.21	-82.86	150.25	-71.53	-7.78	-65.87	-39.33
RSCLA	σ	3.52	16.86	7.82	7.26	8.23	4.24	6.86	7.62
SCLAR	*m*	-62.94	-42.49	-113.42	138.62	-92.09	123.14	-73.90	-17.91
SCLAR	σ	5.46	3.91	16.09	7.29	1.31	0.25	7.24	5.71
CLARF	*m*	-118.44	145.19	-107.33	-13.62	-69.11	-41.79	-57.12	-44.13
CLARF	σ	2.67	6.65	22.25	70.62	4.53	11.74	1.86	3.89

APÊNDICE H – Agrupamento das tuplas molde da proteína 1GB1

Tabela 39: Agrupamento das tuplas-molde associadas a um fragmento alvo s_i da proteína cujo código PDB é 1GB1: (m) é o valor médio e (σ) é o desvio padrão estimado de cada grupo.

		Grupo 01		Grupo 02		Grupo 03		Grupo 04	
Frag		phi	psi	phi	psi	phi	psi	phi	psi
TIDQW	m	-65.79	-38.49	-68.77	137.17	-89.53	-25.87	-130.68	123.41
TIDQW	σ	6.26	7.92	28.74	0.00	19.61	17.66	10.00	8.16
IDQWL	m	-63.10	-42.86	-44.24	-53.48	-63.13	-15.56	-144.97	113.53
IDQWL	σ	7.42	5.05	9.87	5.18	8.26	8.96	21.53	36.32
DQWLL	m	-96.43	134.88	-26.27	-91.80	-119.12	136.12	-67.20	-45.31
DQWLL	σ	12.23	0.71	24.85	93.74	2.68	9.34	10.57	13.70
QWLLK	m	36.65	-28.56	-62.31	-42.96	-67.29	-37.13	-32.11	-38.57
QWLLK	σ	31.10	4.88	1.71	1.83	2.18	1.70	7.34	0.76
WLLKN	m	76.75	-58.98	-71.43	-64.32	-133.51	137.20	-112.63	126.78
WLLKN	σ	6.81	63.79	19.58	45.42	8.72	23.98	5.70	4.25
LLKNA	m	-62.53	144.2	62.77	32.42	-54.17	-50.73	-66.32	-35.74
LLKNA	σ	1.05	3.48	1.03	3.22	8.62	8.22	4.18	5.42
LKNAK	m	-99.12	-2.29	-77.04	-40.79	53.88	45.63	-86.80	146.54
LKNAK	σ	11.27	8.81	22.25	9.69	12.07	24.32	34.39	12.63
KNAKE	m	-129.41	164.7	-64.45	-33.12	64.84	25.29	-85.43	127.66
KNAKE	σ	6.81	2.83	10.82	13.38	7.93	6.31	9.17	13.72
NAKED	m	-170.56	164.33	-60.13	-50.89	-63.69	141.25	-176.71	19.54
NAKED	σ	39.53	98.51	4.54	13.86	13.55	9.88	6.89	7.75
AKEDA	m	-67.74	-55.08	-129.46	-179.43	-59.54	159.11	-150.24	-2.89
AKEDA	σ	13.08	45.06	38.91	92.47	9.31	12.50	22.21	3.94
KEDAI	m	-62.53	133.45	-64.01	-43.31	-80.39	-17.98	66.64	2.94
KEDAI	σ	20.16	5.38	9.16	7.27	16.44	19.13	29.38	46.92
EDAIA	m	-98.99	-43.83	-158.00	150.10	-62.71	-42.68	-69.81	-41.36
EDAIA	σ	0.57	0.04	14.63	27.46	3.52	4.50	3.46	7.62
DAIAE	m	-135.27	122.43	-70.32	144.12	-76.03	-35.25	-62.76	-41.91
DAIAE	σ	3.62	16.10	2.30	7.19	4.61	16.25	6.51	6.00
AIAEL	m	-79.23	-1.18	-47.67	-43.32	-67.50	-35.14	-58.51	-41.64
AIAEL	σ	2.26	28.23	3.70	7.55	2.19	12.15	2.67	10.44
IAELK	m	-61.58	-46.52	-69.80	-38.92	-84.72	-9.56	-66.04	-55.14
IAELK	σ	4.71	3.18	5.91	4.28	11.23	5.74	3.01	0.21
AELKK	m	-65.61	-41.77	-123.59	117.93	-98.44	-38.95	-82.76	82.67
AELKK	σ	4.55	5.49	6.66	0.88	9.17	11.15	16.72	42.13
ELKKA	m	-89.24	-29.10	-71.98	120.31	-57.31	-41.31	-65.90	-37.48
ELKKA	σ	9.17	14.36	9.78	25.12	4.79	10.69	4.82	9.67
LKKAG	m	-96.27	-12.44	-63.24	122.75	-73.62	-24.38	-60.33	-36.07
LKKAG	σ	11.81	18.16	7.42	19.81	3.36	26.69	5.88	8.63
KKAGI	m	54.01	-97.58	-106.03	139.72	-58.30	125.96	-83.50	-14.27
KKAGI	σ	6.52	75.03	4.79	5.85	12.71	13.96	14.41	21.39

continua na próxima página

		Grupo 01		Grupo 02		Grupo 03		Grupo 04	
Frag		phi	psi	phi	psi	phi	psi	phi	psi
KAGIT	m	89.95	3.72	147.58	-160.64	-94.77	-165.04	-66.41	-37.51
KAGIT	σ	14.61	17.60	80.77	56.69	8.16	6.32	1.69	7.49
AGITS	m	-63.35	-39.96	-54.03	107.89	-79.93	124.22	-57.91	-45.88
AGITS	σ	5.21	13.05	8.83	63.90	4.00	8.40	3.95	7.76
GITSD	m	54.60	119.14	-105.09	170.31	-120.32	-173.99	-103.48	-14.15
GITSD	σ	31.94	139.28	16.07	4.69	6.28	3.73	26.49	24.31
ITSDF	m	-66.11	-40.95	-109.16	139.55	90.01	-13.14	-104.39	18.02
ITSDF	σ	10.18	20.56	23.72	18.99	29.54	54.88	17.19	12.14
TSDFY	m	-72.64	75.66	-100.20	153.13	-69.03	-27.37	-136.72	-57.91
TSDFY	σ	6.33	1.88	20.02	7.70	11.56	5.23	26.11	89.57
SDFYF	m	-89.22	134.70	-59.77	-46.53	-123.14	140.48	-148.44	126.08
SDFYF	σ	34.96	74.88	34.96	74.88	0.47	3.79	3.58	19.57
DFYFN	m	-87.68	-30.26	-79.15	136.04	-119.40	133.79	-147.68	153.36
DFYFN	σ	20.51	19.46	7.68	16.42	4.72	6.93	8.87	17.26
FYFNA	m	-58.04	-38.58	-127.73	138.80	-43.49	134.01	-104.27	157.81
FYFNA	σ	36.72	55.93	1.71	12.97	3.63	5.96	4.38	0.99
YFNAI	m	-123.87	178.82	-90.71	-19.01	-63.49	-36.83	-77.61	-27.64
YFNAI	σ	15.77	52.98	2.01	0.65	5.34	6.07	2.52	5.37
FNAIN	m	-121.30	52.32	-72.11	-25.64	64.11	-161.98	82.83	9.19
FNAIN	σ	2.23	69.27	9.96	13.44	0.55	3.96	8.59	12.63
NAINK	m	-74.34	-40.03	-59.98	-45.46	-60.76	-31.77	-103.57	15.75
NAINK	σ	7.54	7.21	2.65	3.94	5.87	4.59	9.28	11.71
AINKA	m	-67.29	-26.18	-35.33	-59.76	-73.98	146.37	-58.79	-44.46
AINKA	σ	6.92	11.86	5.42	4.61	7.15	16.02	7.38	7.32
INKAK	m	-151.29	118.10	-96.53	-11.21	-71.52	-16.93	-59.08	-33.11
INKAK	σ	2.96	3.91	6.93	10.96	0.17	4.16	10.27	7.06
NKAKT	m	-78.87	-28.20	-42.14	-106.07	-63.42	137.92	84.65	-170.22
NKAKT	σ	14.24	13.75	4.27	49.50	10.22	15.58	36.05	81.42
KAKTV	m	-105.92	130.57	-58.50	-49.46	92.00	-1.19	-101.09	-33.68
KAKTV	σ	23.24	20.39	8.19	10.40	41.18	82.26	7.84	19.98
AKTVE	m	-77.25	139.01	-69.07	-43.19	-151.03	152.22	-112.54	-166.29
AKTVE	σ	5.33	24.99	7.29	3.17	14.25	9.58	14.35	9.76
KTVEE	m	-55.50	-41.58	8.82	141.90	-68.63	146.98	-67.53	-34.94
KTVEE	σ	3.89	9.56	12.93	39.85	12.93	39.85	4.31	4.78
TVEEV	m	-91.83	110.67	-50.07	-49.99	-83.18	-35.01	-61.42	-40.44
TVEEV	σ	13.29	49.36	3.03	7.70	15.58	13.01	4.14	4.24
VEEVN	m	-103.68	128.93	167.59	-2.27	-76.33	-42.23	-154.27	130.83
VEEVN	σ	10.53	16.10	58.07	90.14	16.32	34.48	5.22	5.19
EEVNA	m	-111.16	135.98	-62.70	-47.41	-83.07	98.84	-53.07	-57.86
EEVNA	σ	4.47	5.37	3.35	5.33	3.97	8.65	5.86	6.80
EVNAL	m	-107.86	18.78	-123.47	110.73	-75.23	-33.28	-69.99	131.44
EVNAL	σ	8.23	17.10	8.95	39.55	21.04	9.05	13.48	9.36
VNALK	m	-70.74	-37.78	-91.82	-54.61	-64.36	-41.75	-55.81	-45.68
VNALK	σ	5.55	15.42	12.84	4.18	3.17	2.67	3.93	5.39
NALKN	m	-57.97	-51.75	-90.26	-25.04	-73.31	-9.19	-66.51	-44.56

continua na próxima página

		Grupo 01		Grupo 02		Grupo 03		Grupo 04	
Frag		phi	psi	phi	psi	phi	psi	phi	psi
NALKN	σ	1.92	2.29	4.63	11.08	3.41	5.41	7.41	10.79
ALKNE	*m*	-87.04	108.19	58.37	-126.30	-62.43	-38.43	-149.35	140.54
ALKNE	σ	4.41	5.25	38.20	81.71	4.40	11.67	7.85	27.64
LKNEI	*m*	80.92	-168.02	-78.32	115.64	57.15	47.22	-70.88	-36.99
LKNEI	σ	35.89	82.85	6.98	30.89	35.89	82.85	15.59	31.78
KNEIL	*m*	-55.65	-40.77	-69.76	-37.25	-91.00	138.66	-111.76	147.04
KNEIL	σ	0.25	0.34	5.42	2.23	2.12	5.52	17.93	93.47
NEILK	*m*	-55.20	-48.61	-81.65	126.63	-64.57	-41.30	-110.46	-29.20
NEILK	σ	4.59	9.47	18.86	31.07	4.76	11.80	4.26	29.63
EILKA	*m*	-58.61	-32.82	-58.30	-48.38	-65.15	-42.11	-66.17	-26.75
EILKA	σ	20.21	0.19	4.54	6.30	5.31	3.82	6.19	1.84
ILKAH	*m*	-116.25	-41.94	47.50	43.02	-57.15	-43.14	-70.84	-28.53
ILKAH	σ	6.62	20.30	8.81	8.89	7.17	8.29	2.48	7.08
LKAHA	*m*	-62.84	-40.81	-119.06	166.07	-93.52	32.48	-103.90	106.05
LKAHA	σ	3.53	4.57	21.47	73.40	1.88	3.90	6.33	1.03

APÊNDICE I – Agrupamento das tuplas molde da proteína 1GAB

Tabela 40: Agrupamento das tuplas-molde associadas a um fragmento alvo s_i da proteína cujo código PDB é 1GAB: (m) é o valor médio e (σ) é o desvio padrão estimado de cada grupo.

		Grupo 01		Grupo 02		Grupo 03		Grupo 04	
Frag		phi	psi	phi	psi	phi	psi	phi	psi
TIDQW	m	-65.79	-38.49	-68.77	137.17	-89.53	-25.87	-130.68	123.41
TIDQW	σ	6.26	7.92	28.74	0.00	19.61	17.66	10.00	8.16
IDQWL	m	-63.10	-42.86	-44.24	-53.48	-63.13	-15.56	-144.97	113.53
IDQWL	σ	7.42	5.05	9.87	5.18	8.26	8.96	21.53	36.32
DQWLL	m	-96.43	134.88	-26.27	-91.8	-119.12	136.12	-67.20	-45.31
DQWLL	σ	12.23	0.71	24.85	93.74	2.68	9.34	10.57	13.70
QWLLK	m	36.65	-28.56	-62.31	-42.96	-67.29	-37.13	-32.11	-38.57
QWLLK	σ	31.10	4.88	1.71	1.83	2.18	1.7	7.34	0.76
WLLKN	m	76.75	-58.98	-71.43	-64.32	-133.51	137.20	-112.63	126.78
WLLKN	σ	6.81	63.79	19.58	45.42	8.72	23.98	5.70	4.25
LLKNA	m	-62.53	144.20	62.77	32.42	-54.17	-50.73	-66.32	-35.74
LLKNA	σ	1.05	3.48	1.03	3.22	8.62	8.22	4.18	5.42
LKNAK	m	-99.12	-2.29	-77.04	-40.79	53.88	45.63	-86.80	146.54
LKNAK	σ	11.27	8.81	22.25	9.69	12.07	24.32	34.39	12.63
KNAKE	m	-129.41	164.70	-64.45	-33.12	64.84	25.29	-85.43	127.66
KNAKE	σ	6.81	2.83	10.82	13.38	7.93	6.31	9.17	13.72
NAKED	m	-170.56	164.33	-60.13	-50.89	-63.69	141.25	-176.71	19.54
NAKED	σ	39.53	98.51	4.54	13.86	13.55	9.88	6.89	7.75
AKEDA	m	-67.74	-55.08	-129.46	-179.43	-59.54	159.11	-150.24	-2.89
AKEDA	σ	13.08	45.06	38.91	92.47	9.31	12.50	22.21	3.94
KEDAI	m	-62.53	133.45	-64.01	-43.31	-80.39	-17.98	66.64	2.94
KEDAI	σ	20.16	5.38	9.16	7.27	16.44	19.13	29.38	46.92
EDAIA	m	-98.99	-43.83	-158.00	150.10	-62.71	-42.68	-69.81	-41.36
EDAIA	σ	0.57	0.04	14.63	27.46	3.52	4.50	3.46	7.62
DAIAE	m	-135.27	122.43	-70.32	144.12	-76.03	-35.25	-62.76	-41.91
DAIAE	σ	3.62	16.10	2.30	7.19	4.61	16.25	6.51	6.00
AIAEL	m	-79.23	-1.18	-47.67	-43.32	-67.50	-35.14	-58.51	-41.64
AIAEL	σ	2.26	28.23	3.70	7.55	2.19	12.15	2.67	10.44
IAELK	m	-61.58	-46.52	-69.8	-38.92	-84.72	-9.56	-66.04	-55.14
IAELK	σ	4.71	3.18	5.91	4.28	11.23	5.74	3.01	0.21
AELKK	m	-65.61	-41.77	-123.59	117.93	-98.44	-38.95	-82.76	82.67
AELKK	σ	4.55	5.49	6.66	0.88	9.17	11.15	16.72	42.13
ELKKA	m	-89.24	-29.10	-71.98	120.31	-57.31	-41.31	-65.90	-37.48
ELKKA	σ	9.17	14.36	9.78	25.12	4.79	10.69	4.82	9.67
LKKAG	m	-96.27	-12.44	-63.24	122.75	-73.62	-24.38	-60.33	-36.07
LKKAG	σ	11.81	18.16	7.42	19.81	3.36	26.69	5.88	8.63
KKAGI	m	54.01	-97.58	-106.03	139.72	-58.30	125.96	-83.50	-14.27
KKAGI	σ	6.52	75.03	4.79	5.85	12.71	13.96	14.41	21.39

continua na próxima página

		Grupo 01		Grupo 02		Grupo 03		Grupo 04	
Frag		phi	psi	phi	psi	phi	psi	phi	psi
KAGIT	*m*	89.95	3.72	147.58	-160.64	-94.77	-165.04	-66.41	-37.51
KAGIT	σ	14.61	17.60	80.77	56.69	8.16	6.32	1.69	7.49
AGITS	*m*	-63.35	-39.96	-54.03	107.89	-79.93	124.22	-57.91	-45.88
AGITS	σ	5.21	13.05	8.83	63.90	4.00	8.4	3.95	7.76
GITSD	*m*	54.60	119.14	-105.09	170.31	-120.32	-173.99	-103.48	-14.15
GITSD	σ	31.94	139.28	16.07	4.69	6.28	3.73	26.49	24.31
ITSDF	*m*	-66.11	-40.95	-109.16	139.55	90.01	-13.14	-104.39	18.02
ITSDF	σ	10.18	20.56	23.72	18.99	29.54	54.88	17.19	12.14
TSDFY	*m*	-72.64	75.66	-100.20	153.13	-69.03	-27.37	-136.72	-57.91
TSDFY	σ	6.33	1.88	20.02	7.70	11.56	5.23	26.11	89.57
SDFYF	*m*	-89.22	134.70	-59.77	-46.53	-123.14	140.48	-148.44	126.08
SDFYF	σ	34.96	74.88	34.96	74.88	0.47	3.79	3.58	19.57
DFYFN	*m*	-87.68	-30.26	-79.15	136.04	-119.40	133.79	-147.68	153.36
DFYFN	σ	20.51	19.46	7.68	16.42	4.72	6.93	8.87	17.26
FYFNA	*m*	-58.04	-38.58	-127.73	138.80	-43.49	134.01	-104.27	157.81
FYFNA	σ	36.72	55.93	1.71	12.97	3.63	5.96	4.38	0.99
YFNAI	*m*	-123.87	178.82	-90.71	-19.01	-63.49	-36.83	-77.61	-27.64
YFNAI	σ	15.77	52.98	2.01	0.65	5.34	6.07	2.52	5.37
FNAIN	*m*	-121.30	52.32	-72.11	-25.64	64.11	-161.98	82.83	9.19
FNAIN	σ	2.23	69.27	9.96	13.44	0.55	3.96	8.59	12.63
NAINK	*m*	-74.34	-40.03	-59.98	-45.46	-60.76	-31.77	-103.57	15.75
NAINK	σ	7.54	7.21	2.65	3.94	5.87	4.59	9.28	11.71
AINKA	*m*	-67.29	-26.18	-35.33	-59.76	-73.98	146.37	-58.79	-44.46
AINKA	σ	6.92	11.86	5.42	4.61	7.15	16.02	7.38	7.32
INKAK	*m*	-151.29	118.10	-96.53	-11.21	-71.52	-16.93	-59.08	-33.11
INKAK	σ	2.96	3.91	6.93	10.96	0.17	4.16	10.27	7.06
NKAKT	*m*	-78.87	-28.2	-42.14	-106.07	-63.42	137.92	84.65	-170.22
NKAKT	σ	14.24	13.75	4.27	49.50	10.22	15.58	36.05	81.42
KAKTV	*m*	-105.92	130.57	-58.50	-49.46	92.00	-1.19	-101.09	-33.68
KAKTV	σ	23.24	20.39	8.19	10.40	41.18	82.26	7.84	19.98
AKTVE	*m*	-77.25	139.01	-69.07	-43.19	-151.03	152.22	-112.54	-166.29
AKTVE	σ	5.33	24.99	7.29	3.17	14.25	9.58	14.35	9.76
KTVEE	*m*	-55.5	-41.58	8.82	141.90	-68.63	146.98	-67.53	-34.94
KTVEE	σ	3.89	9.56	12.93	39.85	12.93	39.85	4.31	4.78
TVEEV	*m*	-91.83	110.67	-50.07	-49.99	-83.18	-35.01	-61.42	-40.44
TVEEV	σ	13.29	49.36	3.03	7.70	15.58	13.01	4.14	4.24
VEEVN	*m*	-103.68	128.93	167.59	-2.27	-76.33	-42.23	-154.27	130.83
VEEVN	σ	10.53	16.10	58.07	90.14	16.32	34.48	5.22	5.19
EEVNA	*m*	-111.16	135.98	-62.70	-47.41	-83.07	98.84	-53.07	-57.86
EEVNA	σ	4.47	5.37	3.35	5.33	3.97	8.65	5.86	6.80
EVNAL	*m*	-107.86	18.78	-123.47	110.73	-75.23	-33.28	-69.99	131.44
EVNAL	σ	8.23	17.10	8.95	39.55	21.04	9.05	13.48	9.36
VNALK	*m*	-70.74	-37.78	-91.82	-54.61	-64.36	-41.75	-55.81	-45.68
VNALK	σ	5.55	15.42	12.84	4.18	3.17	2.67	3.93	5.39
NALKN	*m*	-57.97	-51.75	-90.26	-25.04	-73.31	-9.19	-66.51	-44.56

continua na próxima página

		Grupo 01		Grupo 02		Grupo 03		Grupo 04	
Frag		phi	psi	phi	psi	phi	psi	phi	psi
NALKN	σ	1.92	2.29	4.63	11.08	3.41	5.41	7.41	10.79
ALKNE	*m*	-87.04	108.19	58.37	-126.30	-62.43	-38.43	-149.35	140.54
ALKNE	σ	4.41	5.25	38.20	81.71	4.40	11.67	7.85	27.64
LKNEI	*m*	80.92	-168.02	-78.32	115.64	57.15	47.22	-70.88	-36.99
LKNEI	σ	35.89	82.85	6.98	30.89	35.89	82.85	15.59	31.78
KNEIL	*m*	-55.65	-40.77	-69.76	-37.25	-91.00	138.66	-111.76	147.04
KNEIL	σ	0.25	0.34	5.42	2.23	2.12	5.52	17.93	93.47
NEILK	*m*	-55.20	-48.61	-81.65	126.63	-64.57	-41.30	-110.46	-29.20
NEILK	σ	4.59	9.47	18.86	31.07	4.76	11.80	4.26	29.63
EILKA	*m*	-58.61	-32.82	-58.30	-48.38	-65.15	-42.11	-66.17	-26.75
EILKA	σ	20.21	0.19	4.54	6.30	5.31	3.82	6.19	1.84
ILKAH	*m*	-116.25	-41.94	47.50	43.02	-57.15	-43.14	-70.84	-28.53
ILKAH	σ	6.62	20.30	8.81	8.89	7.17	8.29	2.48	7.08
LKAHA	*m*	-62.84	-40.81	-119.06	166.07	-93.52	32.48	-103.90	106.05
LKAHA	σ	3.53	4.57	21.47	73.4	1.88	3.90	6.33	1.03

APÊNDICE J – Agrupamento das tuplas molde da proteína 1UTG

Tabela 41: Agrupamento das tuplas-molde associadas a um fragmento alvo s_i da proteína cujo código PDB é 1UTG: (m) é o valor médio e (σ) é o desvio padrão estimado de cada grupo.

		Grupo 01		Grupo 02		Grupo 03		Grupo 04	
Frag		phi	psi	phi	psi	phi	psi	phi	psi
GICPR	m	-61.38	-50.05	-76.01	128.63	-122.46	118.10	-97.29	-24.60
GICPR	σ	12.19	6.53	13.84	11.14	14.26	28.86	25.17	88.96
ICPRF	m	67.11	20.89	-112.84	132.22	-71.95	-33.93	-135.82	126.07
ICPRF	σ	40.76	72.01	10.22	4.32	19.39	20.62	11.29	5.11
CPRFA	m	-82.64	-11.47	-92.99	101.97	-109.05	139.38	-65.10	-37.02
CPRFA	σ	14.21	58.22	4.17	23.95	2.76	2.47	6.16	4.39
PRFAH	m	-108.92	34.08	-74.2	153.42	-70.18	-32.83	102.35	-29.06
PRFAH	σ	58.42	87.67	21.43	1.82	13.15	6.86	0.68	0.41
RFAHV	m	-80.75	149.05	-120.36	-31.66	-122.70	149.24	-63.16	-27.44
RFAHV	σ	1.13	2.04	4.12	8.96	17.14	15.77	5.14	10.78
FAHVI	m	-80.24	137.53	-61.43	-35.13	-143.37	147.91	-117.46	146.24
FAHVI	σ	9.54	14.22	5.38	9.62	5.92	9.81	4.82	27.17
AHVIE	m	-49.76	-62.75	-60.71	-46.60	-73.18	-17.23	-90.71	123.87
AHVIE	σ	0.46	1.44	6.84	6.60	9.75	11.41	6.64	8.78
HVIEN	m	-55.48	-42.33	-121.82	142.18	-65.93	-32.06	-83.11	118.73
HVIEN	σ	5.04	6.20	12.84	22.07	5.23	5.83	14.16	13.26
VIENL	m	-64.18	-41.42	-102.36	132.63	-121.79	-152.72	55.72	-114.12
VIENL	σ	6.23	10.24	26.54	27.67	35.61	9.84	31.85	90.06
IENLL	m	-75.56	-26.65	-61.49	-39.74	-57.74	-63.08	-133.96	126.52
IENLL	σ	4.41	14.73	5.20	5.10	3.05	6.42	15.48	34.71
ENLLL	m	-90.40	29.52	-113.87	112.60	-107.23	102.78	-63.77	-44.04
ENLLL	σ	8.30	46.51	5.76	7.44	1.02	8.97	6.97	11.43
NLLLG	m	-60.12	132.43	-103.66	162.60	-75.89	-31.60	-114.08	161.12
NLLLG	σ	19.42	94.59	3.37	3.19	6.02	4.37	5.98	9.45
LLLGT	m	-77.76	-15.55	-121.55	149.94	-126.6	-6.93	-146.71	120.75
LLLGT	σ	9.94	10.34	8.78	9.62	4.27	5.06	12.84	5.78
LLGTP	m	-99.28	97.87	107.13	151.03	-83.01	-169.92	84.83	15.03
LLGTP	σ	29.14	57.96	12.50	14.97	54.63	4.15	14.20	16.55
LGTPS	m	-141.56	130.76	-67.85	123.68	-133.97	65.80	-87.56	159.45
LGTPS	σ	4.26	5.07	30.74	36.56	1.91	3.43	11.49	7.02
GTPSS	m	-72.72	168.93	-56.47	142.76	-49.15	-43.27	-66.69	154.82
GTPSS	σ	8.72	90.48	1.25	2.20	1.93	3.99	0.35	2.12
TPSSY	m	94.23	-40.07	-61.11	137.34	-57.75	-38.00	-97.68	47.66
TPSSY	σ	8.98	45.95	2.82	0.87	7.22	14.44	18.85	31.56
PSSYE	m	-102.02	19.83	-59.77	-37.05	-113.92	137.89	95.08	-11.60
PSSYE	σ	15.83	14.46	10.49	33.39	43.55	15.67	7.41	5.39
SSYET	m	-50.97	-30.67	-162.49	84.15	-54.32	-47.62	-61.72	-48.55
SSYET	σ	0.05	0.32	25.93	32.36	3.67	1.49	2.40	5.99

continua na próxima página

		Grupo 01		Grupo 02		Grupo 03		Grupo 04	
Frag		phi	psi	phi	psi	phi	psi	phi	psi
SYETS	*m*	-96.18	56.95	59.52	31.33	-122.05	158.63	-72.38	-20.36
SYETS	σ	3.67	8.63	48.12	66.78	32.87	9.54	10.13	20.84
YETSL	*m*	-124.72	-82.98	176.79	35.92	-61.51	-31.65	-125.64	-2.13
YETSL	σ	48.89	60.19	0.42	0.17	4.93	12.95	25.28	98.03
ETSLK	*m*	-100.49	-26.45	-63.90	-19.66	-160.15	150.66	-58.49	-42.47
ETSLK	σ	28.46	52.62	3.77	2.49	2.09	16.11	4.55	6.20
TSLKE	*m*	-97.79	-1.80	-60.07	-41.51	-68.31	-12.88	-51.24	129.99
TSLKE	σ	5.89	4.23	6.48	5.74	6.79	6.60	9.00	7.53
SLKEF	*m*	-111.69	13.29	-83.85	148.46	-57.06	-35.84	-90.18	-14.27
SLKEF	σ	14.62	51.65	3.24	0.33	4.32	6.68	14.62	51.65
LKEFE	*m*	-90.95	-7.14	-62.59	-38.17	-138.46	147.83	-117.23	-3.83
LKEFE	σ	7.37	5.80	5.45	8.36	13.46	19.45	5.54	19.43
KEFEP	*m*	-67.03	147.62	-136.47	149.12	63.12	38.93	-112.07	-3.25
KEFEP	σ	6.62	9.00	7.84	14.27	2.43	1.45	21.54	19.27
EFEPD	*m*	-144.74	139.88	-96.86	112.75	-91.17	148.77	54.74	63.33
EFEPD	σ	4.21	13.74	39.67	16.14	16.75	7.58	5.47	4.95
FEPDD	*m*	-52.72	-22.61	-58.87	140.70	-31.09	-48.78	-78.37	155.12
FEPDD	σ	6.65	16.96	6.21	6.45	8.83	12.42	2.16	3.15
EPDDT	*m*	-68.41	-12.85	-83.97	23.98	-69.07	173.35	-62.31	-33.63
EPDDT	σ	1.08	3.44	6.66	79.74	4.72	3.31	1.36	13.74
PDDTM	*m*	99.91	-15.87	-139.51	1.10	-108.11	2.96	-65.28	-33.63
PDDTM	σ	12.13	2.76	2.63	3.67	29.99	5.42	2.24	6.34
DDTMK	*m*	114.82	175.62	-118.47	38.33	-66.77	-22.85	-62.57	-47.96
DDTMK	σ	22.75	29.19	22.75	29.19	7.01	10.50	5.44	5.14
DTMKD	*m*	-117.56	-2.40	-51.26	-64.55	-154.56	111.05	-69.43	-38.10
DTMKD	σ	10.25	12.88	12.49	19.65	24.61	28.88	3.92	8.65
TMKDA	*m*	-67.65	-40.11	-75.54	154.74	59.30	38.52	-52.31	-59.90
TMKDA	σ	6.75	5.67	8.70	14.74	32.74	70.08	4.25	11.41
MKDAG	*m*	-95.51	25.86	48.24	36.29	-94.56	-148.36	-62.40	-34.24
MKDAG	σ	23.50	40.58	23.44	37.00	3.64	30.05	8.84	12.31
KDAGM	*m*	53.67	41.43	-104.72	-14.55	-69.96	-22.01	-60.11	84.99
KDAGM	σ	28.72	28.10	12.58	31.28	5.80	11.69	28.72	28.10
DAGMQ	*m*	-65.33	-45.81	102.28	-173.73	79.19	78.07	-114.15	139.00
DAGMQ	σ	6.20	6.28	9.32	1.40	13.76	68.98	10.23	21.34
AGMQM	*m*	-107.15	-9.19	86.76	-145.35	-135.78	160.11	-65.54	-27.74
AGMQM	σ	23.30	58.93	43.23	96.39	24.68	22.45	6.91	18.19
GMQMK	*m*	-56.70	-39.22	-94.84	89.20	-99.16	89.82	-62.42	-40.25
GMQMK	σ	2.94	11.10	1.29	1.10	3.82	6.92	2.08	2.95
MQMKK	*m*	-67.03	-39.53	-69.37	130.52	-89.83	12.87	-113.20	135.89
MQMKK	σ	7.87	7.03	24.42	1.90	22.01	83.11	9.39	21.19
QMKKV	*m*	-134.98	147.71	-86.60	-40.07	-62.10	-34.27	-69.90	159.60
QMKKV	σ	17.14	16.37	12.04	3.97	4.18	8.28	8.12	5.26
MKKVL	*m*	169.06	-18.53	-83.28	138.90	-62.02	-48.70	-99.03	-18.33
MKKVL	σ	37.74	78.67	26.65	16.56	7.06	8.22	21.37	11.13
KKVLD	*m*	-71.26	-33.39	-99.24	117.69	-60.89	-44.99	-97.99	-50.64

continua na próxima página

		Grupo 01		Grupo 02		Grupo 03		Grupo 04	
Frag		phi	psi	phi	psi	phi	psi	phi	psi
KKVLD	σ	7.68	10.08	5.67	3.14	6.03	9.42	7.9	15.74
KVLDS	*m*	-99.92	135.38	-59.62	145.33	-120.85	136.62	-70.07	-28.65
KVLDS	σ	5.25	7.92	17.98	74.54	1.86	3.45	4.75	17.13
VLDSL	*m*	-60.55	-41.93	-65.21	147.14	-147.95	144.72	-68.7	-27.03
VLDSL	σ	3.26	3.70	16.10	8.36	1.62	14.23	17.49	15.80
LDSLP	*m*	65.78	-70.75	-135.11	152.36	-98.79	-28.81	-82.52	-13.11
LDSLP	σ	6.15	97.82	22.11	8.10	34.44	51.38	7.46	9.98
DSLPQ	*m*	-110.98	151.10	-62.75	-45.91	-160.81	101.27	-76.29	143.05
DSLPQ	σ	8.10	3.49	0.64	0.27	11.43	27.2	13.53	5.54
SLPQT	*m*	-49.62	142.98	-59.48	152.45	-	-	-	-
SLPQT	σ	6.97	6.69	6.97	6.69	-	-	-	-
LPQTT	*m*	79.07	-28.80	-98.92	155.76	-58.39	-28.35	-77.09	-0.23
LPQTT	σ	34.41	78.41	26.21	6.05	7.81	13.17	22.56	17.43
PQTTR	*m*	-64.47	-41.13	50.03	49.39	-145.71	-161.79	-114.27	119.03
PQTTR	σ	10.36	9.43	42.85	91.71	42.85	91.71	25.39	36.15
QTTRE	*m*	-77.41	-34.45	-163.06	142.12	-83.07	128.82	-60.29	-44.95
QTTRE	σ	12.79	4.74	1.51	1.82	14.48	17.74	3.43	2.82
TTREN	*m*	-166.12	162.07	-79.67	56.43	-51.32	-27.37	-52.66	-38.19
TTREN	σ	43.08	78.52	0.01	0.05	11.63	24.01	8.18	8.95
TRENI	*m*	-60.22	-43.14	-104.73	134.84	114.6	-14.65	-119.59	-115.93
TRENI	σ	7.43	11.76	6.17	16.72	32.89	45.33	18.32	50.77
RENIM	*m*	-160.41	165.01	55.74	36.49	-104.79	145.84	-64.49	-38.15
RENIM	σ	53.52	84.27	8.90	11.45	5.61	2.68	10.34	10.90
ENIMK	*m*	-57.51	-50.93	-75.69	-1.61	-58.08	-9.72	-68.58	-39.86
ENIMK	σ	0.72	2.58	4.53	7.93	7.40	21.77	2.40	5.40
NIMKL	*m*	-118.38	66.68	-65.55	-40.15	-57.71	129.41	19.61	-169.79
NIMKL	σ	17.41	72.05	5.15	7.01	5.45	9.99	31.93	87.97
IMKLT	*m*	-67.58	-33.96	-64.2	-48.66	-93.17	127.06	-110.77	123.76
IMKLT	σ	3.02	4.09	4.64	5.41	4.35	6.44	7.71	9.41
MKLTE	*m*	-91.27	-14.26	-60.47	-34.83	-69.72	157.04	-127.09	146.37
MKLTE	σ	12.68	32.99	8.38	7.30	7.75	14.98	7.79	3.49
KLTEK	*m*	-85.43	166.76	-121.58	-176.48	-61.34	-40.68	-122.38	164.90
KLTEK	σ	8.75	4.56	21.42	109.01	8.41	11.01	8.63	14.27
LTEKI	*m*	-84.98	-32.35	-63.25	-36.74	-133.34	157.12	-77.86	144.44
LTEKI	σ	6.84	6.38	8.72	12.38	14.32	91.73	5.61	9.04
TEKIV	*m*	-57.12	-46.90	-79.20	164.72	-120.57	101.43	-126.4	-17.33
TEKIV	σ	11.07	12.05	35.6	73.86	17.37	11.24	14.15	8.25
EKIVK	*m*	-107.25	-66.89	-64.45	-41.97	-93.21	30.48	-87.87	125.68
EKIVK	σ	6.87	8.80	7.63	7.90	0.00	47.20	6.28	15.24
KIVKS	*m*	-70.72	-20.26	-61.35	-43.77	-70.79	-20.6	-50.69	137.78
KIVKS	σ	1.29	6.11	0.96	0.45	1.29	6.11	8.41	76.99
IVKSP	*m*	-103.16	145.71	-34.07	119.01	-114.80	109.91	-97.87	-9.69
IVKSP	σ	7.37	14.30	28.66	64.43	7.06	9.44	16.00	10.03
VKSPL	*m*	-69.95	154.26	-51.56	-51.59	-137.32	92.83	-116.26	129.86
VKSPL	σ	14.65	12.66	7.30	4.37	4.20	30.51	9.81	31.88

continua na próxima página

		Grupo 01		Grupo 02		Grupo 03		Grupo 04	
Frag		phi	psi	phi	psi	phi	psi	phi	psi
KSPLC	m	-113.62	21.75	-77.92	4.23	-59.4	-28.47	-47.09	-46.63
KSPLC	σ	2.97	3.03	29.10	28.06	4.89	5.33	0.08	0.14
SPLCM	m	-118.82	150.15	-66.88	-41.38	-82.17	-11.31	168.12	-49.66
SPLCM	σ	33.78	4.29	9.46	15.74	11.67	5.67	70.32	91.91