

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

**Um estudo sobre a predição da estrutura 3D
aproximada de proteínas utilizando o método
CReF com refinamento**

Karina Cristina da Motta Dall'Agno

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

Orientador: Prof. Dr. Osmar Norberto de Souza

Porto Alegre, janeiro de 2012.

Dados Internacionais de Catalogação Pública (CIP)

D147e Dall'Agno, Karina Cristina da Motta
Um estudo sobre a predição da estrutura 3 D aproximada de
proteínas utilizando o método CReF com refinação / Karina
Cristina da Motta. – Porto Alegre, 2012.
132 f.

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

1. Informática. 2. Biologia Computacional. Mineração de
Dados (Informática). I. Souza, Osmar Norberto de. II. Título.

CDD 005.74

**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

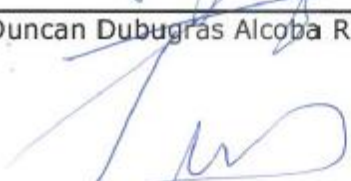
Dissertação intitulada "Um Estudo Sobre a Predição da Estrutura 3D Aproximada de Proteínas Utilizando o Método CReF com Refinamento", apresentada por Karina Cristina da Motta Dall'Agno como parte dos requisitos para obtenção do grau de Mestre em Ciência da Computação, Bioinformática e Modelagem Computacional, aprovada em 22/03/2012 pela Comissão Examinadora:



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



Prof. Dr. Duncan Dubugras Alcoba Ruiz - PPGCC/PUCRS



Prof. Dr. Luiz Augusto Basso - PPGBCM/PUCRS

Homologada em 24/04/2012, conforme Ata No. 009 pela Comissão Coordenadora.



Prof. Dr. Paulo Henrique Lemelle Fernandes
Coordenador.

PUCRS

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

AGRADECIMENTOS

Gostaria de agradecer em primeiro lugar à minha mãe Mercedes por seu apoio incondicional, por seu imenso carinho e dedicação, pela compreensão dos momentos em que estive ausente, por toda a ajuda que me dispensou para que eu pudesse dedicar meu tempo livre neste trabalho. Além disso, obrigada por tudo o que fez por mim e por tudo o que sou.

Agradeço ao professor Osmar pela orientação, pelo conhecimento compartilhado, pela compreensão e pela parceria na realização deste trabalho. Muito obrigada à colega Renata pela troca de informações, conhecimentos e, especialmente, pela amizade. Agradecimentos aos colegas Christian e Luís Fernando pela contribuição e apoio técnico. Um especial agradecimento ao amigo e grande incentivador para a realização deste mestrado, Cristiano Galina. Obrigada à empresa Tlantic, em especial à gerente Angélica, pela compreensão e incentivo que me possibilitou tempo adicional para dedicar a este trabalho enquanto desempenhava minha atividade profissional. De forma geral, meus agradecimentos aos parentes, amigos e colegas que me ajudaram, entenderam e torceram por mim.

RESUMO

Um dos principais desafios da Bioinformática Estrutural é entender como a informação decodificada em uma sequência linear de aminoácidos, ou estrutura primária de uma proteína, possibilita a formação de sua estrutura tridimensional. Muitos algoritmos buscam propor soluções para o problema complexo da classe NP-completo. Dentre eles, está o método CReF (*Central Residue Fragment-based method*) que realiza a predição da estrutura 3D aproximada de proteínas ou polipeptídios. O método usa técnicas de mineração de dados para agrupar dados de estruturas, apresentando boa predição de estruturas secundárias, bom desempenho em máquina de baixo custo, mas enfrenta problemas na predição das regiões de voltas e alças e na usabilidade. Valorizando as características diferenciadas do método e buscando sua evolução, este trabalho propôs-se a realizar melhorias no CReF. Após uma etapa inicial de entendimento e adaptações para tornar a ferramenta executável na situação atual dos bancos de dados e ferramentas de apoio, foram identificadas duas categorias de melhorias. As melhorias técnicas tiveram por objetivo automatizar a ferramenta, adaptá-la ao ambiente e ao usuário enfatizando usabilidade. Para melhorias no método realizaram-se testes com variação na quantidade de grupos identificados na etapa de mineração de dados com o algoritmo *Expectation Maximization* (EM) no Weka. Os testes indicaram que as melhores conformações iniciais eram obtidas com quatro e seis grupos, assim, optou-se por permitir ao usuário a escolha dos grupos a considerar. Um novo mapeamento do mapa de Ramachandran indicou ajustes que foram corrigidos e decidiu-se descartar grupos identificados nas regiões não permitidas na análise do resultado da mineração de dados. A nova versão do CReF, gerada pela implementação dessas melhorias, também padronizou o método de predição de estrutura secundária, passando a utilizar o método Porter. Como consequência, as regras para escolha do grupo resultante da mineração a representar cada aminoácido foram adaptadas e ampliadas para atender novas situações. A nova versão manteve o desempenho de predição e execução iniciais do CReF, entretanto, manteve o problema das voltas e alças. Este problema de otimização das regiões de voltas e alças foi endereçado por meio do desenho e aplicação de um protocolo de refinamento, baseado em simulações pelo método da dinâmica molecular, o qual apresentou um resultado expressivo para a proteína alvo de código PDB 1ZDD.

Palavras-chave: Bioinformática, predição de estrutura 3D de proteínas, mineração de dados, refinamento de estruturas proteicas, simulação por dinâmica molecular, predição de estrutura 2D.

ABSTRACT

One of the most important problems in Structural Bioinformatics is to understand how the information coded in linear sequence amino acids, or primary structure, is translated into the three-dimensional structure of a protein. Many algorithms proposed solutions to this complex problem of NP-complete class. One of them is the CReF method (Central Residue Fragment-based) which makes prediction of approximate 3-D structure of proteins and polypeptides. The method uses data mining techniques to group data structures, showing good secondary structure prediction, good performance at low machine cost, but has problems in the prediction of turns and loops regions and usability. Valuing the different characteristics of CReF and seeking to evolve it, this work proposes improvements to CReF. After the initial stage of understanding the tool and making changes to turn it executable on the current state of data banks and support tools, two categories of improvements to make were identified. The technical improvements aimed to automate CReF, adapting it to the environment and emphasizing usability. In the method's improvements variations on the amount of groups were tested for data mining with the Expectation Maximization algorithm in Weka. Tests indicated that the best results for the initial conformation were for four and six groups, hence we decided to allow the user to select the amount of groups. A new mapping of the data in the Ramachandran plot indicated some problems that had to be fixed. In the analysis of data mining results, we decided that groups in regions not allowed would be discarded. The new version of CReF generated by the implementation of these improvements standardized the method of secondary structure prediction to use Porter. As a consequence, the rules of selection of data mining groups to represent each amino acids have been changed and extended. The new version has the same initial performance of CReF in prediction and execution, however, the problem of correct predictions of turns and loops remained. This problem was addressed through a refinement protocol, based on simulations by the molecular dynamics method, which presented a significant result for the target protein 1ZDD.

Keywords: Bioinformatics, three-dimensional protein structure prediction, data mining, refinement of protein structures, molecular dynamics simulations, secondary structure prediction.

LISTA DE FIGURAS

Figura 1 – Estrutura química de um aminoácido.	20
Figura 2 – Ligação peptídica: dois aminoácidos se ligam entre o átomo C da carbonila do primeiro e o átomo N da amina do segundo, ocorrendo uma desidratação que libera água e forma-se a ligação peptídica.....	24
Figura 3 – Representação esquemática de um peptídeo identificando os ângulos de torção da cadeia principal ϕ_i , ψ_i e ω_i	25
Figura 4 – Mapa de Ramachandran: região mais favorável em vermelho, região permitida em amarelo, região ainda aceitável em amarelo claro e região não permitida em branco.....	27
Figura 5 – Definições dos estados conformacionais no mapa de Ramachandran segundo A. V. Efimov.	27
Figura 6 – Hélice α mostrando o esqueleto peptídeo.	29
Figura 7 – Representação dos diferentes tipos de hélices. Neste desenho foi usada a representação CPK para os átomos e a hélice foi desenhada passando pelos átomos da cadeia principal.	30
Figura 8 – Representação da folha β : em (A) folha paralela e em (B) folha antiparalela.....	31
Figura 9 – Volta I da Thermolysin: 12-15 (Gly-Val-Leu-Gly).....	34
Figura 10 – Volta I' da Actinidina: 172-175 (Gly-Gly-Glu-Val).....	34
Figura 11 – Volta II' da Elastase: 36B-37 (Gly-Ser-Ser-Ser).	34
Figura 12 – Voltas da actinidina (código PDB: 2ACT - à esquerda) e da elastase (código PDB: 3EST - à direita) no mapa de Ramachandran.....	35
Figura 13 – Representação (A) das espirais desordenadas (<i>loop</i> ou <i>random coil</i>) e (B) da estrutura secundária irregular volta (<i>turn</i> em inglês).	35
Figura 14 – Exemplos das quatro estruturas supersecundárias. As combinações de estruturas secundárias (alças e hélices) dão nome aos motivos.....	36
Figura 15 – Representação do tipo <i>Ribbons</i> da estrutura terciária da proteína Crambina (código PDB: 1CRN) composta por duas hélices α e duas estruturas de fitas β , sendo uma delas antiparalela e que estão conectadas por uma estrutura irregular do tipo volta.....	37
Figura 16 – Estrutura quaternária da hemoglobina (código PDB: 1A00), sem o grupo heme, em representação do tipo <i>Ribbons</i> , identificando as quatro cadeias: A em roxo, B em amarelo, C em verde e D em ciano.	38
Figura 17 – (A) Proteína classe α : calponina da utrofina (código PDB: 1BHD), (B) Proteína classe β : módulo de adesão celular tipo III-10 da fibronectina (código PDB: 1FNA).	39
Figura 18 – Proteína classe $\alpha + \beta$: domínio principal da TBP (código PDB: 1CDW).	39
Figura 19 – Proteínas Classe α/β : em A a acilfosfatase (código PDB: 2ACY); em B a tioredoxina (código PDB: 1THX); e em C a proteína Chey (código PDB: 3CHY).....	40
Figura 20 – Proteína classe barril α/β : Glicolato oxidase de espinafre (código PDB: 1GOX).....	40
Figura 21 – Crescimento anual do número total de estruturas 3D de proteínas no PDB.....	42
Figura 22 – Esquema com as nove etapas do método CReF. As etapas dois e sete são executadas remotamente e as demais localmente (Dorn, 2008).	47
Figura 23 – Esquema representando uma sequência alvo hipotética K dividida em p fragmentos s_i . Cada fragmento, por sua vez, é caracterizado pelo duplete, ou ângulos de torção ϕ e ψ , de seu resíduo central.	47
Figura 24 – Cálculo dos ângulos de torção de cada duplete do aminoácido central da sequência alvo K	49
Figura 25 – Representação no mapa de Ramachandran: (A) tuplas ocupando regiões do mapa, em vermelho a região mais favorável, em amarelo a região permitida, em amarelo claro a região ainda aceitável e em branco a região não permitida. (B) representa a delimitação de um intervalo entre os ângulos de torção mínimo e máximo de ϕ e ψ , indicados pelos pontos P_1 e P_2	51

Figura 26 – Esquema da escolha dos grupos k_i que representam os ângulos de torção dos aminoácidos de uma sequência alvo K	53
Figura 27 – Fluxograma do método de otimização de voltas de polipeptídeos representados na forma de intervalos de variação angular do método CReF.	55
Figura 28 – Fluxograma da versão inicial do CReF.	59
Figura 29 – Fluxograma da versão atual do CReF.....	60
Figura 30 – Comparativo da conformação inicial das proteínas 1ZDD e 1GB1 com quatro e seis grupos na mineração dos dados: (A) estrutura experimental da 1ZDD, (B) estrutura inicial da 1ZDD com 4 grupos (RMSD: 9,82 Å), (C) estrutura inicial da 1ZDD com 6 grupos (RMSD: 6,64 Å), (D) estrutura experimental da 1GB1, (E) estrutura inicial da 1GB1 com 4 grupos (RMSD: 21,05 Å), (F) estrutura inicial da 1GB1 com 6 grupos (RMSD: 14,85 Å).	64
Figura 31 – Proteínas pequenas: (A) 2ERL, (B) 1YWJ, (C) 1GPT.....	72
Figura 32 – Proteínas médias: (A) 1CSP, (B) 1CTF, (C) 1C5A, (D) 1OPD.	73
Figura 33 – Proteínas grandes: (A) 2EZK, (B) 1KSR, (C) 1ERV.....	74
Figura 34 – Parâmetros para execução na nova versão do CReF.	75
Figura 35 – Mapa de Ramachandran da proteína alvo 1ZDD: (A) mapa da estrutura experimental, (B) mapa da estrutura predita sem otimização com base em 4 grupos na etapa de mineração de dados, (C) mapa da estrutura predita sem otimização com base em 6 grupos na mineração.....	80
Figura 36 – Comparação da conformação inicial da proteína 1ZDD com variação de grupos na mineração dos dados: (A) estrutura experimental da 1ZDD, (B) conformação inicial com 4 grupos (RMSD: 9,82 Å) e (C) conformação inicial com 6 grupos (RMSD: 6,64 Å).	80
Figura 37 – Estrutura 3D predita (conformação inicial) pela versão inicial do CReF antes da otimização de voltas (RMSD: 5,51 Å).	81
Figura 38 – Mapa de Ramachandran da proteína alvo 1GB1: (A) mapa da estrutura experimental, (B) mapa da estrutura predita sem otimização com base em 4 grupos na etapa de mineração de dados, (C) mapa da estrutura predita sem otimização com base em 6 grupos na mineração.....	84
Figura 39 – Comparação da conformação inicial da proteína 1GB1 com variação de grupos na mineração dos dados: (A) estrutura experimental da 1GB1, (B) conformação inicial com 4 grupos (RMSD: 21,05 Å) e (C) conformação inicial com 6 grupos (RMSD: 14,85 Å).	85
Figura 40 – Estrutura 3D predita (conformação inicial) pela versão inicial do CReF antes da otimização de voltas (RMSD: 12,31 Å).	86
Figura 41 – Mapa de Ramachandran da proteína alvo 1C5A: (A) mapa da estrutura experimental, (B) mapa da estrutura predita sem otimização com base em 4 grupos na etapa de mineração de dados, (C) mapa da estrutura predita sem otimização com base em 6 grupos na mineração.....	88
Figura 42 – Comparação da conformação inicial da proteína 1C5A com variação de grupos na mineração dos dados: (A) estrutura experimental da 1C5A, (B) conformação inicial com 4 grupos (RMSD: 12,29 Å) e (C) conformação inicial com 6 grupos (RMSD: 9,56 Å).	89
Figura 43 – Mapa de Ramachandran da proteína alvo 1OPD: (A) mapa da estrutura experimental, (B) mapa da estrutura predita sem otimização com base em 4 grupos na etapa de mineração de dados, (C) mapa da estrutura predita sem otimização com base em 6 grupos na mineração.....	92
Figura 44 – Comparação da conformação inicial da proteína 1OPD com variação de grupos na mineração dos dados: (A) estrutura experimental da 1OPD, (B) conformação inicial com 4 grupos (RMSD: 25,67 Å) e (C) conformação inicial com 6 grupos (RMSD: 15,19 Å).	93
Figura 45 – RMSD da trajetória dinâmica, durante o refinamento da estrutura predita 1ZDD_P em relação à estrutura experimental 1ZDD, como função do tempo de simulação.....	100
Figura 46 – Representação do tipo <i>Ribbons</i> da cadeia principal ($C\alpha$ átomos) da estrutura experimental da proteína 1ZDD (preto) e de conformações representativas da trajetória dinâmica durante o refinamento da estrutura predita 1ZDD_P (cinza).....	101
Figura 47 – Comparação de conformações predita e refinada da proteína 1ZDD: (A) estrutura experimental da 1ZDD, (B) conformação inicial com 4 grupos na mineração de dados (RMSD: 9,82 Å) e (C) conformação predita refinada (RMSD: 1,29 Å).....	102

Figura 48 – Estrutura 3D final predita pelo CReF: (A) estrutura 3D após a etapa de otimização da versão inicial (RMSD: 5,00 Å), (B) estrutura 3D refinada da nova versão (RMSD: 1,29 Å) . . 109

LISTA DE TABELAS

Tabela 1 – Classificação e informações dos aminoácidos.	20
Tabela 2 – Número de ângulos χ presentes em cada aminoácido.	26
Tabela 3 – Ângulos de torção para tipos de hélices.	30
Tabela 4 – Tendência e classificação dos resíduos para hélices α ou folhas β	33
Tabela 5 – Alguns exemplos de tipos de voltas conforme a nomenclatura de Venkatachalam.	34
Tabela 6 – Regiões conformacionais do mapa de Ramachandran segundo Thornton e colaboradores.	52
Tabela 7 – Exemplo de predição de estrutura secundária para o código PDB: 1ZDD.	52
Tabela 8 – Conjunto inicial de proteínas teste utilizado nos experimentos de predição com o método CReF.	71
Tabela 9 – Quantidade de moldes por fragmento da proteína alvo 1ZDD.	78
Tabela 10 – Análise do RMSD em Å das estruturas secundárias e conformação inicial das predições da 1ZDD.	79
Tabela 11 – Quantidade de moldes por fragmento da proteína alvo 1GB1.	82
Tabela 12 – Análise do RMSD em Å das estruturas secundárias e conformação inicial das predições da 1GB1.	84
Tabela 13 – Quantidade de moldes por fragmento da proteína alvo 1C5A.	87
Tabela 14 – Análise do RMSD em Å das estruturas secundárias e conformação inicial das predições da 1C5A.	88
Tabela 15 – Quantidade de moldes por fragmento da proteína alvo 1OPD.	90
Tabela 16 – Análise do RMSD em Å das estruturas secundárias e conformação inicial das predições da 1OPD.	91
Tabela 17 – Resumo dos demais experimentos realizados com a nova versão do CReF com valores de RMSD calculados em Å.	94
Tabela 18 – Resultados dos experimentos com a nova versão do CReF. Em negrito estão indicados os melhores valores de RMSD para cada proteína alvo.	106
Tabela 19 – Comparação entre a estrutura experimental e as estruturas preditas considerando quatro e seis grupos na mineração de dados para o conjunto de proteínas teste da nova versão do CReF.	119
Tabela 20 – Avaliação dos métodos de predição de estrutura secundária.	127

LISTA DE FÓRMULAS

Fórmula 1.....	32
Fórmula 2.....	32
Fórmula 3.....	49
Fórmula 4.....	50

LISTA DE SIGLAS

2D	Secundária
3D	Tridimensional
AMBER	Assisted Model Building with Energy Refinement
CASP	Critical Assessment of Techniques for Protein Structure Prediction
CReF	Central Residue Fragment-based method
DM	Dinâmica Molecular
EM	Expectation Maximization
EP	Energia Potencial
GB	Generalized Born
NMR	Nuclear Magnetic Resonance
PDB	Protein Data Bank
ps	Picossegundo
RMSD	Root Mean Square Deviation

SUMÁRIO

1. Introdução	15
2. Bioinformática estrutural	17
2.1 Resumo do capítulo	18
3. Proteínas e sua organização estrutural	19
3.1 Aminoácidos.....	19
3.2 Ligação peptídica e ângulos de torção	23
3.3 O diagrama de Sasisekharan-Ramakrishnan-Ramachandran.....	26
3.4 Hierarquia estrutural das proteínas.....	28
3.5 Classificação Estrutural de Proteínas	39
3.6 Bancos de dados de estruturas de proteínas	41
3.7 Resumo do capítulo	44
4. Predição <i>ab initio</i> ou <i>de novo</i> da estrutura 3D de proteínas com o método CReF	45
4.1 Etapa 1: fragmentação da sequência alvo.....	47
4.2 Etapa 2: busca por proteínas molde.....	48
4.3 Etapa 3: cálculo dos ângulos de torção dos dupletos	48
4.4 Etapa 4: agrupamento de dupletos.....	49
4.5 Etapa 5: representação dos ângulos de torção na forma de intervalos	50
4.6 Etapa 6: classificação dos grupos em regiões ocupadas no mapa de Ramachandran	51
4.7 Etapa 7: predição da estrutura secundária	52
4.8 Etapa 8: construção da conformação inicial.....	53
4.9 Etapa 9: otimização das regiões de volta.....	54
4.10 Resumo do capítulo	54
5. Desenvolvimento de nova versão do CReF.....	57
5.1 Melhorias implementadas no CReF	58
5.1.1 Melhorias técnicas	58
5.1.2 Alterações no método	63
5.2 O problema da otimização de voltas e alças.....	66
5.3 Resumo do capítulo	69
6. Experimentos	71
6.1 Definição de novo conjunto de proteínas teste.....	71
6.2 Materiais e métodos.....	74
6.3 Experimentos com a nova versão do CReF.....	77
6.3.1 Estudo de caso 1: 1ZDD	77
6.3.2 Estudo de caso 2: 1GB1.....	82
6.3.3 Estudo de caso 3: 1C5A.....	86

6.3.4	Estudo de caso 4: 1OPD	89
6.3.5	Demais estudos de caso	93
6.4	Desempenho do CReF.....	96
6.5	Refinamento de conformações iniciais.....	97
6.5.1	Detalhes da simulação por DM da 1ZDD_P	98
6.5.2	Resultados do refinamento da 1ZDD_P	99
6.6	Resumo do capítulo	102
7.	Considerações Finais	105
7.1	Principais contribuições	110
7.2	Trabalhos futuros.....	110
	Referências.....	111
	Apêndice A – Comparação da conformação inicial predita pelo CReF	119
	Apêndice B – Análise dos métodos de predição de estrutura secundária.....	127

1. Introdução

A Bioinformática é uma ciência interdisciplinar, na qual todos estão engajados em desvendar o mistério da vida: o genoma. Com este objetivo os projetos Genoma, dentre outros, têm gerado um volume impressionante de dados biológicos nos últimos tempos (Cochrane e Galperin, 2010). Esses dados, armazenados em bancos de dados, foram organizados de forma a permitir aos pesquisadores a realização de buscas e submissão de dados (Luscombe et al., 2001). Ferramentas e recursos foram desenvolvidos para diversas tarefas de análise desses dados (Luscombe et al., 2001).

Vencida a etapa de desenvolvimento e de melhoria dessas ferramentas, o que permitiu uma maior confiabilidade e disponibilidade das informações, apresenta-se agora uma nova etapa: a construção do conhecimento. Isso compreende interpretar e atribuir significado aos dados, gerando conhecimento (Prosdocimi et al., 2002). Essa nova fase tem sido chamada de “Era Pós-Genômica” e seu fruto é o proteoma que mapeia uma determinada função a uma proteína. Essa informação torna-se um instrumento poderoso para identificação de mecanismos celulares e a descoberta de novos fármacos mais eficientes e específicos (Desenvolvimento Racional de Fármacos) (Mandal, 2009). Um dos principais repositórios e fonte dessas informações é o *Protein Data Bank* (PDB), o mais volumoso banco de dados com informações de proteínas (Berman et al., 2000).

No processo de descoberta das funções de cada proteína, a identificação de sua estrutura tridimensional (3D) tem papel de destaque. A estrutura terciária é obtida de forma mais eficaz por meio de experimentos, mas estes são caros, de difícil execução e possuem limitações técnicas (Prosdocimi et al., 2002). Alternativamente a esta abordagem, é possível prever a conformação de uma proteína comparando-a com outra proteína homóloga de estrutura 3D conhecida (Prosdocimi et al., 2002). A modelagem por homologia, “baseada no conhecimento”, é um dos recursos mais eficientes para esta tarefa (Lesk, 2008; Prosdocimi et al., 2002; Martí-Renom et al., 2000). Ainda assim, a modelagem de proteínas é uma técnica heurística. Mesmo atendendo a todas as restrições, não há garantias de que seja a estrutura correta. A relação entre estrutura e função da proteína é complexa. Proteínas com estruturas semelhantes podem estar associadas a funções diferentes e vice-versa (Lesk, 2008). Este é mais um dos “mistérios” instigantes que a Bioinformática dispõe-se a decifrar: como a informação codificada numa sequência de aminoácidos ajuda a predizer qual a estrutura 3D de uma proteína? Muitos estudos buscam responder a esta pergunta, por exemplo: Dong et al., 2009; Kundrotasa et al., 2008; Lai et al., 2004; Rajgaria et al., 2010. Dentre estes estudos também está o de Dorn e Norberto de Souza (2008).

Dorn e Norberto de Souza (2008) propuseram um novo método para predição aproximada de estrutura tridimensional de proteínas, o CReF (*Central Residue Fragment-based method*). Os

resultados promissores que ele apresentou serviram como incentivo para a manutenção e a evolução do método CReF. Nesse sentido este trabalho implementou melhorias no método CReF buscando manter suas principais características e para aprimorar sua execução por meio da sua automatização e de um processo de refinamento.

Para atender a este objetivo o trabalho começa com a revisão de alguns conceitos importantes, como a Bioinformática Estrutural no capítulo 2 e, no capítulo 3, as proteínas e sua organização estrutural. O capítulo 4 traz uma revisão sobre o método CReF indicando o seu propósito e apresentando as nove etapas que o compõe. As melhorias técnicas e as aplicadas à metodologia que deram origem a uma nova versão da ferramenta são descritas no capítulo 5. Este capítulo também trata sobre o problema da otimização das regiões de voltas e alças em proteínas. Os experimentos realizados com a nova versão do CReF são apresentados no capítulo 6. Além disso, o capítulo apresenta o conjunto de proteínas teste considerado, o seu critério de escolha e os materiais e métodos utilizados no desenvolvimento desta dissertação. O desempenho da ferramenta é analisado e uma discussão sobre o refinamento das conformações iniciais geradas pela nova versão tem início também neste capítulo. Por último, no capítulo 7, são apresentadas as considerações finais e as principais contribuições desta dissertação, assim como, são indicados trabalhos que podem ser realizados futuramente com base neste estudo.

2. Bioinformática estrutural

A Biologia deu grandes passos apoiada por outras áreas, uma delas foi a Química. Essa conjunção deu origem à Bioquímica. A necessidade de explicar os fenômenos em nível atômico deu origem à Biofísica, e dessa mesma forma e unindo diferentes áreas surgiu a Bioinformática com especial contribuição da Ciência da Computação. O trabalho colaborativo entre as diferentes áreas traz benefícios para ambas as ciências (Cohen, 2004). Cohen em 2004, dizia que o trabalho com biólogos inspiraria os cientistas da computação a fazerem descobertas que melhorariam também a Ciência da Computação. Analisando o cenário atual pode-se dizer que ele estava certo. A Computação tem apoiado fortemente a Biologia e tornou-se um recurso essencial para a manutenção de diversos recursos, como acesso via web e programas para análises e processamentos, e também para a solução e a compreensão dos desafios. Áreas interdisciplinares como esta exigem esforços e disposição dos envolvidos para compreender o ponto de vista do colega de outra área, para que assim seja possível construir uma teoria ou uma solução para um problema. As habilidades essenciais a cada área precisam ser consideradas e adaptadas para se atingir o objetivo principal.

O crescimento e a abrangência dos projetos biológicos em todo o mundo criaram necessidades de análise e armazenamento de grandes volumes de dados e a disponibilização dessas informações para acesso público. Na era pós-genômica, a proteômica juntamente com o entendimento estrutural e funcional das proteínas tem sido o foco dos estudos (da Silveira, 2005). Esses estudos permitiram a separação, identificação e caracterização das proteínas (Dorn e Norberto de Souza, 2010; Lesk, 2008) e o próximo desafio da bioinformática é a predição da sua estrutura tridimensional (da Silveira, 2005). O aumento do volume de informações sobre estruturas tridimensional (3D) obtidas por meio de experimentos estimulou a criação de uma sub-disciplina na Bioinformática: a Bioinformática Estrutural.

Bioinformática Estrutural é a conceituação da biologia em termos de moléculas, no sentido físico-químico, e a aplicação de técnicas de informática (derivadas de disciplinas como: matemática, ciência da computação e estatística) para entender, organizar e explorar a informação estrutural associada a essas moléculas em grande escala (Luscombe et al., 2001). Em outros termos corresponde à intersecção de três grandes áreas: a Bioinformática, a Modelagem Molecular e a Biologia Molecular Estrutural. Seu principal foco é em representação, armazenamento, recuperação, análise e visualização da informação estrutural das proteínas (da Silveira, 2005).

Sabe-se que a função de uma proteína é, em grande parte, determinada pela sua estrutura 3D e não por sua sequência de aminoácidos apenas. A estrutura de uma proteína é peça de um interessante quebra-cabeça chamado proteoma. Quando uma proteína se dobra os resíduos importantes orientam-se em posições corretas para formar as regiões funcionais (Lesk, 2001). A maioria dessas regiões funcionais serve de ligação a outras moléculas. Proteínas pertencentes a uma mesma família e, mesmo não apresentando grande similaridade entre suas sequências, conservam a mesma estrutura 3D (Lesk, 2001). Muito mais importante do que o percentual de identidade entre duas sequências é a identidade de resíduos-chave, os quais são os principais responsáveis pela função da proteína. Por isso, a análise pura da sequência não pode ser levada em consideração a fim de evitar conclusões erradas sobre a função. Nesse sentido faz-se importante conhecer a estrutura tridimensional de uma proteína para determinar a sua função. Esse cenário apresenta outro desafio que é como predizer corretamente estruturas 3D para novas proteínas obtidas a partir de projetos genoma, por exemplo. Para isso é necessário compreender bem como se organizam as proteínas.

2.1 Resumo do capítulo

Neste capítulo foi apresentada a conceituação de Bioinformática Estrutural e o seu contexto nesta dissertação. Este conceito e aqueles envolvidos com proteínas e sua estruturação, a serem apresentados no capítulo seguinte, compõem o embasamento teórico deste trabalho.

3. Proteínas e sua organização estrutural

As proteínas são polímeros longos compostos pela combinação de aminoácidos e que apresentam um esqueleto repetitivo uniforme com uma cadeia específica para cada um dos 20 aminoácidos existentes. Mesmo sendo moléculas grandes, apenas uma pequena porção da estrutura das proteínas funciona de maneira precisa (o sítio ativo). O restante da estrutura tem o papel de criar e manter relações espaciais entre os resíduos do sítio ativo (Lesk, 2008).

A sequência de aminoácidos forma uma cadeia polipeptídica por meio de um processo de condensação e entre os aminoácidos ocorrem ligações peptídicas. Em ambiente nativo uma cadeia polipeptídica adota uma estrutura tridimensional (3D) única ou conformação nativa. Essa conformação determina a função da proteína na célula (Baxevanis e Ouellette, 2005; Branden e Tooze, 1998), dentre elas estão: proteínas estruturais, proteínas que catalisam reações químicas (enzimas), proteínas de transporte e de armazenamento (hemoglobina), proteínas reguladoras (hormônios), proteínas que controlam a transcrição gênica, proteínas envolvidas em reconhecimento (anticorpos), etc. (Lesk, 2008). O conhecimento da estrutura 3D implica no conhecimento da sua função e com este conhecimento é possível influenciar a ação da proteína no organismo por meio de fármacos ou drogas.

A maioria das cerca de 77.400 estruturas 3D conhecidas atualmente (conforme dados do PDB em novembro de 2011) foi determinada por cristalografia por difração de raios X ou por ressonância magnética nuclear. A partir dessas estruturas foi construído o conhecimento sobre as funções individuais e dos princípios gerais de estrutura e enovelamento de proteínas (Lesk, 2008). O enovelamento é o processo que faz com que a proteína atinja sua conformação no estado nativo. Os principais fatores desse processo são a sequência de aminoácidos e o seu ambiente.

Nessa seção pretende-se apresentar os principais conceitos relacionados a proteínas e sua organização estrutural.

3.1 Aminoácidos

As proteínas são formadas por unidades menores: os aminoácidos. Quimicamente um aminoácido compõe-se de um átomo de carbono central denominado $C\alpha$ que possui quatro ligantes: um grupamento amino ($-NH_2$), um grupamento carboxílico ($-COOH$), um átomo de hidrogênio (H)

e um grupamento orgânico R também chamado de cadeia lateral ou radical (Voet e Voet, 2006). Essa estrutura apresenta-se na Figura 1:

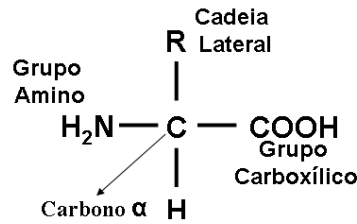


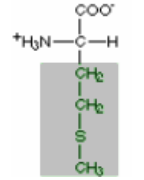
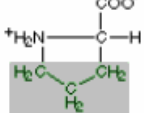
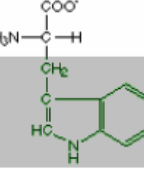
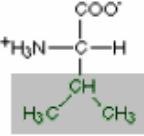
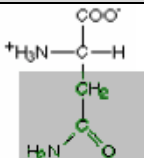
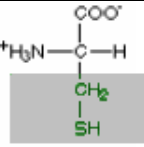
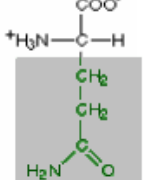
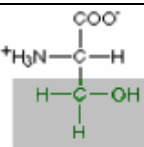
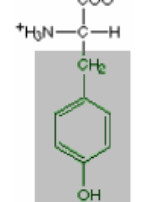
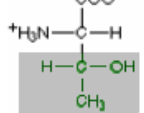
Figura 1 – Estrutura química de um aminoácido.

A cadeia lateral caracteriza as propriedades físico-químicas de cada um dos 20 diferentes aminoácidos. Por convenção internacional, os aminoácidos são identificados por abreviações de três letras (derivadas dos nomes em inglês) ou por um código de uma letra (Lehninger et al., 2002; Voet e Voet, 2006). A tabela a seguir apresentará as abreviaturas, a estrutura química e os ângulos da cadeia lateral de cada aminoácido:

Tabela 1 – Classificação e informações dos aminoácidos.

Aminoácido	Estrutura Química	Número ângulos χ da cadeia lateral
Apolares		
Alanina Ala A		χ_1
Fenilalanina Phe F		χ_1, χ_2
Glicina Gly G		Não possui cadeia lateral
Isoleucina Ile I		χ_1, χ_2
Leucina Leu L		χ_1, χ_2

Continua na próxima página

Aminoácido	Estrutura Química	Número ângulos χ da cadeia lateral
Metionina Met M		χ_1, χ_2, χ_3
Prolina Pro P		Cadeia principal
Triptofano Trp W		χ_1, χ_2
Valina Val V		χ_1
Polares		
Asparagina Asn N		χ_1, χ_2
Cisteína Cys C		χ_1
Glutamina Gln Q		χ_1, χ_2, χ_3
Serina Ser S		χ_1
Tirosina Tyr Y		χ_1, χ_2
Treonina Thr T		χ_1

Continua na próxima página

Polares Básicos		
Aminoácido	Estrutura Química	Número ângulos χ da cadeia lateral
Arginina Arg R		$\chi_1, \chi_2, \chi_3, \chi_4$
Histidina His H		χ_1, χ_2
Lisina Lys K		$\chi_1, \chi_2, \chi_3, \chi_4$
Polares Ácidos		
Ácido Aspártico Asp D		χ_1, χ_2
Ácido Glutâmico Glu E		χ_1, χ_2, χ_3

Segundo Lehninger (2002), a natureza química dos grupos R (cadeia lateral) ainda classifica os aminoácidos em quatro grupos:

- **Apolares:** a cadeia lateral desses aminoácidos não tem a capacidade de receber ou doar prótons, de participar de ligações iônicas ou de formar pontes de hidrogênio (Champe et al., 2006). Contudo apresenta uma propriedade oleosa que favorece as interações hidrofóbicas e contribuem na estabilização da estrutura e, por ser estritamente hidrofóbica e apolar, não se dissolve na água. Nas proteínas, as cadeias laterais apolares tendem a agrupar-se no seu interior. Isso é explicado pelo fenômeno da hidrofobicidade que faz com que esse grupo preencha o interior da proteína compondo sua forma tridimensional (Champe et al., 2006). Apesar de ser apolar, a glicina não contribui para a existência de interações hidrofóbicas. No caso da prolina, seu grupo amino secundário apresenta uma conformação rígida o que reduz sua flexibilidade estrutural.

- **Polares:** abrange aminoácidos não carregados, mas polares cujo tipo de cadeia lateral permite participar na formação de pontes de hidrogênio (Champe et al., 2006). Aminoácidos deste tipo são mais solúveis em água do que os não-polares, pois possuem grupos funcionais que formam ligações de hidrogênio com a água.
- **Polares Básicos:** as cadeias laterais básicas são aceptoras de prótons. As cadeias laterais da lisina e da arginina em pH fisiológico apresentam-se ionizadas positivamente. Já a histidina é fracamente básica e seu aminoácido livre não tem carga em pH fisiológico (Champe et al., 2006). Em contrapartida a isso, quando a histidina é incorporada a uma proteína sua cadeia lateral pode apresentar carga positiva ou neutra dependendo do ambiente iônico proporcionado pela cadeia polipeptídica (Champe et al., 2006).
- **Polares Ácidos:** são doadores de prótons e apresentam-se ionizados negativamente em pH neutro (Champe et al., 2006).

A compreensão das propriedades físico-químicas dos aminoácidos é importante, pois estas propriedades contribuem para que a proteína encontre estabilidade em seu estado nativo (Lehninger et al., 2002).

3.2 Ligação peptídica e ângulos de torção

Uma sequência linear de aminoácidos ligados contém a informação necessária para indicar a forma tridimensional de uma proteína. Essa ligação gera um polímero ou cadeia polipeptídica que é formado quando as proteínas se ligam de forma sequencial e covalente. Essa ligação, chamada de ligação peptídica, ocorre entre o átomo de carbono (C) de um aminoácido e o átomo de nitrogênio (N) de outro aminoácido (Voet e Voet, 2006). As ligações peptídicas não são rompidas facilmente, é necessária uma exposição prolongada a um ácido ou a uma base forte em elevadas temperaturas para que essas ligações sejam hidrolisadas de forma não-enzimática (Champe et al., 2006). Essa reação gera como co-produto água, formando-se a partir do $-OH$ da carboxila de um aminoácido e de um átomo H do grupo $-NH_2$ do outro aminoácido (Voet e Voet, 2006). A Figura 2 a seguir representa este processo:

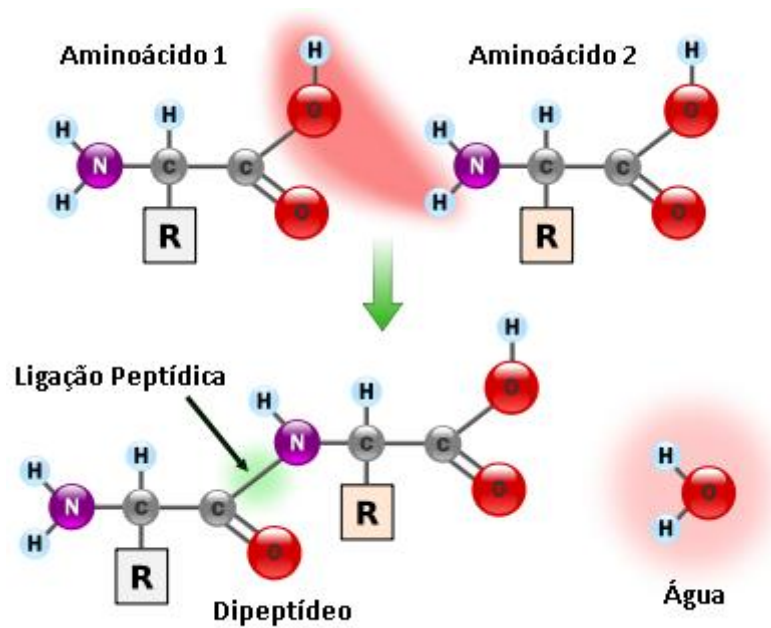


Figura 2 – Ligação peptídica: dois aminoácidos se ligam entre o átomo C da carbonila do primeiro e o átomo N da amina do segundo, ocorrendo uma desidratação que libera água e forma-se a ligação peptídica (figura adaptada de (Answers.com, 2011)).

Muitos aminoácidos associados por ligações peptídicas compõem uma cadeia não ramificada chamada de polipeptídeo, onde cada aminoácido é denominado de resíduo (Champe et al., 2006). Ao longo de uma cadeia apresenta-se um padrão que não se altera formado pelo conjunto $-N-C\alpha-C$, chamado de cadeia principal da proteína. A direção dessa cadeia é dada a partir do grupamento amina (N-terminal) até o grupamento carboxila terminal (C-terminal) (Voet e Voet, 2006). Essa convenção é usada para numerar os resíduos de uma sequência. Essa convenção facilita a análise de diversas características da sequência proteica.

Os grupos peptídicos, com poucas exceções, assumem uma configuração tal que os átomos $C\alpha$ sucessivos ficam em lados opostos da ligação peptídica que os une. Essa e outras observações indicam que o esqueleto de uma proteína compõe-se de uma sequência de grupos peptídicos planares rígidos e ligados (Voet e Voet, 2006). Assim, o enovelamento da proteína ou o enovelamento do esqueleto polipeptídico depende dos ângulos de torção que essa cadeia pode assumir. A rotação somente é permitida nas ligações simples de todos os resíduos: $N-C\alpha$ e $C\alpha-C$ (exceto prolina) (Lesk, 2008). O enovelamento de uma proteína é dado pelos ângulos ϕ (*phi*) e ψ (*psi*) dessas ligações e pelo ângulo ω (*ômega*) de rotação em torno da ligação peptídica (Lesk, 2008). Os valores dos ângulos ϕ e ψ podem variar de -180° a 180° . O ângulo diedro ω representa a rigidez e a planaridade de uma ligação peptídica e indica a rotação em torno da ligação entre o C da carboxila e o N da amina. Este ângulo não é livre para rotar, em virtude disso seus ângulos de torção variam próximos a 0° (*cis*) e a 180° (*trans*). Isso ocasiona uma restrição importante no

número de conformações que uma proteína pode adotar. Os maiores responsáveis pela torção da cadeia principal é o duplete (ϕ e ψ). Devido à maior liberdade de torção desses ângulos, pequenas mudanças podem resultar em alterações significantes na conformação (Dorn, 2008). Uma conformação totalmente estendida é obtida quando esses ângulos são fixados em 180° (Voet e Voet, 2006). Os ângulos ϕ , ψ e ω (triplete) da cadeia principal representam de forma única a conformação de uma proteína (Dorn, 2008).

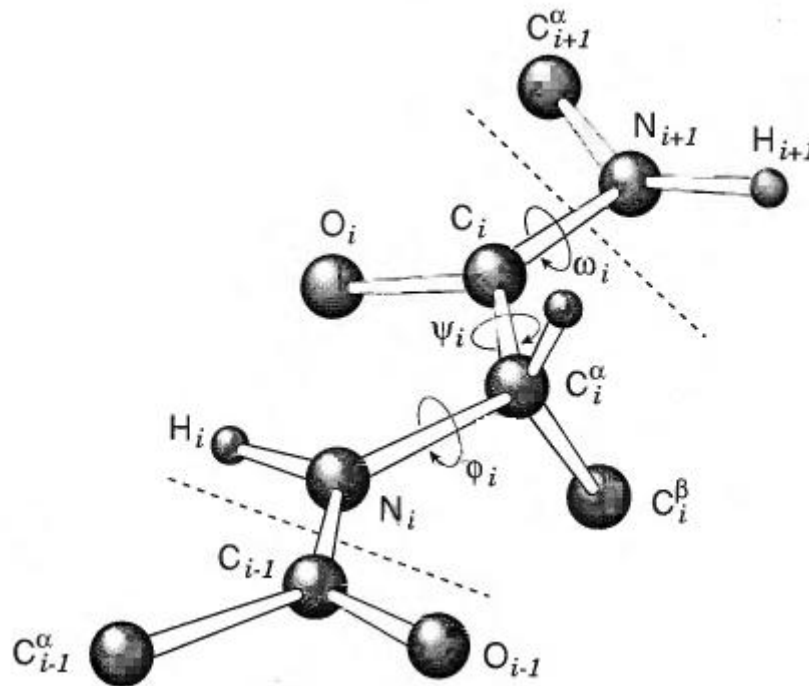


Figura 3 – Representação esquemática de um peptídeo identificando os ângulos de torção da cadeia principal ϕ_i , ψ_i e ω_i (Lesk, 2001).

A cadeia lateral também apresenta ângulos de torção que interferem na conformação da cadeia principal e auxiliam na estabilidade da molécula. A análise estatística do padrão conformacional dos ângulos de cadeias laterais com estruturas determinadas experimentalmente produziu bibliotecas de rotâmeros que indicam a conformação preferencial das cadeias laterais. As bibliotecas podem ser dependentes ou não da estrutura primária. Nas bibliotecas independentes, os ângulos χ (*qui*) são classificados sem considerar a estrutura regular (hélice ou folha) onde o aminoácido aparece e nas bibliotecas dependentes isso é considerado. A biblioteca de rotâmeros Dunbrack (Dunbrack e Karplus, 1993) é a mais utilizada. Essas bibliotecas são importantes na otimização de modelagens para indicar uma atribuição correta de ângulos de torção às cadeias laterais. Uma atribuição incorreta pode ocasionar choques estereoquímicos entre átomos da cadeia principal e da cadeia lateral e, conseqüentemente, a proteína assumirá uma conformação indevida

(Voet e Voet, 2006). A demonstração gráfica dos valores proibidos e permitidos para esses ângulos de torção é apresentada no Mapa de Ramachandran (Ramachandran e Sasisekharan, 1968).

Os ângulos de torção χ ocorrem em número diferente e dependem do tipo de resíduo de aminoácido. A tabela a seguir apresenta um resumo do que foi apresentado em uma das colunas da Tabela 1 sobre as cadeias laterais:

Tabela 2 – Número de ângulos χ presentes em cada aminoácido.

Resíduo	Número de ângulos χ
Ala, Gly, Pro	Cadeia principal
Cys, Ser, Thr, Val	χ_1
Asn, Asp, His, Ile, Leu, Phe, Trp, Tyr	χ_1, χ_2
Gln, Glu, Met	χ_1, χ_2, χ_3
Arg, Lys	$\chi_1, \chi_2, \chi_3, \chi_4$

3.3 O diagrama de Sasisekharan-Ramakrishnan-Ramachandran

A conformação de uma proteína pode ser descrita, quantitativamente, em termos dos ângulos internos de rotação em torno das ligações entre os átomos da cadeia principal (Figura 3). As conformações estericamente proibidas são aquelas onde qualquer distância interatômica entre átomos não-ligados é menor que a distância de van der Waals correspondente. Essas informações foram descritas pela primeira vez por V. Sasisekharan, C. Ramakrishnan e G. N. Ramachandran no que é hoje conhecido como mapa de Ramachandran (Voet e Voet, 2006 e Figura 4).

O mapa apresenta a variação possível dos ângulos ϕ (*phi*) e ψ (*psi*), de -180° a 180° . O ângulo de torção ω (ω), em torno da ligação peptídica, normalmente assume o valor de 180° (conformação *trans*) e, ocasionalmente (frequentemente, antes de um resíduo de prolina) $\omega = 0^\circ$ (conformação *cis*) (Lesk, 2001). As regiões do mapa que identificam as conformações permitidas dependem do raio de van der Waals escolhido para calculá-las (Voet e Voet, 2006). Em termos de enovelamento, as regiões do mapa representam padrões de torção da cadeia polipeptídica para elementos da estrutura secundária como folhas β e hélices α (Voet e Voet, 2006).

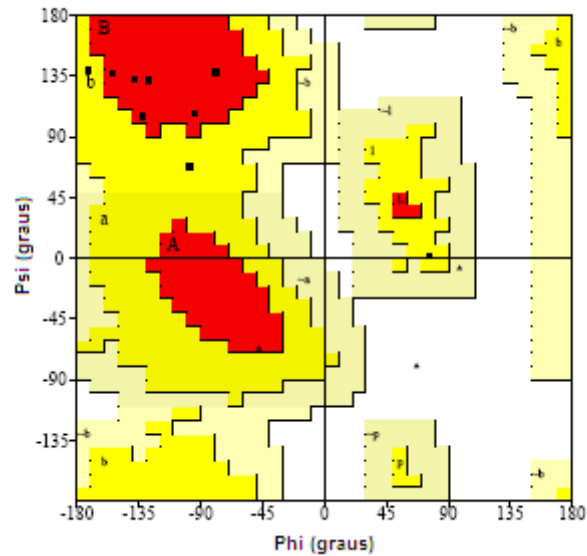


Figura 4 – Mapa de Ramachandran: região mais favorável em vermelho, região permitida em amarelo, região ainda aceitável em amarelo claro e região não permitida em branco. O canto superior em vermelho trata-se de região favorável para folhas β e no centro direito e esquerdo em vermelho para hélices α , respectivamente. Modelo adotado por Thornton e colaboradores (Laskowski et al., 1993).

A figura anterior foi produzida pelo programa PROCHECK (Laskowski et al., 1993), para definir regiões permitidas e proibidas ele usa a estrutura cristalográfica de 118 proteínas resolvidas numa resolução melhor que 2,00 Å. As regiões não permitidas podem ser ocupadas pela glicina, já que sua cadeia lateral tem apenas um átomo de hidrogênio e isso permite maior flexibilidade para os ângulos ϕ e ψ (Lesk, 2001).

As regiões do mapa de Ramachandran também estão associadas a conformações de resíduos, isso é especificado pela nomenclatura de A. V. Efimov (Efimov, 1993) conforme demonstrado na Figura 5:

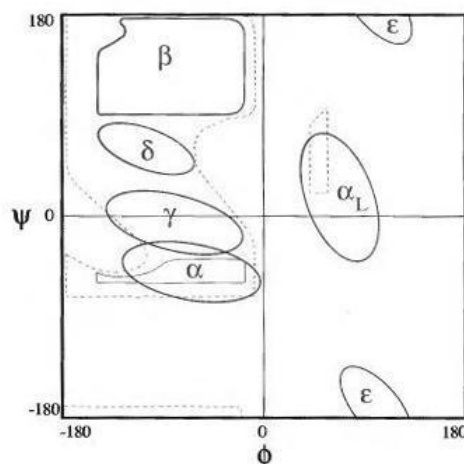


Figura 5 – Definições dos estados conformacionais no mapa de Ramachandran segundo A. V. Efimov (Lesk, 2001).

A nomenclatura indica regiões do mapa para as conformações:

- β : folhas β ;
- α : hélices α ;
- α_L : hélices α à esquerda;
- γ : volta γ ;
- ε : volta ε ;
- δ : volta δ ;

3.4 Hierarquia estrutural das proteínas

A alta complexidade da estrutura proteica foi melhor compreendida a partir da análise em quatro níveis hierárquicos: estrutura primária, estrutura secundária, estrutura terciária e estrutura quaternária. Essa análise mostrou que certos elementos repetem-se numa ampla variedade de proteínas, isso sugere a existência de regras que determinam como as proteínas se organizam (Champe et al., 2006). Esses elementos são combinados das formas mais simples até as mais complexas em domínios e, por isso, é importante compreender do que consistem estas estruturas.

Estrutura primária: é a própria sequência linear de aminoácidos ligados através de ligações peptídicas.

Estrutura secundária: são os padrões regulares nas estruturas das proteínas. Essa regularidade na conformação espacial mantém-se pelas ligações de hidrogênio entre os hidrogênios dos grupos amino e os oxigênios dos grupos carboxílicos de outros aminoácidos (Voet e Voet, 2006). Hélices α e folhas β são as conformações mais comuns. Além destas, há estruturas aleatórias com a função de conectar as estruturas secundárias regulares, como a volta e a alça (Voet e Voet, 2006).

A hélice α apresenta uma estrutura helicoidal, que consiste de um esqueleto polipeptídico em espiral no centro e com as cadeias laterais dos aminoácidos estendendo-se para fora do eixo central para evitar a interferência estérica entre si (Champe et al., 2006). A figura a seguir apresenta este esqueleto:

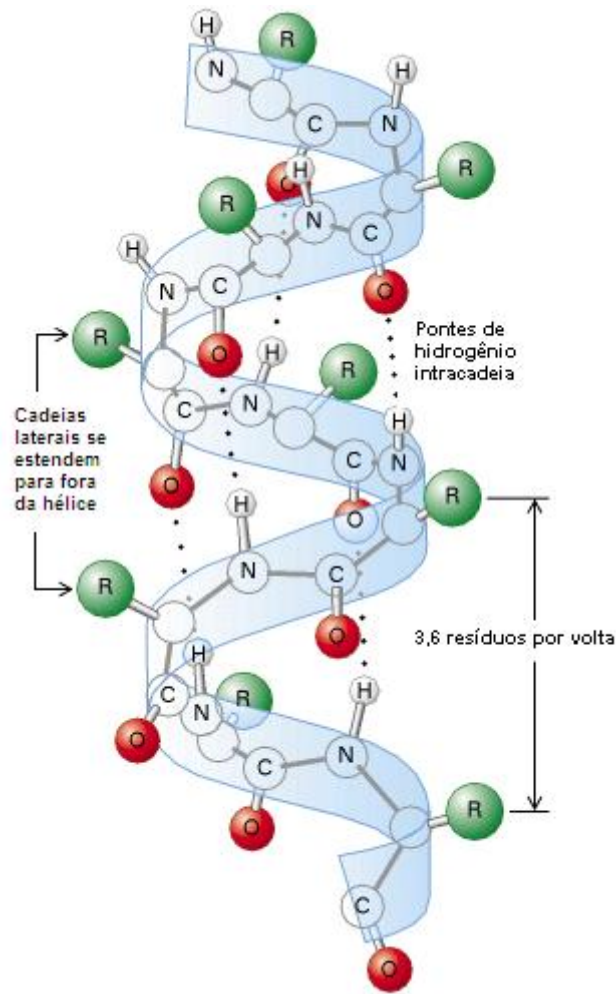


Figura 6 – Hélice α mostrando o esqueleto peptídico (figura adaptada de (Machine, 2011)).

Na Figura 6 anterior também se observam as pontes de hidrogênio entre os átomos de oxigênio e de hidrogênio que estabilizam a formação de uma hélice α (Champe et al., 2006). As pontes de hidrogênio estendem-se na espiral e são ligações individualmente fracas, mas em conjunto tem a função de estabilização (Champe et al., 2006). Em torno de 3,6 aminoácidos compõem uma volta completa de uma hélice, portanto, os resíduos separados por outros três ou quatro resíduos na sequência primária ficam espacialmente próximos quando dobrados na hélice (Champe et al., 2006). As ligações de hidrogênio podem acontecer entre o resíduo i e o resíduo $i + 4$, analisando este padrão de ligação verifica-se que são envolvidos 13 átomos da cadeia principal. Essa característica fixa da hélice deu origem a uma notação alternativa para sua representação: hélice $3,6_{13}$. Esta notação indica que na hélice α há 3,6 resíduos por volta e 13 átomos entre as ligações de hidrogênio (Lesk, 2008).

Existem ainda outras duas hélices possíveis, mas menos comuns que a hélice α . A hélice 3_{10} que tem três voltas e 10 átomos da cadeia principal entre as ligações de hidrogênio que envolve a carbonila do resíduo i com o nitrogênio do resíduo $i + 3$. E a hélice Pi , com a notação $4,4_{16}$, que

tem 4,4 resíduos por volta e 16 átomos entre as ligações de hidrogênio que envolve o resíduo i e o resíduo $i + 5$ (Lesk, 2001).

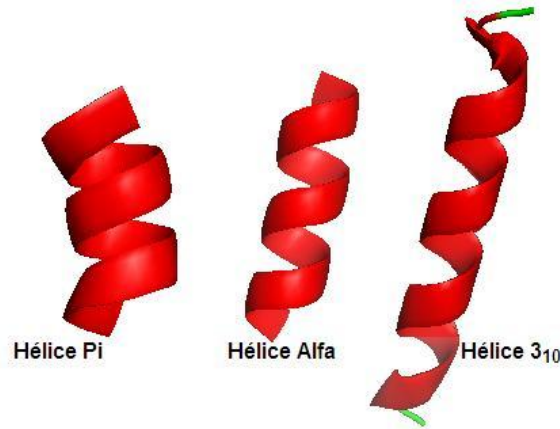


Figura 7 – Representação dos diferentes tipos de hélices. Neste desenho foi usada a representação CPK para os átomos e a hélice foi desenhada passando pelos átomos da cadeia principal. Figura gerada com o *software* PyMOL 1.2r1 (Delano, 2002).

A estrutura de uma hélice α pode ser comparada a uma mola em equilíbrio (hélice do centro na Figura 7), onde não são aplicadas forças. Ao se comprimir uma hélice α , há um encolhimento da mola, o que resulta numa hélice Pi (hélice à esquerda). Já o estiramento dessa mola, gera uma hélice 3₁₀ (hélice à direita) (Lesk, 2001). Essas estruturas também apresentam diferenças nos ângulos de torção:

Tabela 3 – Ângulos de torção para tipos de hélices (Lesk, 2001).

Tipo de Hélice	ϕ	ψ	ω
hélice α	-57,0	-47,0	180
hélice 3 ₁₀	-49,0	-26,0	180
hélice π	-57,1	-69,7	180

Propriedades diferentes influenciam diferentes sequências de aminoácidos a formar hélices α . Os resíduos de aminoácidos mais propícios a formar hélices são: alanina, glutamina, metionina, lisina e leucina. Enquanto isso, os resíduos prolina, isoleucina, glicina e serina dificilmente são encontrados em hélices α . O número total de aminoácidos em uma hélice α varia entre 5 e 40 resíduos, as hélices com 10 resíduos são as mais frequentes (Lesk, 2001).

Diferentemente da estrutura helicoidal das hélices, as folhas β compõem-se de duas ou mais cadeias peptídicas (fitas β) ou segmentos de cadeias polipeptídicas que se apresentam em

zigue-zague e arranjadas lado a lado assemelhando-se a uma série de fitas (Lehninger et al., 2002). A folha β também pode ser chamada de “folha β pregueada” devido a sua aparência. Numa folha β todos os componentes da ligação peptídica envolvem-se em pontes de hidrogênio (Champe et al., 2006). Suas cadeias adjacentes podem ser paralelas ou antiparalelas e, nesses casos, os padrões das ligações de hidrogênio são diferentes (Pauling e Corey, 1951).

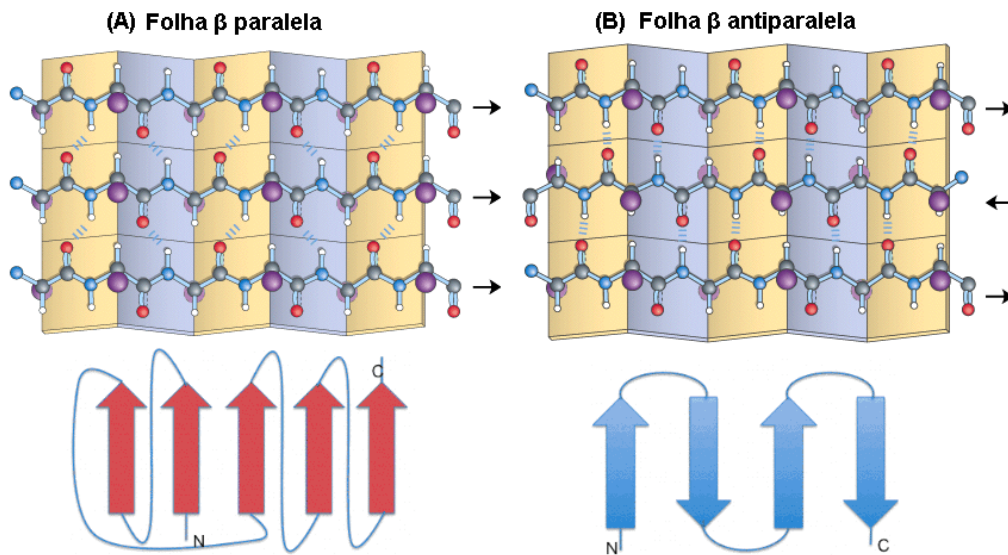


Figura 8 – Representação da folha β : em (A) folha paralela e em (B) folha antiparalela. A primeira representação mostra o padrão das ligações de hidrogênio em cada tipo de folha: perpendiculares em relação ao esqueleto na folha paralela e paralelas na folha antiparalela. A segunda representação usa vetores para representar as fitas que formam a folha (figura adaptada de (Al-Karadaghi, 2011) e (Biochemistry 462a, 2011)).

Nas proteínas globulares, as folhas β podem ser constituídas de duas a 15 fitas, sendo encontradas em média seis fitas. As cadeias das folhas β , tipicamente, apresentam até 15 resíduos de extensão e uma média de seis resíduos na cadeia. Observou-se que folhas paralelas com menos de cinco resíduos são raras, isso indica que as folhas antiparalelas são mais estáveis que as folhas paralelas (Voet e Voet, 2006). Esse fato deve justificar-se pelas ligações de hidrogênio que, quando comparadas à folha antiparalela, se mostram distorcidas na folha paralela. As folhas β mistas (paralela-antiparalela) são comuns, mas apenas 20% das fitas possuem ligação paralela de um lado e antiparalela do outro (considerando uma mistura aleatória, a expectativa seria de 50%) (Voet e Voet, 2006).

Quando não se encontra estruturas homólogas à sequência molde, podem ser utilizados métodos empíricos para prever estruturas secundárias. Certas sequências de aminoácidos limitam as conformações disponíveis para uma cadeia. Por exemplo, um resíduo Pro não pode se ajustar no

interior de uma hélice α ou de uma folha β comum, porque seus anéis pirolidínicos preencheriam um espaço ocupado por um segmento adjacente da cadeia e a falta do grupo N-H que contribui na ligação de hidrogênio compromete estabilidade da hélice (Voet e Voet, 2006).

Peter Chou e Gerald Fasman (Chou e Fasman, 1978) propuseram um método empírico de predição de estrutura que é bastante utilizado por ser de fácil execução (Voet e Voet, 2006). Esse método se baseia em algumas informações como a frequência (f_α) na qual um dado resíduo aparece em uma hélice α em um dado conjunto de estruturas:

$$f_\alpha = \frac{n_\alpha}{n} \quad (\text{Fórmula 1})$$

Onde n_α é o número de resíduos de um dado tipo que aparece em hélices α e n é o número total de resíduos deste tipo no conjunto de proteínas consideradas (Voet e Voet, 2006). A probabilidade de um dado resíduo aparecer em uma hélice α é definida por:

$$P_\alpha = \frac{f_\alpha}{\langle f_\alpha \rangle} \quad (\text{Fórmula 2})$$

Onde $\langle f_\alpha \rangle$ é o valor médio de f_α para todos os 20 aminoácidos. Um valor $P_\alpha > 1$ indica que um resíduo ocorre em hélice α em frequência maior do que em outras regiões da proteína. De forma semelhante é definida a probabilidade de um resíduo ocorrer em folhas β (P_β) (Voet e Voet, 2006).

Com base em 29 estruturas determinadas por raios x foi obtida a probabilidade de cada resíduo aparecer em hélices α ou em folhas β . Conforme essa probabilidade, cada resíduo recebe uma classificação (apresentada na Tabela 4) para estrutura secundária: altamente formadora (H), formadora (h), pouco formadora (I), indiferentemente formadora (i), não-formadora (b), ou altamente não-formadora (B) (Voet e Voet, 2006).

Com base nesses dados, Chou e Fasman desenvolveram o método Chou-Fasman para prever estruturas secundárias através de regras empíricas (Voet e Voet, 2006):

Um grupo de quatro resíduos formadores de hélice (H_α ou h_α , com I_α valendo meio h_α) em um conjunto de seis resíduos contíguos formará uma hélice. O segmento de hélice se propaga em ambas as direções da cadeia até que $P_\alpha < 1,00$ para um segmento tetrapeptídico.

Um grupo de três resíduos formadores de folhas β (H_β ou h_β) em um segmento de cinco resíduos contíguos formará uma folha. A folha também se propaga em ambas as direções até que o $P_\beta < 1,00$ para um segmento tetrapeptídico.

Para regiões que formem tanto α como β , a região de sobreposição é prevista como helicoidal se $P_\alpha > P_\beta$; do contrário, supõe-se a ocorrência de folha.

Essas regras predizem segmentos de hélice α e folhas β em proteínas com uma confiabilidade média de 50% e, em casos mais favoráveis, de 80%.

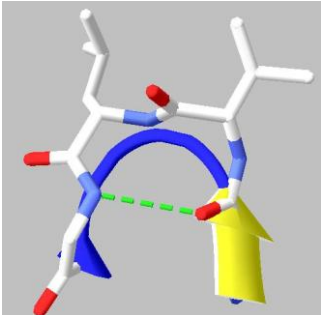
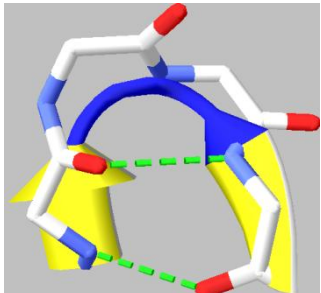
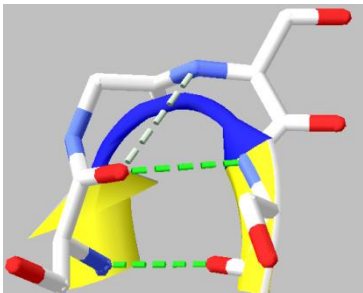
Tabela 4 – Tendência e classificação dos resíduos para hélices α ou folhas β (Chou e Fasman, 1978).

Resíduo	P_α	Classificação hélice	P_β	Classificação folhas
Ala	1,42	H_α	0,83	i_β
Arg	0,98	I_α	0,93	i_β
Asn	0,67	B_α	0,89	i_β
Asp	1,01	I_α	0,54	B_β
Cys	0,70	i_α	1,19	h_β
Gln	1,11	h_α	1,10	h_β
Glu	1,51	H_α	0,37	B_β
Gly	0,57	B_α	0,75	b_β
His	1,00	I_α	0,87	h_β
Ile	1,08	h_α	1,60	H_β
Leu	1,21	H_α	1,30	h_β
Lys	1,16	h_α	0,74	b_β
Met	1,45	H_α	1,05	h_β
Phe	1,13	h_α	1,38	h_β
Pro	0,57	B_α	0,55	B_β
Ser	0,77	i_α	0,75	b_β
Thr	0,83	i_α	1,19	h_β
Trp	1,08	h_α	1,37	h_β
Tyr	0,69	b_α	1,47	H_β
Val	1,06	h_α	1,70	H_β

Além de hélices e folhas, há as estruturas irregulares como as voltas e as alças, que são ditas espirais desorganizadas e conectam sucessivas estruturas secundárias regulares (hélice ou folha) (Voet e Voet, 2006). Uma proteína globular contém, aproximadamente, dois terços de resíduos em hélices e folhas e um terço em estruturas irregulares (Lesk, 2001). Voltas e alças tendem a ser mais flexíveis do que hélices e folhas nas mudanças conformacionais. No mapa de Ramachandran, as conformações em estruturas irregulares podem ocupar qualquer região, inclusive regiões de hélices α e folhas β . Em virtude disso, é difícil prever estas estruturas por meio de métodos computacionais (Dorn, 2008).

As voltas acontecem onde o polipeptídeo muda de direção, isto é, acontecem após uma estrutura secundária regular (Voet e Voet, 2006). A maioria das voltas constitui-se de quatro resíduos sucessivos e que se organizam de diferentes formas. O primeiro a identificar as voltas foi Venkatachalam (1968) que as classificou em três tipos: I, II e III. Quando ocorre uma conformação em imagem-espelho destes tipos obtêm-se os tipos I', II' e III'. O tipo I é o mais frequente e o tipo III é uma volta de hélice 3_{10} . Já I' e II' são mais raros, mas o tipo I' parece ser a preferida em grampos β (estrutura supersecundária que indica a união de folhas β) (Creighton, 1993).

Tabela 5 – Alguns exemplos de tipos de voltas conforme a nomenclatura de Venkatachalam (Venkatachalam, 1968).

Conformação	Código PDB	Tipo
<p>Região do mapa de Ramachandran: β-α-α-ϵ</p>  <p>Figura 9 – Volta I da Thermolysin: 12-15 (Gly-Val-Leu-Gly).</p>	3FV4	I
<p>Região do mapa de Ramachandran: β-α_L-α_L-β</p>  <p>Figura 10 – Volta I' da Actinidina: 172-175 (Gly-Gly-Glu-Val).</p>	2ACT	I'
<p>Região do mapa de Ramachandran: β-ϵ-γ-β</p>  <p>Figura 11 – Volta II' da Elastase: 36B-37 (Gly-Ser-Ser-Ser).</p>	3EST	II'

A conformação dos resíduos apresentados anteriormente segue a nomenclatura conforme A. V. Efimov. Nos mapas de Ramachandran apresentados a seguir é possível realizar uma comparação entre a “trajetória” das voltas dos tipos I’ e II’:

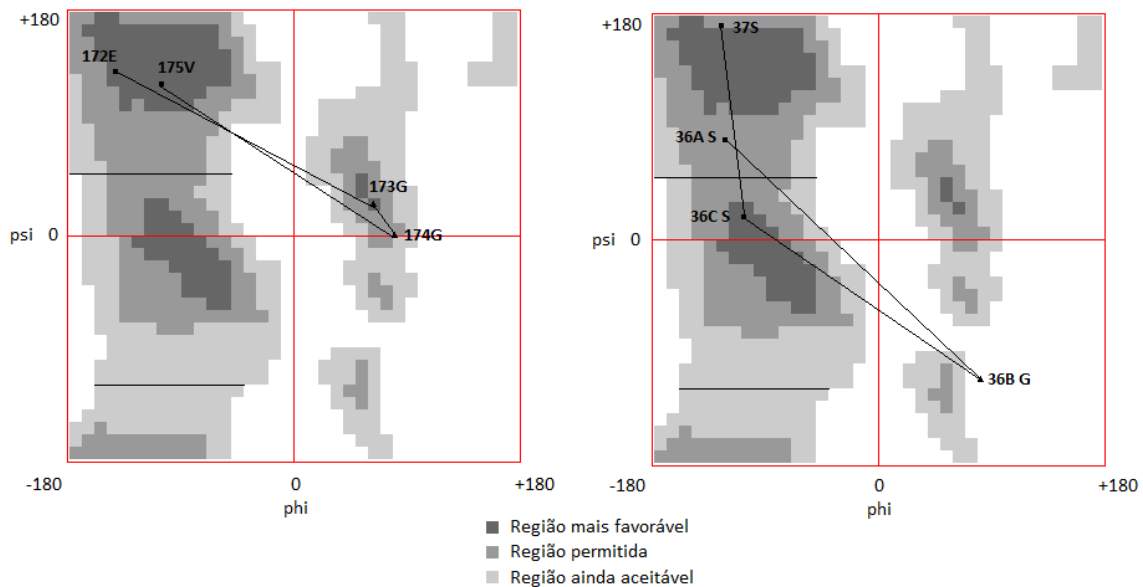


Figura 12 – Voltas da actinidina (código PDB: 2ACT - à esquerda) e da elastase (código PDB: 3EST - à direita) no mapa de Ramachandran. Mapas gerados por (Gopalakrishnan et al., 2010). Exemplos de proteínas obtidos de (Lesk, 2001).

As alças, que também podem ser reversas, estão presentes na maioria das proteínas com mais de 60 resíduos e aparecem uma ou mais vezes em uma quantidade de resíduos entre seis e 16 apresentando distâncias menores de 10,00 Å. Como elementos globulares compactos, pois suas cavidades internas são preenchidas por suas cadeias laterais, as alças, quase sempre, se localizam na superfície. Além disso, elas desempenham importante papel no reconhecimento biológico (Voet e Voet, 2006).

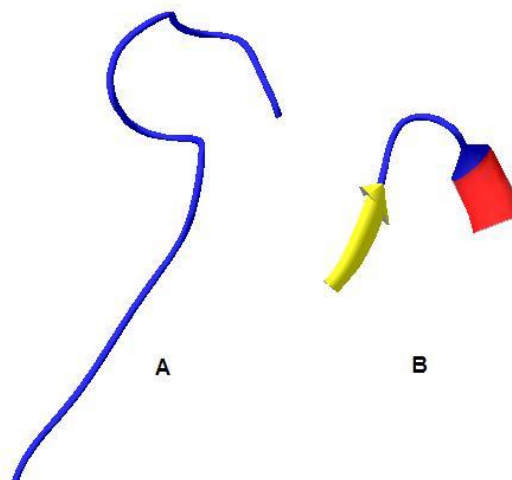


Figura 13 – Representação (A) das espirais desordenadas (*loop* ou *random coil*) e (B) da estrutura secundária irregular volta (*turn* em inglês). Essas estruturas pertencem a 2-*trans*-enoyl-ACP redutase de *Mycobacterium tuberculosis* (código PDB: 1ENY, para A S94 - A111 e B E62 - E68).

A combinação de elementos de estrutura secundária (hélices α e folhas β) pode gerar estruturas supersecundárias (motivos). Esse tipo de combinação costuma aparecer no interior da proteína e são conectadas por alças na superfície (Champe et al., 2006). O **motivo $\beta\alpha\beta$** é formado por uma sequência de: uma fita β , uma hélice α e outra fita β . Já o **grampo α** é formado por duas hélices α antiparalelas e sucessivas posicionadas uma contra a outra com os eixos inclinados para permitir a interação entre as cadeias laterais. Outra estrutura comum é o **grampo β** (β -Hairpin) que é formado por uma folha β antiparalela composta de segmentos sucessivos de uma cadeia por voltas reversas firmes formando um pequeno trecho de folha. Na **chave grega**, que recebe este nome pelo fato do desenho do motivo lembrar o desenho ornamental da Grécia antiga, um grampo β dobra sobre si para formar uma folha β antiparalela com quatro fitas (Voet e Voet, 2006).

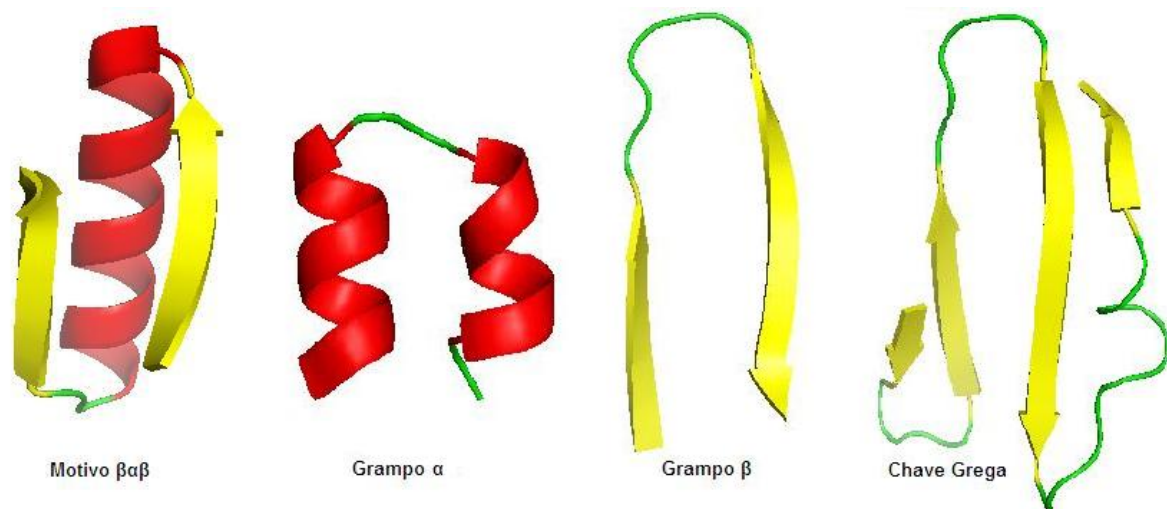


Figura 14 – Exemplos das quatro estruturas supersecundárias. As combinações de estruturas secundárias (alças e hélices) dão nome aos motivos.

Estrutura Terciária: é a representação da distribuição no espaço 3D dos arranjos de estruturas secundárias de uma proteína. Essa estrutura também chamada de estrutura funcional ou estrutura nativa da proteína deve sua conformação a: interações covalentes, ligações de hidrogênio, interações hidrofóbicas, interações hidrofílicas, interações eletrostáticas e forças de van der Waals (Gibas e Jambeck, 2001). A estrutura 3D assumida por uma proteína está diretamente relacionada à sua topologia (ou enovelamento). O tipo de sucessão de estruturas secundárias conectadas e a forma que se dispõem no espaço 3D é o que determina uma topologia. Com base nesta combinação de topologia e estrutura 3D é possível analisar a função da proteína no organismo (Dorn, 2008).

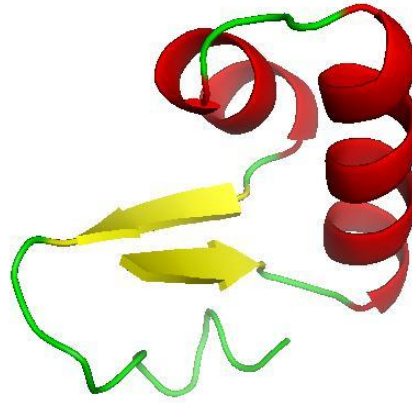


Figura 15 – Representação do tipo *Ribbons* da estrutura terciária da proteína Crambina (código PDB: 1CRN) composta por duas hélices α e duas estruturas de fitas β , sendo uma delas antiparalela e que estão conectadas por uma estrutura irregular do tipo volta (Murzin et al., 1995).

A interação das cadeias laterais dos aminoácidos determina a forma como a cadeia polipeptídica enovela-se, já que elas são atraídas ou repelidas conforme suas propriedades químicas. Por exemplo, cadeias laterais carregadas positiva ou negativamente atraem-se, já as cadeias com cargas semelhantes repelem-se. Além disso, interações do tipo hidrofóbicas e pontes de hidrogênio e dissulfeto influenciam o processo de dobramento (Champe et al., 2006). Nesse processo uma série de possibilidades é testada até que seja encontrado um estado no qual as atrações sobressaíam às repulsões, o que resulta numa proteína dobrada corretamente e com baixo estado energético (Champe et al., 2006).

Dentro da estrutura terciária é possível observar unidades funcionais fundamentais, os domínios. Cadeias com mais de 200 aminoácidos normalmente apresentam dois ou mais **domínios**. Um domínio é composto pela combinação de elementos supersecundários (motivos). O dobramento de um domínio é independente do dobramento de outro domínio e, conseqüentemente, apresenta características de proteínas pequenas e compactas que não possuem dependência estrutural de outros domínios (Champe et al., 2006). Entretanto, nem sempre a estrutura de um domínio é óbvia, podem ocorrer contatos extensos entre os domínios o que dificulta a identificação e fazem com que a proteína pareça ser uma única entidade globular (Voet e Voet, 2006). Analisando em mais detalhe os domínios nota-se uma composição de camadas com elementos da estrutura secundária. Isso se explica pelo fato de que são necessárias, ao menos, duas camadas dessas para proteger o núcleo hidrofóbico de um domínio em ambiente aquoso (Voet e Voet, 2006). Frequentemente, os domínios possuem funções específicas como a ligação para pequenas moléculas. Nas proteínas multidomínios, os sítios de ligação ocorrem nas fendas entre os domínios, isto significa dizer que as pequenas moléculas se ligam a dois domínios. Essa configuração é justificada pela necessidade de

uma interação flexível entre a proteína e a pequena molécula, e dessa forma é possível obter uma conexão maleável entre os domínios (Voet e Voet, 2006).

Ainda é possível que grupos de motivos se combinem sobrepondo-se ou não, para formar a estrutura 3D de um domínio, isso é denominado como **padrão de enovelamento**. Ao analisar o conjunto de proteínas de estruturas conhecidas verifica-se que, diferentemente do que se pensava, o número de padrões de enovelamento é pequeno. Dos 1.000 padrões diferentes de enovelamento que são previstos na natureza, 600 deles já foram observados (Voet e Voet, 2006).

Estrutura quaternária: refere-se ao arranjo de várias estruturas terciárias e mantido pelas mesmas forças que mantém as estruturas hierarquicamente inferiores. As subunidades que compõem esta estrutura podem funcionar de forma independente uma da outra ou de forma cooperativa, como no caso da hemoglobina, onde a ligação do oxigênio a uma subunidade aumenta a afinidade das outras subunidades ao oxigênio (Champe et al., 2006). Esse tipo de estrutura é comum, pois a construção com subunidades fornece a vantagem de reparação mais simples de erros pela substituição de subunidades, além de servir de base para a regulação de suas atividades (Voet e Voet, 2006).

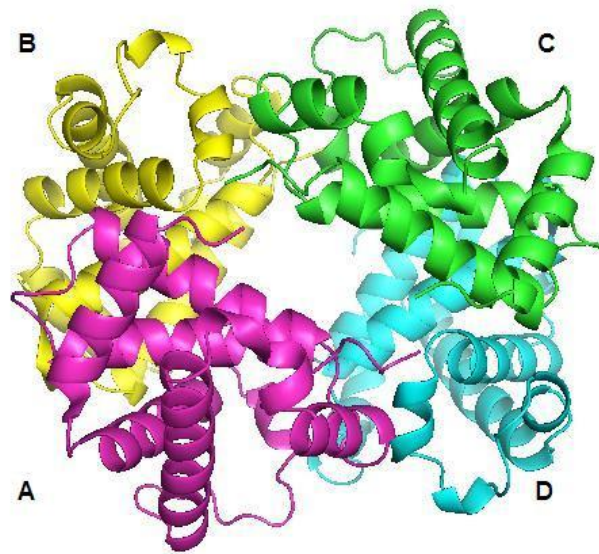


Figura 16 – Estrutura quaternária da hemoglobina (código PDB: 1A00), sem o grupo heme, em representação do tipo *Ribbons*, identificando as quatro cadeias: A em roxo, B em amarelo, C em verde e D em ciano. Cada subunidade é uma estrutura terciária.

3.5 Classificação Estrutural de Proteínas

As proteínas são agrupadas e classificadas conforme seus padrões de enovelamento (Lesk, 2008). Entre proteínas com padrões similares de enovelamento, há famílias com compartilhamento de características em estruturas, sequências e funções (Lesk, 2008). Com base nas estruturas secundárias e terciárias presentes nas proteínas, determinaram-se classes mais genéricas de estruturas:

Classe α : estrutura secundária somente com hélices α ou na sua maioria.

Classe β : estrutura secundária somente com folha β ou na sua maioria.

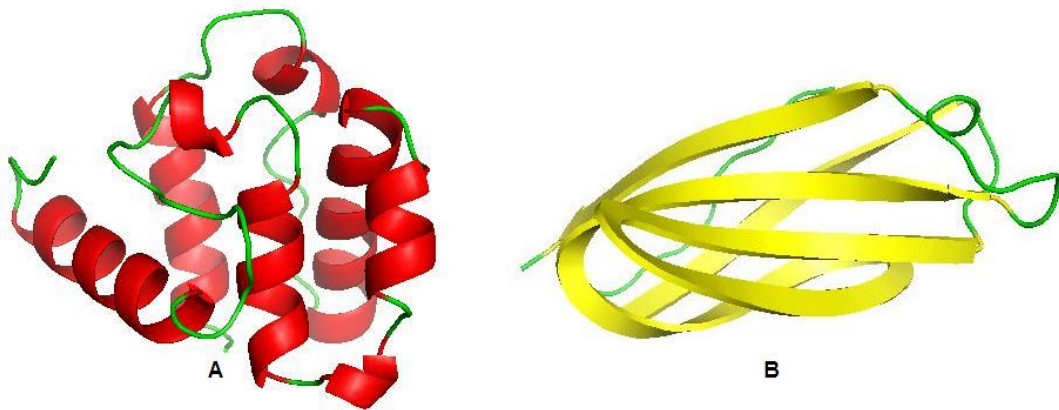


Figura 17 – (A) Proteína classe α : calponina da utrofina (código PDB: 1BHD), (B) Proteína classe β : módulo de adesão celular tipo III-10 da fibronectiva (código PDB: 1FNA).

Classe $\alpha + \beta$: hélices α e folhas β separadas e em partes diferentes da molécula, onde estrutura supersecundária β - α - β esteja ausente.

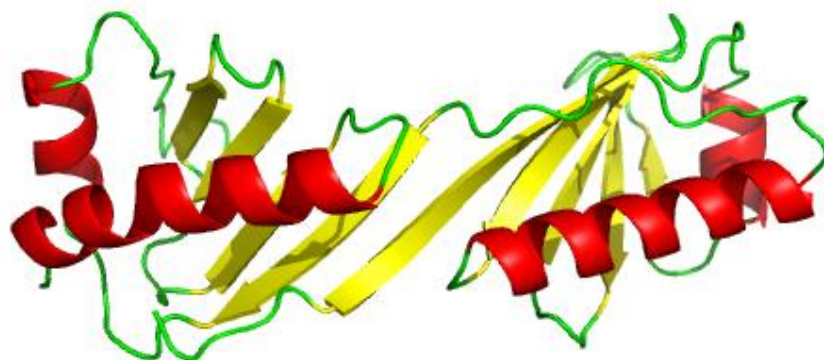


Figura 18 – Proteína classe $\alpha + \beta$: domínio principal da TBP (código PDB: 1CDW).

Classe α/β : hélices e folhas apresentadas a partir de unidades β - α - β . A figura a seguir apresenta três exemplos de proteínas desta classe, mas que possuem topologias diferentes:

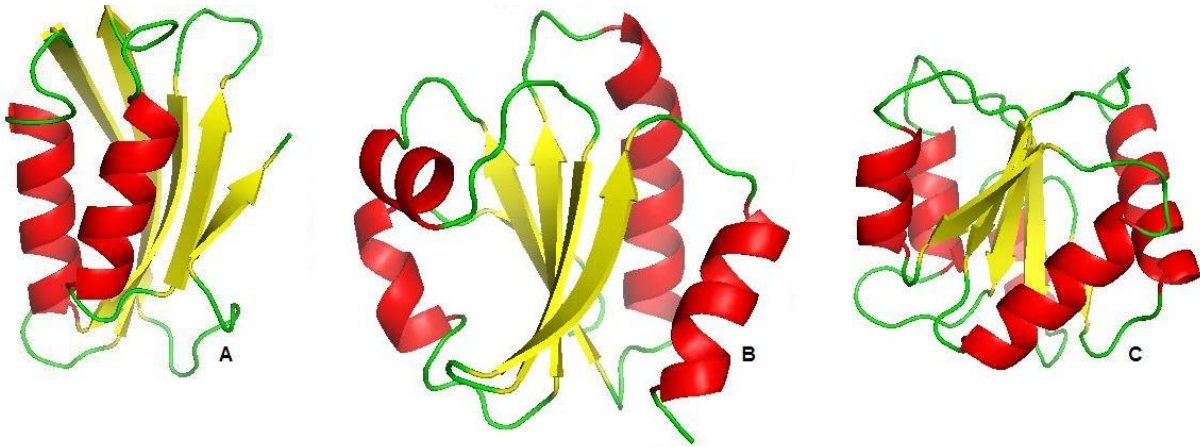


Figura 19 – Proteínas Classe α/β : (A) a acilfosfatase (código PDB: 2ACY); (B) a tioredoxina (código PDB: 1THX); (C) a proteína Chey (código PDB: 3CHY).

Classe α/β linear: uma linha passando pelos centros das folhas é aproximadamente linear. Um exemplo desta classe é a tioredoxina que está representada na Figura 19B acima.

Classe Barril α/β : uma linha passando pelos centros das folhas é aproximadamente circular.

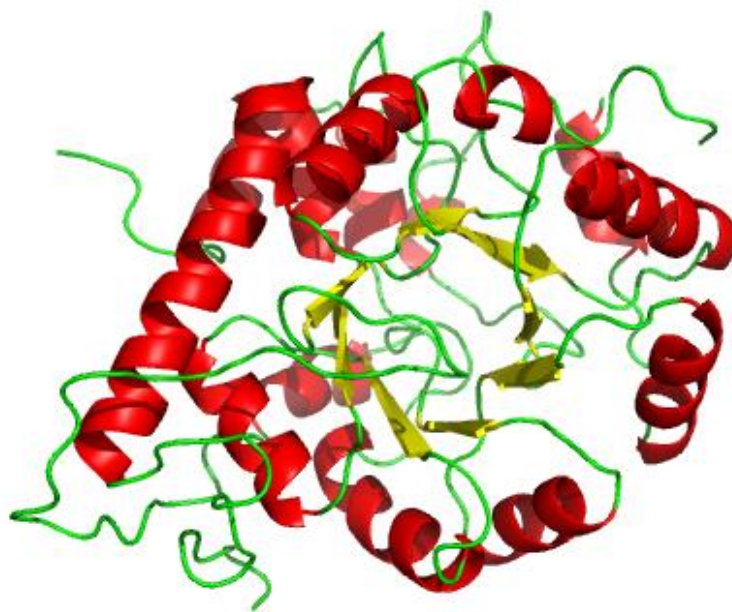


Figura 20 – Proteína classe barril α/β : Glicolato oxidase de espinafre (código PDB: 1GOX).

Com base na classificação de estruturas de proteínas construíram-se bancos de dados derivados do PDB que fornecem uma série de recursos interessantes para busca de informações estruturais (Lesk, 2008). Os principais bancos de dados são:

- **SCOP** (*Structural Classification of Proteins*) (Murzin et al., 1995): organiza a estrutura das proteínas conforme sua origem evolucionária e similaridade estrutural. No nível hierárquico mais baixo estão os domínios que são unidades compactas dentro do padrão de enovelamento que parecem possuir estabilidade independente. Nas famílias de homólogos estão agrupados conjuntos de domínios cujas similaridades implicam numa comum origem evolucionária. Já as superfamílias são formadas pelas famílias com estruturas de proteínas similares, mas cuja relação evolucionária não é conclusiva. Aquelas superfamílias com padrão de enovelamento (ou topologia) comum, ao menos para grande parte do núcleo da estrutura, formam as topologias (*fold*s). Cada grupo de topologia está associado a uma classe geral, dentre as principais classes estão: α , β , $\alpha + \beta$, α / β e uma mista para proteínas pequenas (Lesk, 2008).
- **CATH** (*Class, Architecture, Topology, Homologous superfamily*) (Orengo et al., 1997): usa uma hierarquia semelhante ao SCOP. Uma família sequencial compreende proteínas com estruturas muito similares, sequência e função biológica. As proteínas com clara relação evolucionária (conforme informação da identidade sequencial e da estrutura 3D) compõem a superfamília homóloga. Os conjuntos de superfamílias que compartilham arranjo espacial e conectividade de hélices ou fitas formam uma topologia. Já as proteínas com arranjos similares de hélices e fitas, mas com conectividade entre esses elementos de estrutura secundária diferentes geram uma arquitetura. O último nível desta hierarquia é a classe com divisão semelhante ao SCOP: α , β , α - β (α / β e $\alpha + \beta$ do SCOP) e domínios de pouca estrutura secundária.

3.6 Bancos de dados de estruturas de proteínas

Os bancos de dados desempenham papel essencial nos estudos das estruturas de proteínas, pois é a fonte básica de informações para a predição, análise e o estudo das estruturas. O mais conhecido e volumoso banco de dados de estruturas 3D atualmente é o *Protein Data Bank* (PDB)

(Berman et al., 2000) cujo propósito é armazenar e distribuir estruturas proteicas de forma organizada.

O crescimento dessa base de dados é impressionante. É possível notar esse crescimento pela análise dos dados dos últimos 20 anos apresentado na Figura 21. Em 1991 o PDB continha apenas 695 estruturas, em 2001 já eram 16.430 estruturas apresentando um crescimento de 2.364% em relação a 1991, e em 2011 já são 77.394 estruturas (dados de novembro de 2011) o que é um crescimento de 471% em relação a 1991. Esse crescimento exponencial do volume total de dados está representado em vermelho na Figura 21.

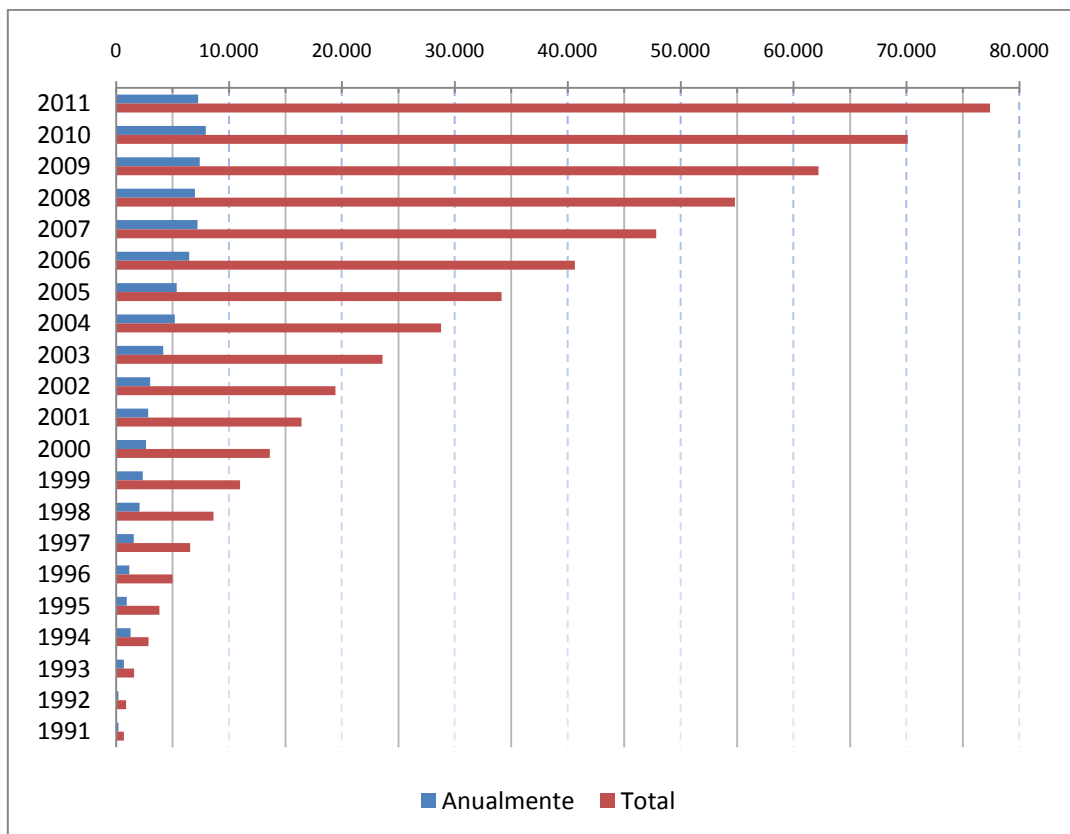


Figura 21 – Crescimento anual do número total de estruturas 3D de proteínas no PDB. Dados acessados em novembro de 2011.

Nessa figura também é possível analisar a evolução da quantidade de estruturas identificadas por ano (em azul). Em 1991 foram submetidas 187 novas estruturas, em 2001 houve um crescimento de 1.514% e esse número foi de 2.831, e em 2011 (até o mês de novembro) foram 7.267 novas estruturas o que representa um crescimento de aproximadamente 4.000% em relação ao número de 1991. Entre 1991 e 2007 (exceto 1995 e 1996 onde houve decréscimo) o número de estruturas submetidas por ano esteve em pleno crescimento estimulado pela variedade de grupos de pesquisas trabalhando na descoberta de estruturas 3D por meio de métodos experimentais e

apoiados, em alguns casos, pelos métodos de predição. O fato do número de estruturas disponíveis aumentar a cada ano e o PDB fornecer diversos recursos para consulta e recuperação de estruturas auxiliou o trabalho dos grupos de pesquisas na predição por modelagem por homologia e na realização de outros estudos. Nos últimos anos há uma tendência de estabilização da quantidade de submissões, em parte isso explica-se pela complexidade das proteínas a terem suas estruturas descobertas.

O PDB é gerenciado pelo *Research Collaboratory for Structural Bioinformatics* (RCSB), uma organização distribuída com base nos Estados Unidos. Sua *homepage* disponibiliza conexões para diferentes materiais: expositivo e tutorial, arquivos de dados, boletim de notícias sobre o banco de dados, recursos para depósito de novas proteínas e *software* para a análise de estruturas (Lesk, 2008).

Cada proteína armazenada no PDB possui um código identificador de quatro caracteres. O primeiro caractere desse código é um número de um a nove, sem significado mnemônico. Em vários casos, uma mesma proteína apresenta mais de uma entrada. Isso ocorre em virtude das diversidades ocorridas na determinação da proteína, seja pelo método cristalografia ou NMR, apresentando diferentes estados de ligação, usando diferentes formas cristalinas ou utilizando diferentes qualidades de cristal (Lesk, 2008).

Num arquivo de dados para uma dada proteína são disponibilizadas várias das informações armazenadas no PDB, tais como: identificador da proteína; espécie a que pertence à proteína; quem determinou a estrutura e referências das publicações relacionadas; detalhes experimentais incluindo a qualidade geral do resultado, resolução e estatísticas estereoquímicas; sequência de aminoácidos; moléculas adicionais, como co-fatores, inibidores, solventes, etc.; associação dos aminoácidos com elementos da estrutura secundária (2D); coordenadas atômicas; entre outros dados (Lesk, 2008).

Complementarmente aos dados disponibilizados no arquivo de dados, outras informações estão disponíveis para consulta através da *homepage* do PDB. Através de conexão com o PubMed, é possível ter acesso à publicação onde a proteína foi descrita, também estão disponíveis figuras da estrutura da proteína. De acordo com a classificação, listas de estruturas relacionadas são apresentadas, da mesma forma, as análises estereoquímicas, como ângulos de torção, podem ser consultadas, além de outras informações (Lesk, 2008).

Além do PDB existem outros bancos de dados que compartilham o mesmo propósito, são eles: TrEMBL (contém traduções dos genes encontrados em sequências de DNA do EMBL – *European Molecular Biology Laboratory* que é um banco de dados de ácidos nucleicos localizado no *European Bioinformatics Institute* na Europa) (Velankar et al., 2005) e PDBj – *Protein Data Bank Japan* (banco de dados de estruturas localizado no *National Institute of Genetics* no Japão). Recentemente estas instituições juntaram-se com o PDB para formar o *Worldwide Protein Data*

Bank (wwPDB) (Berman et al., 2006) que é uma iniciativa para produzir um arquivo unificado de dados (Lesk, 2008).

A variabilidade e o volume das informações sobre estruturas de proteínas contribuem diretamente no desenvolvimento e aprimoramento dos métodos de predição seja por homologia ou *ab initio*.

3.7 Resumo do capítulo

Neste capítulo foram apresentados conceitos importantes sobre proteínas e sua organização estrutural que compõem o embasamento teórico deste trabalho. Em mais detalhes foi tratado sobre aminoácidos e sua divisão em grupos, formação da ligação peptídica e ângulos de torção, mapa de Ramachandran com os padrões de torção da cadeia polipeptídica, organização das proteínas em hierarquia e a classificação estrutural das proteínas e, finalizando, com informações sobre bancos de dados de estruturas de proteínas.

4. Predição *ab initio* ou *de novo* da estrutura 3D de proteínas com o método CReF

Existem diferentes formas de se predizer a estrutura tridimensional de uma proteína partindo-se apenas da sua sequência de aminoácidos, também denominada de estrutura primária. A mais simples delas é a modelagem comparativa por homologia (Martí-Renom et al., 2000). Nela a sequência de uma proteína (sequência alvo) é alinhada com outras sequências de proteínas ortólogas com estruturas 3D conhecidas e armazenadas no PDB (estrutura molde e suas sequências). São ortólogas aquelas proteínas que são homólogas em espécies diferentes e que se originaram de um ancestral comum durante a especiação. Se a sequência alvo for bastante similar à sequência molde, isto é, caso as sequências sejam homólogas, utiliza-se essa estrutura conhecida para modelar a estrutura da sequência alvo (Martí-Renom et al., 2000). Já os métodos de reconhecimento de padrões de enovelamento (Jones et al., 1992) baseiam-se em potenciais estatísticos obtidos pela análise de enovelamento de proteínas com estrutura 3D conhecidas. Para uma dada sequência de aminoácidos é construída uma biblioteca de padrões de enovelamento. Se os fragmentos da proteína alvo se ajustarem a estas formas de enovelamento, pode-se alinhar a sequência alvo ao padrão de enovelamento mais adequado. Com as informações de uma proteína de estrutura conhecida é possível construir modelos que são ordenados conforme um escore que determinará o melhor modelo a ser utilizado. A predição *ab initio* ou *de novo* (Simons et al., 1999; Srinivasan e Rose, 1995) permite predizer novas formas de enovelamento, pois não se baseia somente em proteínas de estrutura conhecida. Como principais métodos deste tipo de predição têm-se: Rosetta (Rohl et al., 2004; Simons et al., 1999), I-TASSER (Wu e Zhang, 2007; Zhang, 2008), LINUS (Srinivasan e Rose, 1995; Srinivasan e Rose, 2002; Srinivasan et al., 2004) e FragFold (Jones, 2001). Uma revisão abrangente dos métodos atuais pode ser obtida dos anais do CASP9 (*Critical Assessment of Techniques for Protein Structure Prediction* - realizado em 2010) que será publicado em 2011. As principais limitações dessa metodologia são a complexidade e o tamanho do espaço conformacional acessível a proteínas. Mesmo que seja considerada uma pequena proteína, por exemplo, com menos de 100 aminoácidos, o problema da predição torna-se intratável computacionalmente, pois é impossível reproduzir o grande número de estados conformacionais acessíveis que esta proteína pode assumir (Paradoxo de Levinthal) (*apud* Karplus, 1997). A fim de reduzir este espaço conformacional são utilizados fragmentos da sequência alvo com estrutura 3D conhecida e estes são combinados de forma a se obter a proteína completa com a estrutura 3D de menor energia potencial a qual representaria o estado nativo da proteína. Este tipo de predição tem sido o mais explorado ultimamente.

Nesse sentido, Dorn & Norberto de Souza (2008, 2010) propuseram um novo algoritmo para prever a estrutura 3D aproximada de uma proteína ou polipeptídeo. O CReF (*Central Residue Fragment-based method*) tem o potencial de prever novas formas de enovelamento, pois não está limitado somente à informação de proteínas molde (Dorn e Norberto de Souza, 2008; Dorn e Norberto de Souza, 2010). Para isso faz uso de técnicas de mineração de dados para agrupar dados de estruturas determinadas experimentalmente, representação de intervalos de variação angular para representar uma conformação e de manipulação das informações estruturais. Esse método compreende princípios dos métodos de predição de novo para prever novas formas de enovelamento e da modelagem comparativa por homologia para garantir alta acurácia nas predições.

O primeiro passo para se definir um método de predição de estrutura 3D de proteínas é determinar a forma de representação da cadeia polipeptídica sendo uma das mais comuns a representação por meio de coordenadas cartesianas ou coordenadas internas, como os ângulos de torção da cadeia principal. O CReF optou pela representação da cadeia polipeptídica através dos ângulos de torção da cadeia principal e da cadeia lateral (ver seção 3.2). Uma vantagem importante desta representação é a redução significativa do número de variáveis para manipulação quando comparada à representação por coordenadas cartesianas. Isso permite uma maior eficiência computacional no tratamento das variáveis (Dorn, 2008).

Neste método a conformação C de uma proteína é representada por meio de um vetor na forma $C = \{x_1, x_2, \dots, x_n\}$. Cada um dos aminoácidos é representado neste vetor por um x_i que é formado por um tripleto de ângulos de torção ϕ (*phi*), ψ (*psi*) e ω (*ômega*). O conjunto desses tripletos consecutivos representa as rotações internas da cadeia principal da proteína. Para representar a cadeia principal de uma proteína, CReF considera somente o duplete de ângulos de torção ϕ , ψ , sendo os ângulos ω mantidos com valor de 180° (Dorn, 2008). Este último valor mantém o plano peptídico numa conformação *cis*, a mais comum em proteínas (Lesk, 2001). A escolha do duplete justifica-se pela liberdade de torção desses ângulos conforme explicado na seção 3.2.

O método CReF consiste de nove etapas conforme apresenta a figura a seguir:

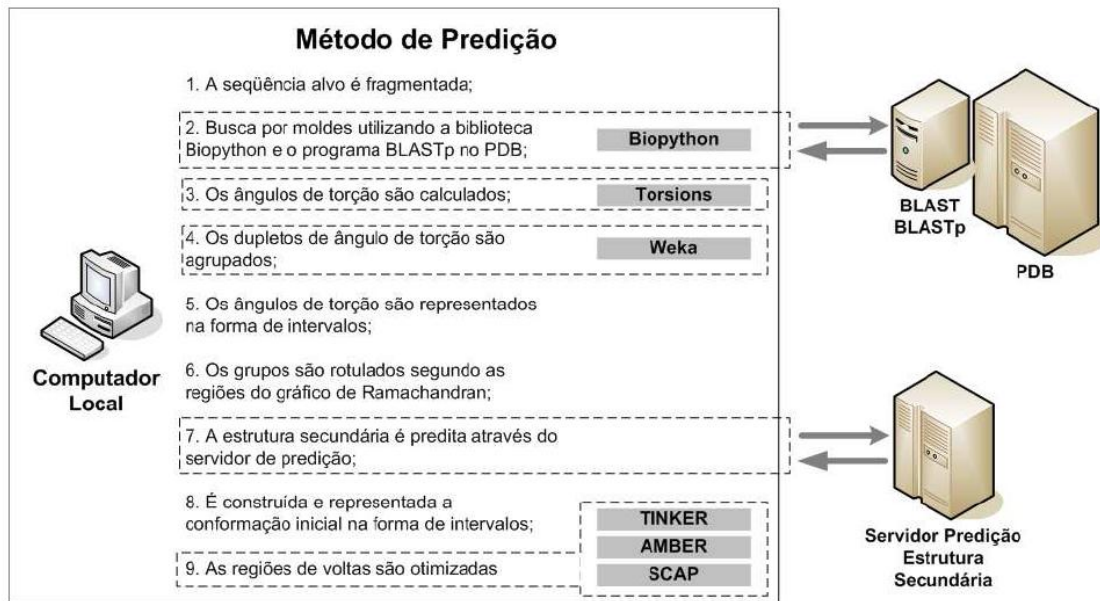


Figura 22 – Esquema com as nove etapas do método CReF. As etapas dois e sete são executadas remotamente e as demais localmente (Dorn, 2008).

As seções seguintes descreverão cada uma das nove etapas do método CReF de predição.

4.1 Etapa 1: fragmentação da seqüência alvo

Nesta primeira etapa, a seqüência alvo de uma proteína (K) é dividida em pequenos fragmentos contíguos e consecutivos (s_i), cada um com l ($l = 5$) resíduos de aminoácidos. O conjunto de todos os possíveis fragmentos com estas características é representado por $S = \{s_i, s_{i+1}, \dots, s_p\}$ (Dorn, 2008). Cada fragmento s_i começa pelo i -ésimo e termina com o j -ésimo resíduo de aminoácido formando um conjunto de tripletos consecutivos de ângulos de torção $\{(w_{i-1}, \varphi_i, \psi_i), \dots, (w_{j-1}, \varphi_j, \psi_j)\}$. Para cada fragmento o método considera a informação do seu resíduo de aminoácido central. (Dorn, 2008).

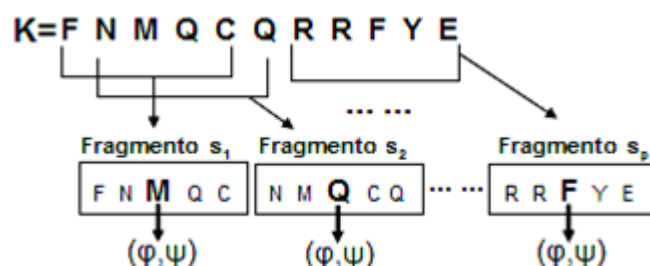


Figura 23 – Esquema representando uma seqüência alvo hipotética K dividida em p fragmentos s_i . Cada fragmento, por sua vez, é caracterizado pelo duplete, ou ângulos de torção φ e ψ , de seu resíduo central (Dorn e Norberto de Souza, 2008).

4.2 Etapa 2: busca por proteínas molde

Para cada fragmento s_i obtido na etapa anterior procura-se por fragmentos molde na base de dados do PDB. Essa busca é realizada com a versão *web* do programa BLASTp (Altschul et al., 1997) que permite a identificação de fragmentos homólogos (molde) ao fragmento submetido s_i (alvo). Somente são considerados os moldes com mesmo tamanho l do alvo (fragmento s_i) e que não possuam relação evolucionária com a sequência alvo K (30% ou menos de similaridade) (Dorn, 2008).

A execução do BLASTp usa a matriz de substituição BLOSUM62 (Henikoff e Henikoff, 1993). A busca ao PDB foi automatizada pela modificação da biblioteca BioPython (Chapman e Chang, 2000) que permitiu a conexão com o PDB e a análise automática de arquivos XML (*Extensible Markup Language*) resultante das consultas (Dorn, 2008).

O resultado desta etapa é uma lista de códigos PDB de moldes para cada um dos fragmentos alvo s_i de tamanho l . Para esses fragmentos alvo são recuperados seus arquivos PDB e armazenados localmente (Dorn, 2008).

4.3 Etapa 3: cálculo dos ângulos de torção dos dupletos

Para cada um dos arquivos PDB obtidos para cada fragmento s_i (etapa 2) são calculados os ângulos de torção do aminoácido central do fragmento molde. Estes dupletos são calculados com o programa *Torsions* (Grupo do Dr. Andrew C. R. Martim) e representados pela tupla $t_i = (\varphi, \psi)$.

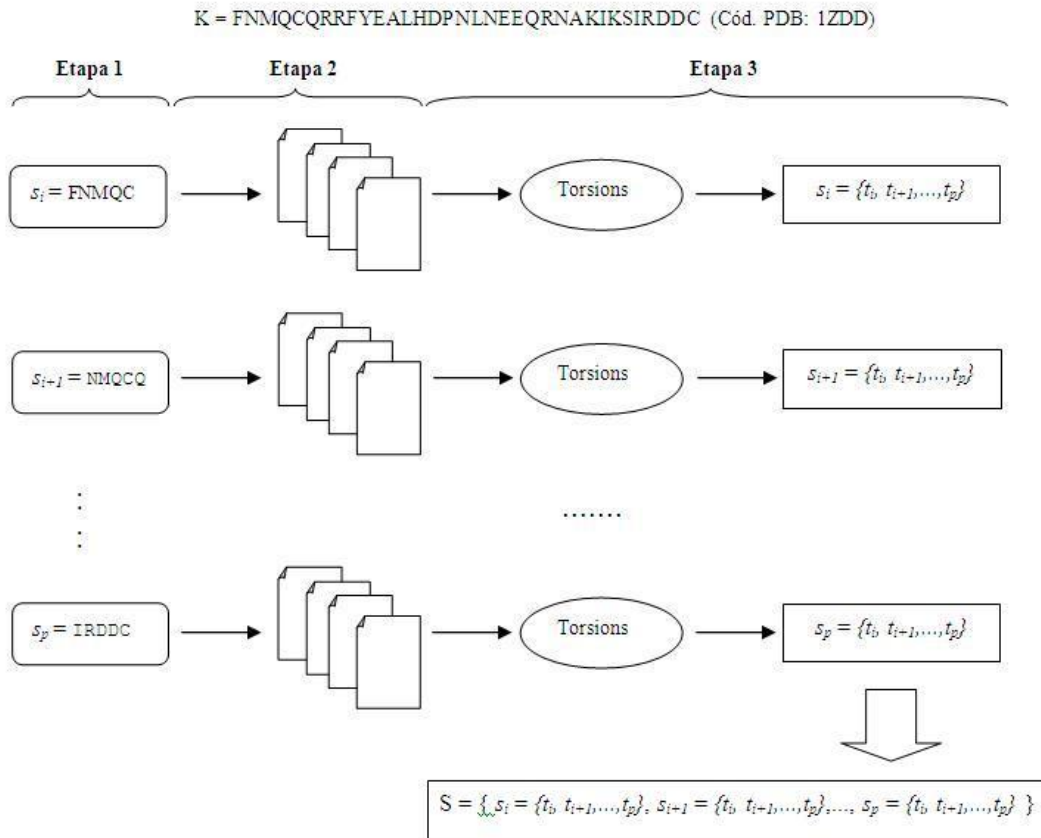


Figura 24 – Cálculo dos ângulos de torção de cada dupletto do aminoácido central da sequência alvo K .

4.4 Etapa 4: agrupamento de dupletos

O conjunto de tuplas de um fragmento é submetido a um processo de agrupamento que busca identificar as regiões onde as tuplas molde concentram-se no mapa de Ramachandran (ver seção 3.3). Esse agrupamento baseia-se no método probabilístico EM (*Expectation Maximization*) (Witten e Frank, 2005) que analisa as diferentes distribuições de probabilidades para cada grupo, buscando identificar o conjunto de grupos mais favoráveis em uma coleção de dados. O algoritmo EM não-supervisionado começa agrupando os dupletos com base no algoritmo k -médias (MacQueen, 1967) que minimiza uma função de erro quadrático (Fórmula 3) onde existem f grupos $k_i, j = 1, 2, \dots, f$ com base no ponto médio $m(k_i)$ de todos os dupletos, gerando uma solução inicial:

$$V = \sum_{i=1}^f \sum_{t_j \in k_i} |t_j - m(k_i)|^2 \quad (\text{Fórmula 3})$$

O valor médio $m(k_i)$ dos n dupletos t_j de um grupo $k_i, j = 1, 2, \dots, n$ é obtido por:

$$m(k_i) = \frac{1}{n} \sum_{j=1}^n \sum_{t_j \in k_i} t_j \quad (\text{Fórmula 4})$$

Em seguida, na fase *Expectation*, as probabilidades de cada grupo são calculadas para cada duplete t_j (Dorn, 2008). A maximização (*Maximization*) das distribuições é realizada a partir dos parâmetros de distribuição calculados a partir das probabilidades determinadas anteriormente (maior detalhamento do algoritmo EM em (Witten e Frank, 2005)).

Por influência das quatro regiões favoráveis do mapa de Ramachandran (Ramachandran e Sasisekharan, 1968) a fase de agrupamento busca encontrar quatro grupos ($f = 4$). Tendo sido identificados os f k_i grupos de um conjunto t_j de s_i calcula-se a média e desvio padrão estimado para cada ângulo (φ, ψ) em um grupo k_i .

No fim desta etapa obtém-se para cada fragmento s_i um conjunto de quatro grupos k_i descrito por 4-upla $s_i = \{k_i = (m\varphi, \sigma\varphi, m\psi, \sigma\psi), \dots, k_f = (m\varphi, \sigma\varphi, m\psi, \sigma\psi)\}$ onde $i = 1, 2, \dots, f$ em k, m representa a média do ângulo referenciado e σ , o desvio padrão do ângulo referenciado (Dorn, 2008). Esse processo de mineração de dados foi automatizado com o uso do pacote Weka (Witten e Frank, 2005). A média e o desvio padrão estimado obtidos para cada grupo $k_i \in s_i$ são utilizados para obter os intervalos de variação angular.

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

A representação das torções dos ângulos diedro na forma de intervalos de variação reduz drasticamente o espaço de busca conformacional. Estes intervalos são constituídos pelo valor médio $m(k_i, \theta)$ e pelo valor do desvio padrão $\sigma(k_i, \theta)$ calculados a partir dos dupletos (φ, ψ) de um grupo k_i . O intervalo de um ângulo diedro é representado por $[\theta] = [\underline{\theta}, \overline{\theta}]$, onde θ representa o intervalo para o ângulo φ ou ψ , $\underline{\theta}$ é o limite inferior e $\overline{\theta}$ é o limite superior do intervalo de ângulo de torção $[\theta]$.

São construídos intervalos para $[\varphi]$ e $[\psi]$ para cada grupo k_i de s_i . Com isso, cada fragmento passa a ser representado por grupos de intervalos de variação para φ e para ψ , $s_i = \{k_i = (\underline{\varphi}_i, \overline{\varphi}_i, \underline{\psi}_i, \overline{\psi}_i), \dots, k_f = (\underline{\varphi}, \overline{\varphi}, \underline{\psi}, \overline{\psi})\}$, onde $\underline{\varphi}$ e $\overline{\varphi}$ são, respectivamente, limite inferior e superior para φ e, $\underline{\psi}$ e $\overline{\psi}$ limite inferior e superior para ψ e f representa o número de grupos k (Dorn, 2008).

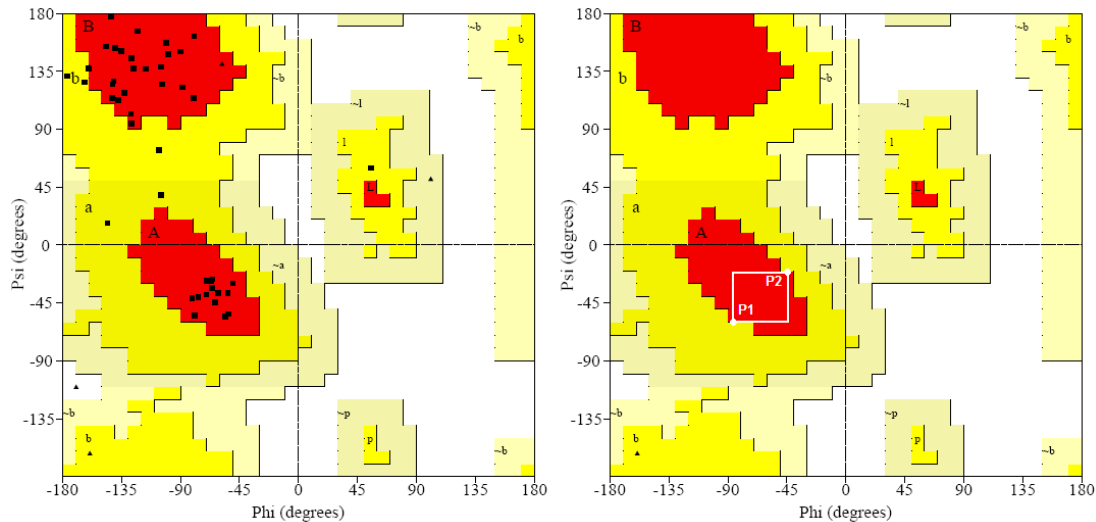


Figura 25 – Representação no mapa de Ramachandran: (A) tuplas ocupando regiões do mapa, em vermelho a região mais favorável, em amarelo a região permitida, em amarelo claro a região ainda aceitável e em branco a região não permitida. (B) representa a delimitação de um intervalo entre os ângulos de torção mínimo e máximo de φ e ψ , indicados pelos pontos P_1 e P_2 (Dorn, 2008).

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

Cada representação na forma de intervalo de variação angular dos grupos k_i de s_i são rotulados nesta etapa. O rótulo é criado a partir dos pontos médios $m(k_i, \varphi)$ e $m(k_i, \psi)$ de um grupo k_i e tem a função de relacionar o grupo à região do mapa de Ramachandran que ele ocupa. Esta classificação baseia-se nos trabalhos de Thornton e seus colaboradores (Morris et al., 1992; Laskowski et al., 1993).

Por simplificação, essas regiões favoráveis foram reduzidas para oito e foram identificadas por: A, B, L, a, b, l, p e c. As regiões $\sim a$, $\sim b$, $\sim l$, $\sim p$ e o restante da área não favorável do mapa passam a ser representadas, por c que é chamada de região de volta. Os rótulos gerados seguem o padrão: $k_i : rot$, onde rot é a identificação da região. Desta forma, cada fragmento s_i passa a ser representado por $s_i = \{k_1: rot, k_2: rot, k_3: rot, k_4: rot\}$. Após o rotulamento, os grupos k_i são ordenados de acordo com o número de elementos t_j molde de cada grupo, começando pelo grupo k_i com maior número de elementos (Dorn, 2008).

Tabela 6 – Regiões conformacionais do mapa de Ramachandran segundo Thornton e colaboradores.

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
~a	região aceitável	hélice- α
~b	região aceitável	folha- β
~l	região aceitável	hélice- α à esquerda
~p	região aceitável	extensão de hélice- α
restante da área	região não permitida	exceto para a Glicina

4.7 Etapa 7: predição da estrutura secundária

Nesta etapa é realizada a predição da estrutura secundária da sequência alvo K , por meio da determinação da região do mapa de Ramachandran que os ângulos de torção (ϕ e ψ) de cada aminoácido possivelmente estarão ocupando. O CReF utilizou um consenso obtido entre três métodos de predição (Dorn, 2008). A tabela a seguir apresenta um exemplo deste consenso usando os métodos DSC (King e Sternberg, 1996), PHD (Rost e Sander, 1993) e PREDATOR (Frishman e Frishman, 1996) para a cadeia A da proteína A estabilizada por ponte de sulfeto (código PDB: 1ZDD).

Tabela 7 – Exemplo de predição de estrutura secundária para o código PDB: 1ZDD.

Método	FNMQCQRRFYREALHDPNLNEEQRNAKIKSIRDCC
DSC	Ccchhhhhhhhhhhcccchhhhhhhhhhhccccc
PHD	Ccchhhhhhhhhhhcccchhhchhhhhhhccc
PREDATOR	Ccchhhhhhhhhhhcccchhhhhhhhhhhccc
Consenso	Ccchhhhhhhhhhhcccchhhhhhhhhhhccc

A informação da predição da estrutura secundária obtida nesta etapa é utilizada na escolha dos intervalos de variação angular para construção da conformação inicial da proteína alvo na etapa seguinte.

4.8 Etapa 8: construção da conformação inicial

A construção da conformação inicial baseia-se na informação dos grupos k_i de cada fragmento s_i . Para isso é necessário escolher um grupo de s_i para representar o i -ésimo aminoácido de uma sequência K (Dorn, 2008). Essa escolha é guiada pela predição da estrutura secundária e deve respeitar duas regras:

Regra 1: o(s) grupo(s) k_i deve ter *rot* igual ao rótulo identificado para o i -ésimo aminoácido no consenso da predição da estrutura secundária, além de ser o grupo mais significativo com o rótulo buscado.

Regra 2: se na predição da estrutura secundária for predito um estado conformacional de estrutura regular (h ou e) para o i -ésimo resíduo e não existir um grupo k_i com rótulo *rot* igual ao identificado para este resíduo procede-se calculando o valor médio entre os ângulos presentes nos resíduos $i - 1$ e $i + 1$ e o i -ésimo resíduo. O valor obtido é substituído na sequência alvo K no i -ésimo resíduo (Dorn, 2008). O processo de construção da conformação inicial representada como intervalos é apresentado no esquema a seguir:

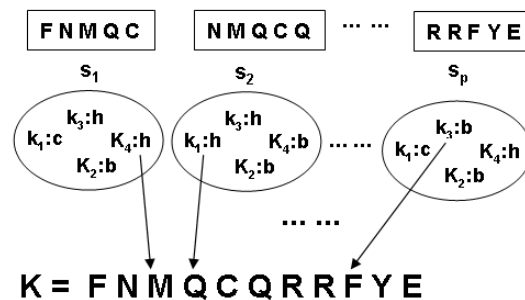


Figura 26 – Esquema da escolha dos grupos k_i que representam os ângulos de torção dos aminoácidos de uma sequência alvo K (Dorn, 2008).

4.9 Etapa 9: otimização das regiões de volta

As regiões de volta possuem grande influência na determinação da forma de enovelamento das proteínas (Fiser et al., 2000). Além disso, as estruturas irregulares (voltas e alças) são os tipos de estruturas secundárias mais difíceis de serem previstas, pois elas podem aparecer em qualquer área do mapa de Ramachandran. Em virtude disso, o método optou por otimizar, pela redução do intervalo da conformação inicial, somente as regiões de volta identificadas na predição da estrutura secundária (Dorn, 2008). Essa otimização acontece conforme o fluxograma da Figura 27.

4.10 Resumo do capítulo

Este capítulo apresentou o método CReF de predição de estrutura 3D aproximada de proteínas desenvolvido por Dorn & Norberto de Souza (2008, 2010). As nove etapas do CReF (fragmentação, busca por proteínas molde, cálculo dos ângulos de torção, agrupamento de dupletos, representação na forma de intervalos, classificação dos grupos no mapa de Ramachandran, predição da estrutura secundária, construção da conformação inicial e otimização das regiões de volta) foram descritas de forma objetiva juntamente com os experimentos realizados e os resultados obtidos pela versão original do CReF. Os bons resultados alcançados foram o embasamento para o trabalho desenvolvido.

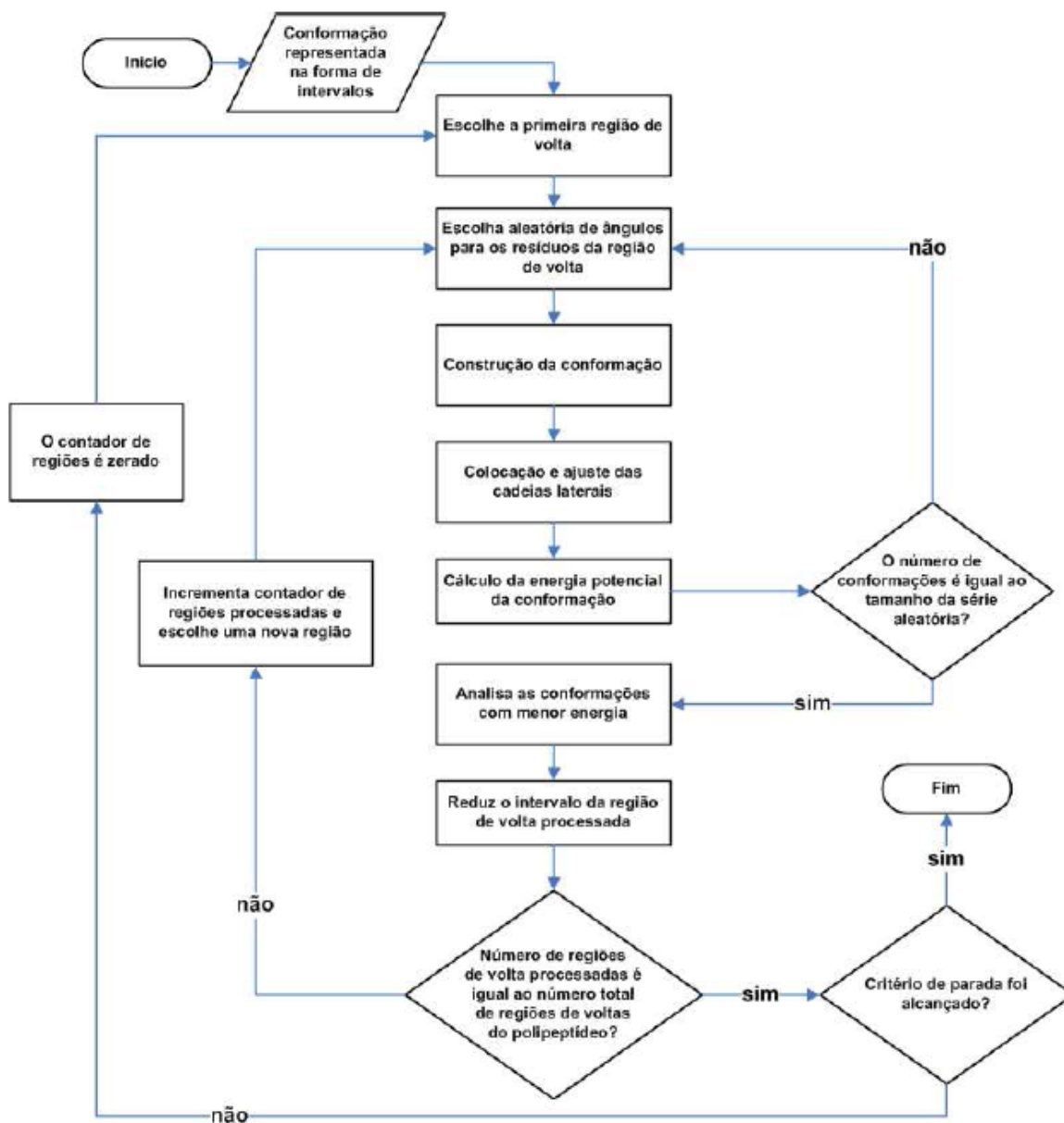


Figura 27 – Fluxograma do método de otimização de voltas de polipeptídeos representados na forma de intervalos de variação angular do método CReF (Dorn, 2008).

5. Desenvolvimento de nova versão do CReF

Os resultados promissores apresentados pelo CReF em sua versão inicial fizeram com que ele se tornasse alvo de novos estudos com o objetivo de aprimorá-lo. A análise dos seus resultados iniciais permitiu identificar seus principais pontos positivos:

- Predição satisfatória de estruturas secundárias.
- Compatibilidade com os resultados de outros métodos (em valores de RMSD) para proteínas com até 200 resíduos.
- Rapidez na execução em plataforma de baixo custo.

Em contrapartida também foram identificados alguns pontos negativos, sendo o principal deles a má formação das estruturas irregulares (voltas e alças) e, conseqüentemente, má organização das estruturas secundárias no espaço 3D.

Buscando aprimorar o CReF e mantendo seus principais diferenciais, gerou-se uma nova versão do método cuja ênfase foi na usabilidade. O conceito de usabilidade surgiu na ciência da computação dentro da engenharia de software como consequência da preocupação em fornecer aos usuários ferramentas fáceis de serem usadas. Atualmente esse conceito foi ampliado para diversas áreas, pois se um produto ou ferramenta é fácil de usar, o usuário tem maior produtividade, aprende mais rápido, memoriza operações e comete menos erros (Pressman, 1995). Acreditando nestes preceitos, buscou-se tornar o CReF uma ferramenta de uso simples e que permite uma interação mais clara e objetiva com o usuário.

A primeira etapa deste trabalho foi simular os resultados iniciais obtidos. Em virtude da falta de um histórico mais detalhado, especialmente sobre as proteínas molde consideradas, não foi possível realizar esta simulação. Após as adaptações necessárias, algumas proteínas foram submetidas, o que permitiu um entendimento mais profundo da ferramenta. O maior entendimento do método possibilitou a identificação de melhorias que poderiam refletir na usabilidade e no resultado da conformação inicial. Esse capítulo apresentará o conjunto de melhorias que foram implementadas.

5.1 Melhorias implementadas no CReF

As melhorias implementadas na nova versão do CReF foram divididas em duas categorias: melhorias técnicas e alterações no método. As melhorias técnicas descrevem mudanças que contribuíram para uma maior eficiência e facilidade na execução da ferramenta. As alterações no método tiveram como objetivo garantir a clareza dos conceitos aplicados, permitir a comunicação adequada entre as etapas e eliminar a subjetividade, aspectos que impossibilitavam a automatização do método e que poderiam ampliar sua confiabilidade.

5.1.1 Melhorias técnicas

Para que o CReF pudesse ser executado após o trabalho de Dorn (2008), algumas adaptações foram necessárias. A etapa de busca por proteínas molde precisou ser readequada aos novos parâmetros do BLASTp (Altschul et al., 1997). Para isso passou a ser usada a biblioteca NCBIWWW na execução da consulta. O ajuste de alguns parâmetros passados ao BLASTp permitiu que as consultas voltassem a ser realizadas com sucesso. Dentre estas alterações está o uso da matriz PAM30 (*Point Accepted Mutation*) (Dayhoff et al., 1978), ao invés da BLOSUM62. Foram realizados testes através da busca Web com a BLOSUM62, mas a busca só retornava resultados quando o indicador “*Short queries*” estava marcado. A marcação deste indicador autoriza o BLASTp a ajustar, automaticamente, os parâmetros de busca para sequências pequenas (com menos de 20 aminoácidos), mas como esta opção não está disponível na biblioteca utilizada, foram necessários os ajustes nos parâmetros e a mudança na matriz de substituição.

Na versão original do CReF, a Etapa 1 de fragmentação da sequência era realizada de forma manual. Para automatizar esta etapa foi criada uma função de fragmentação.

Na etapa de construção da conformação inicial houve problemas com a função `teLeap` do pacote AMBER 9 (Case et al., 2006). Para resolver o problema e manter a compatibilidade do processo passou-se a usar a função `tleap`. Após a realização destas alterações e algumas outras (como: adaptação do tamanho de variáveis, atribuição de índice a variáveis para evitar problemas na concatenação de dados, troca da posição referenciada em vetor, etc.) foi possível criar uma nova versão estável que serviu de base para as análises e realização das demais melhorias.

A execução do CReF baseia-se na leitura de arquivos e na geração de outros que servem de entrada para etapas posteriores. Para esta forma de execução é importante saber rapidamente quais são as entradas e as saídas de cada função para facilitar a execução a partir de uma dada etapa. Neste sentido desenvolveu-se uma documentação que indicava a ordem de execução das funções e,

para cada uma, apresentavam-se algumas informações como: arquivos lidos, arquivos resultantes da sua execução, software necessário, as alterações realizadas e indicava a etapa do CReF à qual a função estava relacionada.

A versão inicial do CReF apresentava uma grande dificuldade de usabilidade, o que tornava onerosa e trabalhosa a realização repetitiva de novas execuções. Devido a inúmeras referências diretas dentro do código aos diretórios onde se localizavam os arquivos com as informações, não era possível adotar uma estrutura diferenciada de diretórios e, desta forma, a preparação do ambiente para testes com outras proteínas alvo precisava ser feita manualmente. Essa preparação do ambiente teve uma melhora inicial através da criação de um script responsável pela realização de uma cópia da última execução e preparação para uma nova execução. O novo script, assim como os demais scripts do método, precisava ser invocado manualmente pelo usuário da ferramenta. Essa condição manual de execução dos scripts tornava o uso da ferramenta CReF complexo. A forma como o CReF foi implementado atribuiu ao usuário toda a responsabilidade pelo funcionamento do método, isto é, sob a responsabilidade do usuário estavam: criação dos arquivos iniciais, garantia da criação dos diretórios necessários, execução dos scripts na ordem correta, etc. Desta forma, qualquer falha do usuário poderia causar erro ou causar alguma distorção nos resultados obtidos. Com o objetivo de facilitar o uso da ferramenta CReF e buscando retirar do usuário essa responsabilidade, decidiu-se implementar a automatização de todas as tarefas possíveis do método.

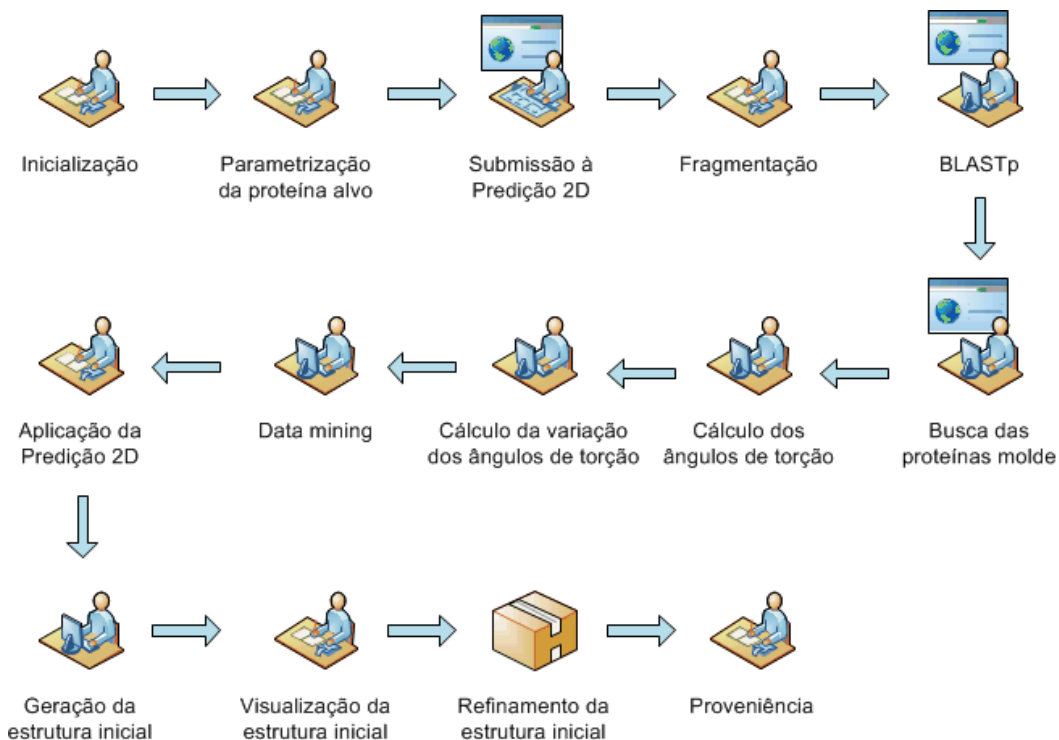


Figura 28 – Fluxograma da versão inicial do CReF.

Como pode ser visto no fluxograma da Figura 28, a versão inicial do CReF era composta de uma série de tarefas manuais. Algumas tarefas eram realizadas com o auxílio de scripts (como: BLASTp, cálculo dos ângulos de torção, *data mining*, etc.), mas sua execução dependia do usuário da ferramenta. A tarefa de refinamento foi considerada como uma caixa neste esquema, pois se trata de outro processo constituído de algumas etapas e para a nova versão do CReF optou-se pela utilização de estratégias diferentes. Devido a esta opção, o processo de refinamento na versão inicial não foi analisado e simulado por este trabalho.

Na nova versão, a maioria das tarefas foi automatizada e a dependência do usuário foi reduzida ao máximo, como demonstra o fluxograma a seguir:

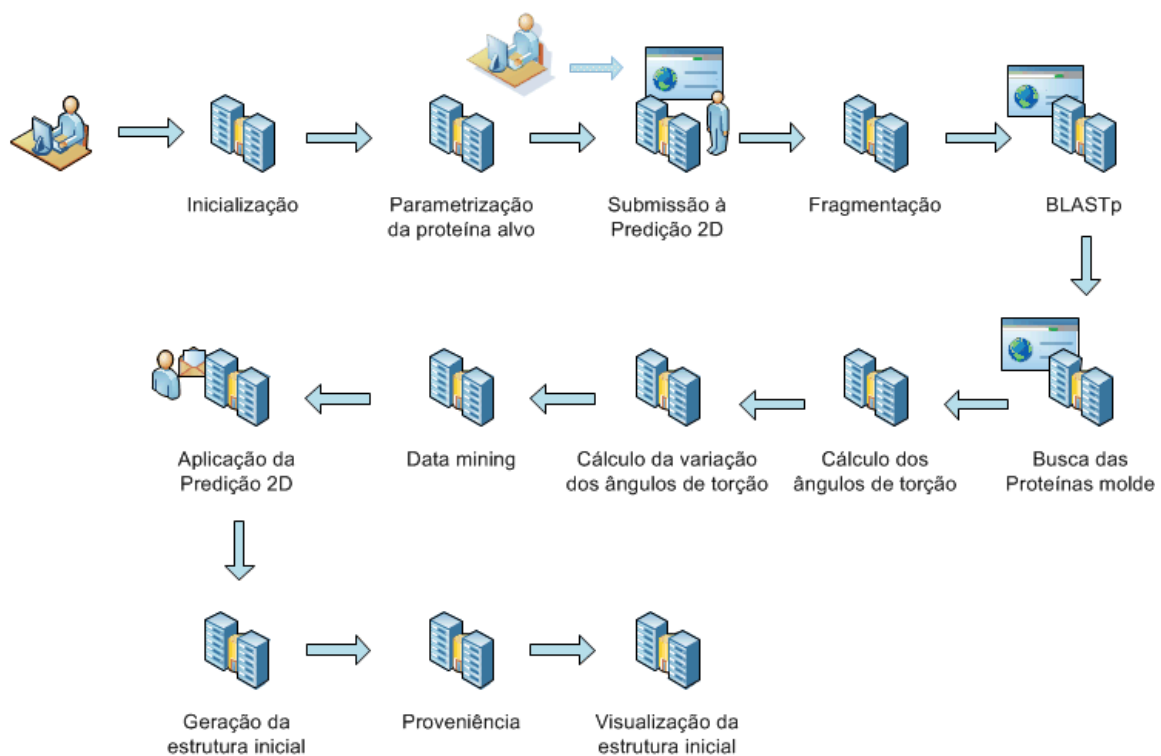


Figura 29 – Fluxograma da versão atual do CReF.

Com as melhorias implementadas, o método é executado por dois scripts principais. O primeiro script, que é o ponto de início da execução do processo, é responsável pelas tarefas de inicialização e parametrização da proteína alvo que correspondem à etapa de pré-execução do método. Invocado automaticamente pelo primeiro, o segundo script é responsável pelas demais tarefas que correspondem às etapas do método CReF (descritas no capítulo 4). Com este modelo o usuário terá facilidade, quando da ocorrência de algum erro ou problema em alguma etapa, re-

executar o método a partir da etapa na qual tenha ocorrido problema até à etapa final. Isso será possível através do segundo script, durante a execução será indicada cada etapa do método em execução. Desta forma, a indicação do erro e da etapa em que ocorreu o problema estará clara. Através da invocação do segundo script com a indicação do número da etapa a partir da qual deverá iniciar a execução, o usuário poderá realizar re-execuções parciais. Esse tipo de execução será possível a partir da tarefa de submissão à predição 2D conforme indicação do fluxograma acima.

Para um melhor entendimento do modelo de implementação da nova versão do CReF, cada uma das tarefas do fluxograma da Figura 29 é descrita a seguir:

- **Inicialização:** realiza a criação da estrutura de diretórios conforme indicação dos parâmetros da biblioteca do CReF. Diferentemente da versão inicial, o usuário passará a ter a possibilidade de indicar uma estrutura de diretórios diferente da indicada pela configuração inicial. A principal melhoria é a possibilidade de indicar um caminho diferente da instalação do CReF para o armazenamento das execuções.
- **Parametrização da proteína alvo:** cria os arquivos iniciais necessários para a execução do método. A partir da informação do código PDB e da sequência da proteína alvo a ser predita, serão gerados automaticamente os arquivos iniciais. Continuará sendo necessário (ou melhor, obrigatório) o usuário do método informar o arquivo com a listagem de proteínas molde que deverão ser desconsideradas pelo método em virtude da relação evolucionária que possuam com a proteína alvo.
- **Submissão à predição 2D:** submete a proteína alvo ao método de predição de estrutura 2D que é uma tarefa antecipada da etapa 7 - Predição da estrutura secundária. A nova versão do CReF passou a utilizar métodos que enviam o resultado da predição por e-mail, o que demanda algum tempo para submissão e envio, por isso essa tarefa foi antecipada. Quando a proteína alvo possuir mais de 20 aminoácidos, a submissão à predição 2D ocorrerá de forma automática, caso contrário, será preciso a intervenção do usuário para submeter à predição. Isso é necessário, pois o método usado para a predição de pequenas proteínas não permite submissão automática e, nesse caso, o usuário deverá submeter a proteína alvo manualmente.
- **Fragmentação:** realiza a etapa 1 de fragmentação da sequência alvo.
- **BLASTp:** submete cada um dos fragmentos ao BLASTp para obter a listagem de proteínas molde que é uma das atividades da etapa 2 – Busca por proteínas molde.
- **Busca das proteínas molde:** cria uma listagem de proteínas molde sem repetição e excluindo as proteínas indicadas pelo usuário que possuem relação evolucionária com a proteína alvo. Para concluir a etapa 2, faz o download dos PDBs das proteínas molde identificadas.

- **Cálculo dos ângulos de torção:** calcula os ângulos *phi* e *psi* das proteínas molde, o que corresponde à primeira parte da etapa 3 – Cálculo dos ângulos de torção dos dupletos.
- **Cálculo da variação dos ângulos de torção:** calcula para o aminoácido central dos fragmentos a variação dos ângulos obtidos de cada proteína molde associada ao fragmento. Essa tarefa complementa a etapa 3 do método CReF.
- **Data mining:** essa tarefa é composta por outras atividades. A primeira delas é a criação de arquivos com as informações da variação dos ângulos para submissão à ferramenta de mineração de dados Weka. Posteriormente, a ferramenta é executada e os dados resultantes do processo de mineração são processados, concluindo a etapa 4 do CReF: o agrupamento de dupletos. Fazendo uso do resultado da mineração de dados, constrói-se uma representação dos ângulos de torção na forma de intervalos por aminoácido da sequência alvo (etapa 5 do método). Os grupos candidatos a representar cada aminoácido da sequência alvo são classificados conforme a região que ocupam no mapa de Ramachandran (etapa 6).
- **Aplicação da predição 2D:** a predição da estrutura secundária da proteína alvo é utilizada para selecionar o melhor agrupamento para representar cada aminoácido da sequência alvo. Para que essa tarefa seja realizada é necessário que o usuário coloque o resultado da predição 2D, recebida por e-mail (o e-mail é enviado para o endereço que consta na biblioteca do CReF), em um arquivo em um local determinado. Desta forma, o processo poderá considerar de forma automática essa informação e finalizar a execução do método. Caso o arquivo não esteja disponível conforme o esperado, o método retornará um erro indicando isso e o usuário poderá realizar nova execução a partir deste ponto posteriormente. Essa tarefa corresponde à etapa 7 – Predição da estrutura secundária.
- **Geração da estrutura inicial:** constrói a estrutura inicial da predição da estrutura terciária da proteína alvo, o que corresponde a etapa 8 – Construção da conformação inicial.
- **Proveniência:** essa tarefa permite o armazenamento das execuções do método, após a conclusão de todas as etapas do CReF. Cada execução é identificada pela data e horário de término e pelo código PDB da proteína alvo. Na versão inicial, o usuário que desejasse criar proveniência deveria fazê-lo conforme seus critérios e de forma manual.
- **Visualização da estrutura inicial:** abre uma ferramenta para a visualização e manipulação 2D da estrutura inicial gerada pelo CReF sem refinamento. Esta tarefa também era realizada manualmente pelo usuário. O objetivo desta tarefa é permitir que, ao final da execução, o usuário já possa visualizar a estrutura inicial e fazer as manipulações que julgar interessante. Isso não impede que o usuário possa fazer uso de outras ferramentas para esta manipulação.

No fluxograma da nova versão do CReF (Figura 29), a tarefa de otimização da estrutura inicial (etapa 9 – seção 4.9) deixa de aparecer em virtude da opção por uma nova estratégia. Por isso, essa etapa será descrita em outra seção.

As melhorias técnicas implementadas permitiram ao CReF tornar-se uma ferramenta automatizada e parametrizável, melhorando substancialmente a sua usabilidade. A automatização não é completa devido à dependência da relação de proteínas moldes a excluir e da predição de estrutura secundária. Através da biblioteca CReF o usuário poderá parametrizar algumas informações tornando-se mais adaptável às necessidades do usuário.

5.1.2 Alterações no método

Os resultados preliminares obtidos com a primeira versão estabilizada foram a base das análises iniciais que evoluíram através de comparações entre resultados e indicaram modificações a serem aplicadas ao método. As alterações foram realizadas em diferentes etapas do processo e são apresentadas nesta seção.

Para construir a estrutura predita o CReF usa a técnica de mineração de dados das proteínas molde identificadas para cada sequência alvo. Dentre os algoritmos de mineração de dados, o *Expectation Maximization* (EM) é o mais adequado a ser aplicado neste contexto. A questão então surgida era relativa à quantidade de grupos (quatro grupos na versão inicial). A fim de esclarecer essa questão foram realizados testes com uma quantidade variada de grupos (de dois a sete grupos) para algumas proteínas do conjunto de teste inicial. Os testes indicaram que, para a maioria das proteínas teste, as conformações obtidas pelas execuções que consideraram seis grupos apresentaram um valor de RMSD menor ao da conformação resultante de quatro grupos (alguns desses resultados podem ser vistos na Figura 30). As conformações obtidas pelas demais proteínas teste podem ser consultadas no Apêndice A. Como isso não foi um consenso e para alguns casos a diferença de RMSD entre as conformações foi pequena, optou-se por oferecer a possibilidade de execução do processo com quatro ou seis grupos na mineração de dados. Na reestruturação técnica do método, essa opção ficou representada por um parâmetro disponível na biblioteca CReF.

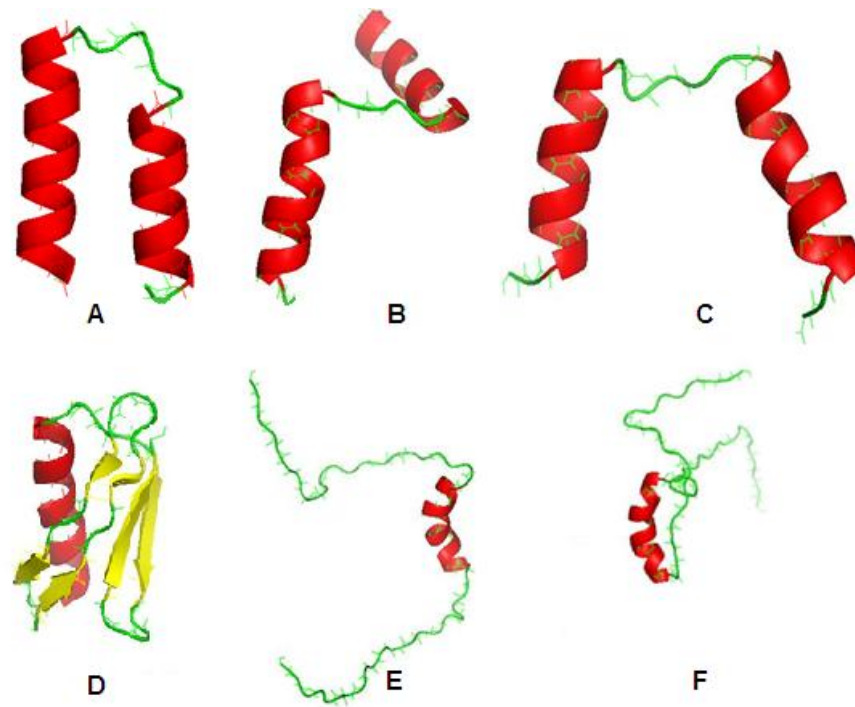


Figura 30 – Comparativo da conformação inicial das proteínas 1ZDD e 1GB1 com quatro e seis grupos na mineração dos dados: (A) estrutura experimental da 1ZDD, (B) estrutura inicial da 1ZDD com 4 grupos (RMSD: 9,82 Å), (C) estrutura inicial da 1ZDD com 6 grupos (RMSD: 6,64 Å), (D) estrutura experimental da 1GB1, (E) estrutura inicial da 1GB1 com 4 grupos (RMSD: 21,05 Å), (F) estrutura inicial da 1GB1 com 6 grupos (RMSD: 14,85 Å).

Na etapa 6 do CReF, os agrupamentos obtidos são rotulados conforme a região que ocupam no mapa de Ramachandran. Com base nos intervalos para os ângulos *phi* e *psi* que determinam cada agrupamento, uma função indica a região que corresponde ao intervalo no mapa de Ramachandran. A análise desta função apontou que os rótulos atribuídos não correspondiam aos descritos na dissertação que apresentava a versão original do CReF. Isso ocasionava dúvidas na aplicação das regras para construção da conformação inicial com base na predição de estrutura secundária obtida (Etapa 8). O modelo utilizado pelo PROCHECK (Laskowski et al., 1993) embasou a realização de um novo mapeamento dos intervalos de ângulos que determinam as regiões do mapa de Ramachandran. Além dos ajustes para garantir o mapeamento correto das regiões, os rótulos também foram alterados para que ficassem de acordo com a descrição do método: de “h” para “A”, de “e” para “B”, de “t” para “L”, de “P” para “p” e de “tl” para “l”.

A análise dos resultados parciais revelou que alguns resíduos da estrutura inicial estavam indevidamente localizados em regiões não permitidas. Para corrigir isso e na busca de melhorias na estrutura inicial, acrescentou-se à função o mapeamento da região não permitida do mapa de Ramachandran. Os agrupamentos com ângulos localizados nesta região passaram a ser rotulados com ‘X’ e foram desconsiderados pelas regras de aplicação do consenso da predição da estrutura secundária.

A versão original do CReF usava diferentes métodos para prever a estrutura secundária conforme a proteína alvo submetida (DPM (Deléage e Roux, 1987), DSC, PHD, PREDATOR e SCRATCH (Cheng et al., 2005)). Para a nova versão definiu-se padronizar o método de predição a aplicar para todas as submissões. Nesse sentido, métodos de predição já usados e outros métodos de predição de estrutura secundária disponíveis foram avaliados. Como o objetivo do CReF é prever estrutura 3D com base em proteínas com origem diferente da proteína alvo, métodos de predição baseados em homologia não foram considerados. Foram então avaliados os seguintes métodos de predição de ES: SOPM (Geourjon e Deléage, 1994), NetSurfP (Petersen et al., 2009), GorV (Sen et al., 2005), CSSP (Gupta et al., 2009), Sable (Adamczak et al., 2004), SAM-T08 (Karplus, 2009), Sspro4 (Cheng et al., 2005) e Porter (Pollastri e McLysaght, 2005). O desempenho individual dos métodos usados na aplicação do consenso da versão original do CReF (DPM, DSC, PHD, PREDATOR e SCRATCH) também foram avaliados. A predição de estrutura secundária obtida em cada um dos métodos avaliados foi comparada, resíduo a resíduo, com a estrutura determinada experimentalmente para determinar a acurácia de cada método (esta análise pode ser consultada no Apêndice B). Nos testes realizados, o método Porter apresentou a melhor acurácia para a maioria das proteínas, exceto para proteínas muito pequenas, como a 1K43 (com menos de 20 aminoácidos). Com base nesses resultados definiu-se o Porter como método a ser utilizado na predição de estrutura secundária para proteínas com 20 ou mais aminoácidos, e, para proteínas com menos de 20 aminoácidos, optou-se por utilizar o método SAM-T08. O SAM-T08 foi o método com a melhor acurácia para mini-proteínas.

Na etapa 8, a predição de estrutura secundária guia a escolha dos grupos retornados pelo Weka para representar cada um dos aminoácidos. A realização de sucessivos testes indicou que as regras que orientam essa escolha deveriam ser alteradas para compreenderem as diferentes situações possíveis. A nova versão do CReF passou a aplicar as regras 1 e 2 da seguinte forma:

Regra 1: o(s) grupo(s) k_i deve ter *rot* igual ao rótulo identificado para o i -ésimo aminoácido no consenso da predição da estrutura secundária. Quando for selecionado mais de um grupo, deve ser escolhido o grupo com maior número de tuplas t_i (Dorn, 2008). Essa escolha varia conforme o tipo de estrutura secundária identificada na predição:

- Hélice α (h): escolher os grupos rotulados por “A” (região mais favorável), caso não exista grupo nessa condição, deve escolher um grupo rotulado por “a” (região favorável);
- Folha β (e): escolher os grupos rotulados por “B” (região mais favorável), caso não exista grupo nessa condição, deve escolher um grupo rotulado por “b” (região favorável);

- Volta ou alça (c): escolher os grupos rotulados por “c” ($\sim a$, $\sim b$, $\sim l$, $\sim p$ e restante da área), em seguida grupos rotulados por: "p" (região favorável), "a" (região favorável), "b" (região favorável), "B" (região mais favorável), "A" (região mais favorável).

Regra 2: se na predição da estrutura secundária for predito um estado conformacional de estrutura regular (h ou e) para o i -ésimo resíduo e não existir um grupo k_i com rótulo *rot* igual ao identificado para este resíduo, deve-se proceder de acordo com a posição do i -ésimo resíduo na sequência:

- Início e fim: se o i -resíduo estiver no início ou no fim da sequência deve-se aplicar a regra 1 para volta ou alça para escolher um grupo k_i .
- Meio: se o i -resíduo estiver entre resíduos para os quais forem encontrados grupos conforme a regra 1, deve-se proceder calculando o valor médio entre os ângulos presentes nos resíduos $i - 1$ e $i + 1$. Os valores obtidos devem representar o i -ésimo resíduo na sequência alvo K , isso se justifica pelo padrão existente entre os ângulos diedros dos aminoácidos que compõem as estruturas secundárias regulares (Dorn, 2008).
- Resíduos contínuos: se o i -resíduo estiver entre um resíduo $i - 1$ para o qual foi encontrado um grupo k_i (a) e um resíduo $i + 1$ também sem grupo correspondente ao tipo de estrutura, deverá ser buscado o próximo resíduo $i + n$ (b) para o qual tenha sido encontrado grupo. Então, deve-se proceder com o cálculo da média entre (a) e (b), o valor obtido deve substituir o i -resíduo. Após isso o cálculo do resíduo $i + 1$ será composto pela média i -resíduo e do resíduo $i + n$. O valor obtido deve substituir o resíduo $i + 1$.

Estas regras eram aplicadas de forma manual para determinar quais agrupamentos representariam cada aminoácido da sequência alvo na versão inicial. Na nova versão do CReF, essas regras passaram a ser aplicadas automaticamente, desde que a informação da predição da estrutura secundária esteja disponível.

5.2 O problema da otimização de voltas e alças

Como já foi visto na seção 3.4, voltas e alças são espirais desorganizadas que conectam estruturas secundárias regulares e podem corresponder a um terço de uma proteína. Na Tabela 5 é

possível observar-se alguns dos diferentes tipos de voltas existentes. As alças aparecem na maioria das proteínas com mais de 60 resíduos, uma ou mais vezes, e se localizam, frequentemente, na superfície para exercer seu papel no reconhecimento biológico. As regiões de volta e alça possuem grande influência na determinação da forma de enovelamento das proteínas, isto é, na forma como as estruturas secundárias regulares organizam-se em nível terciário (Fiser et al., 2000). Isso se deve ao fato de que essas estruturas são bem mais flexíveis nas mudanças conformacionais do que hélices e folhas e, por tanto, podem ocupar qualquer região no mapa de Ramachandran, inclusive regiões de estruturas regulares. Em virtude disso, essas regiões provocam distorções no enovelamento que são difíceis de serem previstas *a priori* (Fiser et al., 2000).

A construção de regiões tão sensíveis quimicamente é influenciada por uma série de restrições (Liu et al., 2009). Por exemplo, as restrições do tipo estérica procuram evitar a ocorrência de colisões entre as cadeias laterais, o que impacta de forma direta na qualidade da conformação. Já as restrições de planaridade limitam os ângulos de torção da cadeia principal das proteínas.

A conformação de uma região de volta é influenciada também por átomos que não fazem parte da região, mas precedem e sucedem essa região (Fiser et al., 2000). Por isso, em uma abordagem que busque por segmentos que combinem com esta estrutura é importante considerar estas regiões (Fiser et al., 2000). A escolha desses segmentos deve ser orientada por critérios de geometria e/ou similaridade com a sequência alvo. Quando há a necessidade de endereçar a modelagem para uma classe específica de voltas que permita o uso de bibliotecas, essa abordagem é mais eficiente. Se a busca for realizada diretamente em uma base de dados de proteínas, como o PDB, é de extrema importância estabelecer critérios para limitar a busca, da mesma forma como é feito na busca de proteínas molde para uma predição.

A dificuldade na predição destas regiões aumenta conforme o tamanho do segmento (Fiser et al., 2000). Para segmentos maiores podem ser encontradas mais conformações incorretas, por isso, a função de energia, ou outro critério adotado, precisa identificar corretamente as melhores conformações. Os testes realizados por Fiser et al. (2000) comprovaram essa dificuldade e demonstraram que o ambiente pode influenciar no resultado final das conformações. Distorções no ambiente geram erros que tem maior impacto em regiões menores do que nas maiores. Isso deve explicar-se pelo fato desses casos já possuírem baixa precisão em ambiente nativo (Fiser et al., 2000).

Os erros na modelagem de voltas e alças podem ser compostos por erros na conformação e por erros na orientação em relação ao resto da proteína (Martin et al., 1997; van Vlijmen e Karplus, 1997). Uma das métricas mais comuns para medir esse erro é o RMSD, que pode ser local ou global. O RMSD local fornece a precisão da conformação, pois não depende da orientação. Já o RMSD global depende da conformação resultante da orientação da proteína predita (Fiser et al.,

2000). Portanto, a melhor conformação precisa apresentar uma combinação de menor valor de ambas as métricas, entre outras. A precisão da modelagem pode ser avaliada também em termos de RMSD e energia potencial, a conformação mais próxima da experimental deve ter baixo RMSD e baixa energia potencial. A obtenção dessa combinação de valores não é uma tarefa fácil, o que foi comprovado pelos experimentos iniciais do CReF.

Há uma grande quantidade de métodos que tem por objetivo a modelagem dessas regiões e que usam diferentes abordagens. Dentre eles, há uma classe de métodos analíticos que determina a conformação adequada através da identificação das possíveis soluções para um conjunto de equações algébricas derivadas da distância geométrica, conforme descrito no trabalho de Go e Sheraga (1970) e outros que se baseiam na teoria cinemática (Kolodny et al., 2005; Cahill et al., 2003). A abordagem chamada *Cyclic Coordinate Descent* – CCD (Canutescu e Dunbrack, 2003) que trata de voltas com diferentes comprimentos através do ajuste iterativo dos ângulos diedro foi incorporada ao Rosetta que apresentou os resultados obtidos em (Hu et al., 2007; Wang et al., 2007). Há métodos que se baseiam nas restrições impostas a estas regiões, como o algoritmo desenvolvido por Liu et al. (2009) que inicia com coordenadas atômicas randômicas e que gera conformações independentes com base em um esquema de ajuste de distâncias.

Os resultados da versão inicial do CReF comprovaram o quão complicado é prever espirais desorganizados. Tendo em conta as especificidades das regiões de voltas e alças, a versão inicial do CReF tratou-as na etapa 9 (seção 4.9). Esse tratamento gerava conformações com base na escolha aleatória dos ângulos de torção para compor uma região de volta. A partir dessa escolha as conformações eram construídas com o ajuste das cadeias laterais. O intervalo de conformações a ser analisado era reduzido através de matemática intervalar e uma função de energia potencial indicava a melhor conformação para representar a região analisada. Na busca pela melhoria na predição dessas regiões decidiu-se, no âmbito deste trabalho, abandonar essa estratégia e adotar uma nova. Seguindo a tendência dos métodos de modelagem de voltas e alças e dos métodos de avaliação de qualidade das conformações, foi escolhida a técnica de Dinâmica Molecular (DM) como estratégia de refinamento da conformação inicial predita pela nova versão do CReF. Para a realização do refinamento dessa conformação buscou-se ensaiar um protocolo com a indicação de parâmetros para as simulações, de forma que pudessem conduzir à melhoria da conformação predita. O detalhamento sobre os parâmetros e os resultados deste refinamento é apresentado no próximo capítulo.

5.3 Resumo do capítulo

Este capítulo discorreu sobre as melhorias realizadas no CReF e originaram uma nova versão da ferramenta. As implementações foram divididas em melhorias técnicas e melhorias do método. A mais significativa das melhorias foi na forma de execução que deixou de ser manual e passou a ser automatizada. Na nova versão o usuário pode disparar as diferentes etapas que geram a conformação inicial através de um único comando. Apenas duas tarefas do método podem necessitar da intervenção manual do utilizador.

O problema da otimização de voltas e alças foi aqui descrito com o intuito de demonstrar como esse tema é importante e para justificar a mudança na forma como essa questão passou a ser tratada na nova versão do CReF. O tratamento das regiões de voltas e alças deixou de ser considerado uma etapa do CReF e passou a ser tratado como um processo pós-CReF de refinamento da conformação inicial.

6. Experimentos

Nesta seção serão descritos os experimentos realizados e os resultados obtidos com a nova versão do método CReF gerada por este trabalho. Os experimentos iniciaram pelo mesmo conjunto de teste utilizado na versão inicial do CReF. Compõem este conjunto seis proteínas cujas estruturas 3D são conhecidas e representam diferentes classes de proteínas:

Tabela 8 – Conjunto inicial de proteínas teste utilizado nos experimentos de predição com o método CReF.

PDB	Nome	Classe SCOP	Enovelamento	Referências
1ZDD	Domínio A da mini-proteína estabilizada por pontes dissulfeto	proteína projetada	grampo α	(Starovasnick et al., 1997)
1K43	Cadeia A da proteína MBH12	proteína projetada	grampo β projetado	(Pastor et al., 2002)
1ROP	Cadeia A da proteína ROP (<i>Escherichia coli</i>)	α	grampo α	(Banner et al., 1987)
1UTG	Cadeia A da Uteroglobina (Coelho)	α	multi hélices	(Morize et al., 1987)
1GAB	Cadeia A da proteína PAB (<i>Escherichia coli</i>)	α	pacote de 3 hélices	(Johansson et al., 1997)
1GB1	Domínio B1 da proteína G do streptococcal (<i>Streptomyces griseus</i>)	$\alpha + \beta$	Mistura α e β	(Gronenborn et al., 1991)

Complementarmente a estas proteínas, foram realizados testes com um novo conjunto de proteínas teste que são apresentadas a seguir.

6.1 Definição de novo conjunto de proteínas teste

No intuito de aprimorar a predição do método CReF foi definido um novo conjunto de dez proteínas para teste. A escolha desse conjunto tomou por base artigos que descreviam a realização de testes de predição de estrutura 3D por outros métodos. Além de considerar proteínas já utilizadas em testes, foram selecionadas aquelas com diferentes classes estruturais (α , β e $\alpha\beta$) e tamanhos diferenciados (pequenas: com menos de 50 resíduos, médias: de 50 a 90, grandes: acima de 90). As proteínas selecionadas apresentam um empacotamento organizado das estruturas secundárias e foram testadas nos trabalhos de Simons et al. (2001), de Zhang et al. (2002) e de Zhang et al. (2003). Somente a proteína de código PDB 1YWJ não foi utilizada em testes prévios. Essa proteína

foi selecionada por meio de consulta ao PDB por pequenas ou mini-proteínas. As dez proteínas selecionadas foram:

Proteínas pequenas

- **2ERL**: Estrutura cristalina do feromônio ER-1 do protozoário ciliado *Euplotes raikovi*, determinada por raios X, considerada proteína pequena (40 aminoácidos), classe SCOP α , (Anderson et al., 1996).
- **1YWJ**: Estrutura do domínio FBPWW1 (*Homo sapiens*), determinada por NMR, considerada proteína pequena (41 aminoácidos), classe SCOP β , (Pires et al., 2005).
- **1GPT**: Estrutura em solução das tioninas gama 1-H e gama 1-P de cevada (*Hordeum vulgare*) e endosperma de trigo determinada por 1H-NMR: um motivo estrutural comum a proteínas tóxicas de artrópodes. Determinada por NMR, considerada proteína pequena (47 aminoácidos), classe SCOP pequenas proteínas ($\alpha\beta$), (Bruix et al., 1993).

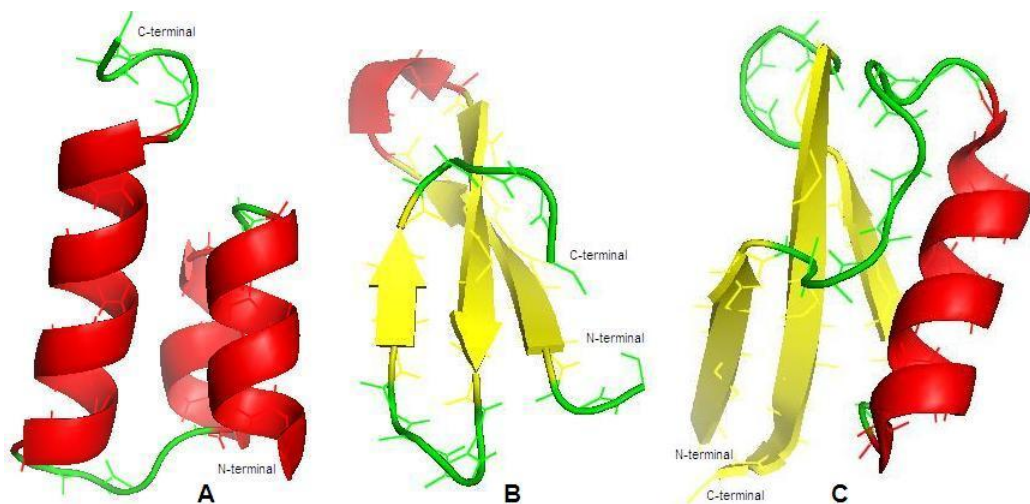


Figura 31 – Proteínas pequenas: (A) 2ERL, (B) 1YWJ, (C) 1GPT.

Proteínas médias

- **1C5A**: Estrutura tridimensional da C5a_{desArg} de porco, determinada por NMR, considerada proteína média (73 aminoácidos), classe SCOP α , (Williamson e Madison, 1990).
- **1CSP**: Estrutura cristalina da proteína principal de choque frio de *Bacillus subtilis* (*Escherichia coli*), determinada por raios X, considerada proteína média (67 aminoácidos), classe SCOP β , (Schindelin et al., 1993).

- **1CTF**: Estrutura do domínio C-terminal da proteína ribossomal L7/L12 de *Escherichia coli*, determinada por raios X, considerada proteína média (74 aminoácidos), classe SCOP $\alpha+\beta$, (Leijonmarck e Liljas, 1993).
- **1OPD**: Proteína rica em histidina (HPr), mutante Ser46Asp de *Escherichia coli*, determinada por raios X, considerada proteína média (85 aminoácidos), classe SCOP $\alpha+\beta$, (Napper et al., 1996).

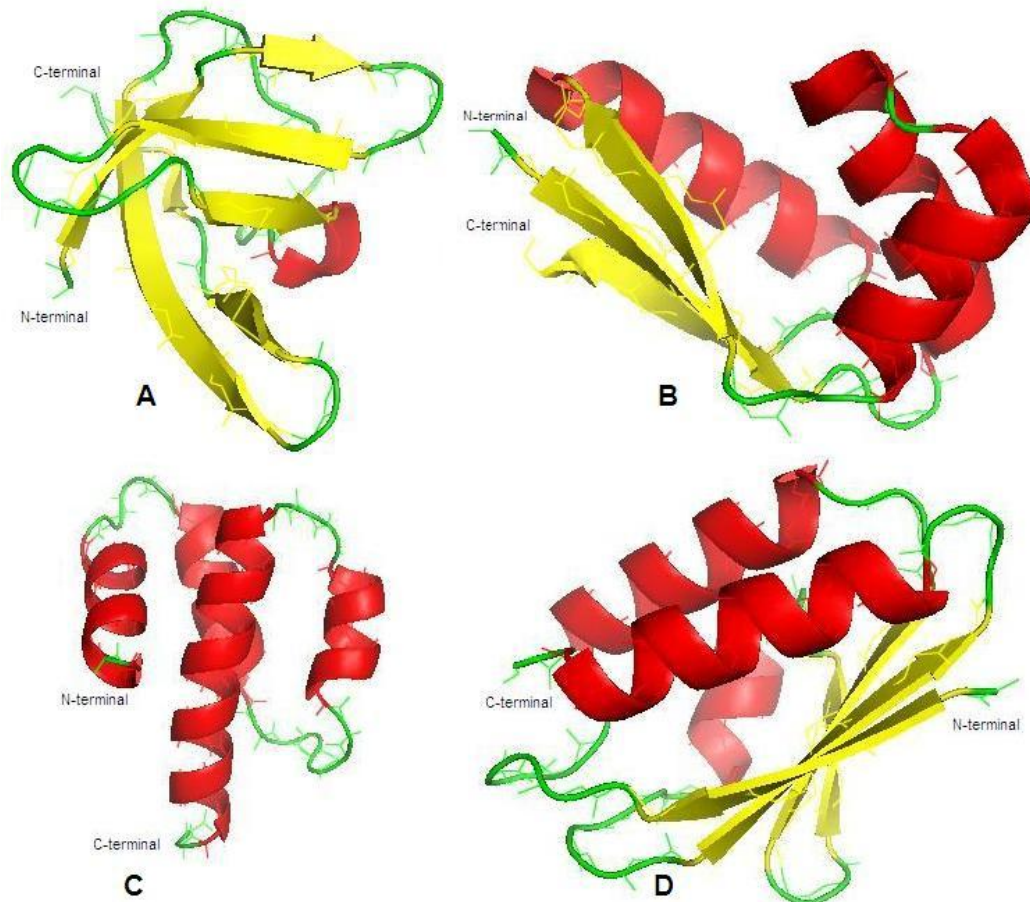


Figura 32 – Proteínas médias: (A) 1CSP, (B) 1CTF, (C) 1C5A, (D) 1OPD.

Proteínas grandes

- **2EZK**: Estrutura em solução do subdomínio Ibeta de ligação ao DNA da extremidade Mu da transposase Mu do fago Entobacteriófago Mu, determinada por NMR, considerada proteína grande (99 aminoácidos), classe SCOP α , (Schumacher et al., 1997).
- **1KSR**: Estrutura em solução do fator de gelificação da actina F (ABP-120) de *Dictyostelium discoideum*, determinada por NMR, considerada proteína grande (100 aminoácidos), classe SCOP β , (Fucini et al., 1997).

- **1ERV**: Estrutura da tioredoxina humana (*Homo sapiens*) mutante Cys73Ser (forma reduzida), determinada por raio X, considerada proteína grande (105 aminoácidos), classe SCOP α/β , (Weichsel et al., 1996).

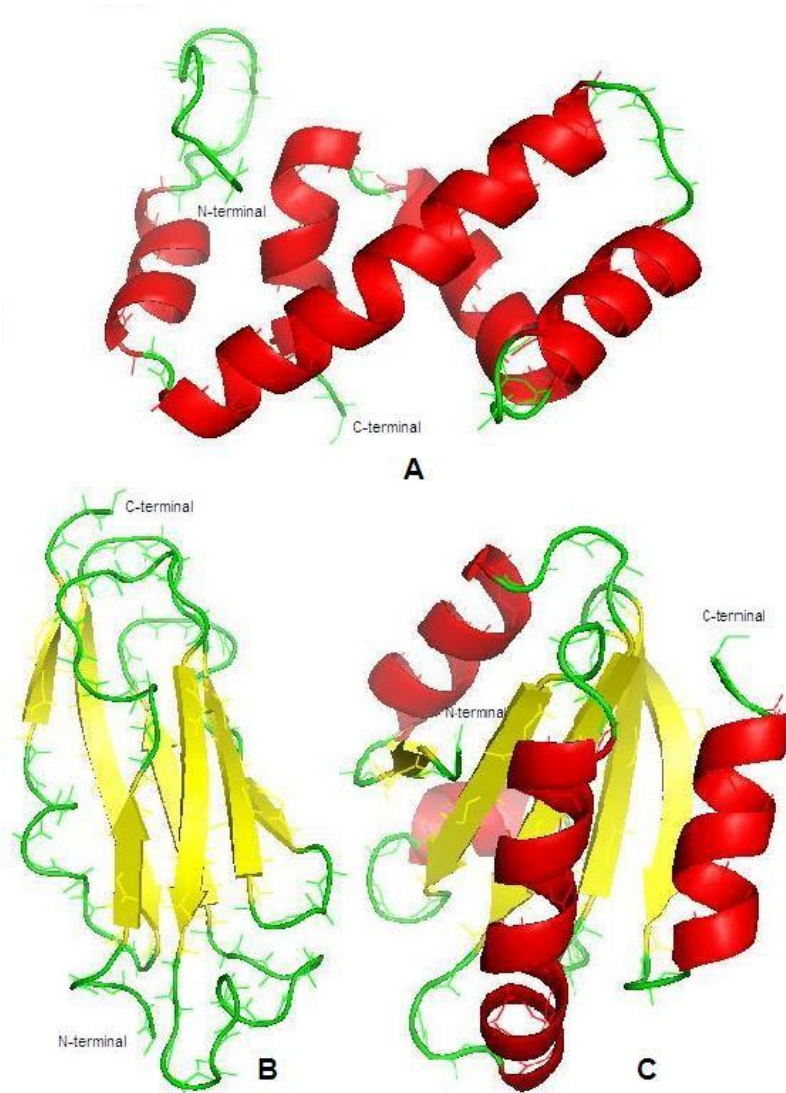


Figura 33 – Proteínas grandes: (A) 2EZK, (B) 1KSR, (C) 1ERV.

6.2 Materiais e métodos

O desenvolvimento da nova versão do CReF envolveu o uso de um conjunto de ferramentas e algumas parametrizações importantes para a execução do método. Quanto a parametrizações, o tamanho dos fragmentos foi mantido o mesmo da versão inicial ($l = 5$ resíduos). Para a seleção de proteínas molde foi adotado como critério a exclusão de todas aquelas proteínas

com mais de 30% de identidade que apresentassem relação evolucionária com a proteína alvo. A relação evolucionária pode não apresentar-se exclusivamente pela identidade, ela pode estar caracterizada por uma origem comum, enovelamento semelhante, etc. Para alguns casos a combinação de uma série de características determinará se a proteína molde deverá ser excluída ou não, portanto, o ideal seria a análise de um especialista. Quanto mais criteriosa for a indicação das proteínas alvo a excluir, maior será a garantia de que a predição do CReF baseia-se em estruturas diferentes da proteína alvo. Na biblioteca criada com as principais funções do CReF também estão disponíveis os parâmetros para execução do método e que podem ser alterados pelo usuário conforme a necessidade. A figura a seguir apresenta estes parâmetros adaptáveis:

```
#!/usr/bin/env python
# -*- coding: latin1 -*-

path_def = "/home/karina/CReF/arquivos/"

path_info_seq = path_def + "info.seq"
path_alvo_blast = path_def + "alvo.blast"
path_alvo_seq = path_def + "alvo.seq"
path_pdb_list = path_def + "pdb.list"
path_pdb_dow = path_def + "pdb.dow"
path_pdb_info = path_def + "pdb_info/"
path_fragm = path_def + "fragm_pdb_list/"
path_pdbs = "/media/SAMSUNG/pdbs/" #path_def + "pdbs/"
path_angulos = path_def + "angulos/"
path_var_ang = path_def + "variacao_angulos/"
path_weka = path_def + "weka/"
path_results = path_def + "weka/results/"
path_organizados = path_def + "weka/results/organizados/"
path_branch = path_def + "branch_bound/"
path_intervalos = path_def + "branch_bound/intervalos/"
folder_estr_ini = "branch_bound/estr_inicial/"
path_estr_ini = path_def + folder_estr_ini

num_grupos_weka = 6 # ou 4
email_pred = "karina.dallagno@acad.pucrs.br"
```

Figura 34 – Parâmetros para execução na nova versão do CReF.

O CReF é executado em ambiente Linux e foi desenvolvido na linguagem de programação Python. A biblioteca BioPython de código aberto foi adaptada para a realização da consulta ao BLASTp (módulo NCBIWWW) e do download do arquivo das proteínas molde junto ao banco de dados PDB. A predição de estrutura secundária é feita através de dois métodos que disponibilizam página web para submissão das sequências e enviam o resultado da predição por e-mail. Para proteínas com 20 ou mais aminoácidos fez-se uso do método Porter (<http://distill.ucd.ie/porter/>) e para pequenas proteínas fez-se uso do método SAM-T08 (http://compbio.soe.ucsc.edu/SAM_T08/T08-query.html). Para a implementação da submissão automática de sequências ao Porter usou-se a linguagem de programação Ruby 1.9.2 (<http://www.ruby-lang.org/pt/>) que é de código aberto.

Para o cálculo dos ângulos de torção foi usada a ferramenta *Torsions* que a partir da leitura do arquivo PDB calcula os ângulos *phi*, *psi* e *omega*. A etapa de *data mining* foi implementada através da ferramenta Weka 3.6.0 que identifica padrões de agrupamento para os dados submetidos conforme os fragmentos da sequência alvo. A construção da conformação inicial com base nos resultados da mineração de dados e da predição de estrutura secundária foi realizada com o AMBER 9. Ao final da execução do CReF, a nova versão permite a visualização da conformação inicial que é feita através da ferramenta Jmol 12.2.5.

A avaliação de um método de predição deve ser realizada com a aplicação de métricas que possam quantificar a qualidade da estrutura predita. Para a avaliação da nova versão do CReF analisou-se as métricas utilizadas na avaliação dos grupos participantes da categoria de refinamento de estruturas do CASP9 (descritas em (MacCallum et al., 2011)). Dentre as métricas consideradas estava o desvio médio quadrático (RMSD - *Root Mean Square Deviation*) que realiza uma comparação átomo a átomo de cada estrutura. Levando em conta que as estruturas preditas a serem avaliadas se tratavam de conformações iniciais, considerou-se que métricas mais complexas não seriam adequadas para esta avaliação. Por isso, optou-se por usar o RMSD como métrica devido à facilidade de cálculo e por permitir uma avaliação global quando aplicado a toda conformação e uma avaliação local quando aplicado a uma estrutura secundária.

As análises das conformações obtidas, a sobreposição da estrutura predita com a estrutura experimental e o cálculo do RMSD foram realizados na ferramenta Swiss-PdbViewer v4.0.1 (Guex e Peitsch, 1997). Esta ferramenta permite que a estrutura possa ser girada e avaliada sob diferentes ângulos, entre outras análises, o que permite uma análise da qualidade da predição complementar ao valor do RMSD. O cálculo do RMSD foi obtido pela sobreposição dos átomos $C\alpha$ da estrutura predita com os $C\alpha$ da estrutura experimental ou de comparação. A criação de mapas e algumas análises foram realizadas por meio do Ramachandran Plot 2.0 (<http://dicsoft1.physics.iisc.ernet.in/rp/index.html>) que permite a geração de mapas com base em um código PDB ou em um arquivo, e também disponibiliza opções de visualização: como da proteína inteira ou de apenas um segmento. As representações gráficas das estruturas 3D foram preparadas através do PyMOL (Delano, 2002).

A implementação da nova versão do CReF foi realizada em um ambiente de máquina virtual com o sistema operacional Ubuntu 10.04. Essa máquina virtual é executada em um notebook HP Intel CPU T2050 1,60 GHz com sistema operacional Windows XP, 103 GB de espaço em disco e 2 GB de memória RAM.

Como o refinamento usa estratégias e ferramentas diferentes em um contexto bem específico, optou-se pela descrição dos materiais e métodos juntamente com a descrição do protocolo na seção 6.5.

6.3 Experimentos com a nova versão do CReF

Ao contrário do que se considerou inicialmente, os experimentos da primeira versão não puderam ser repetidos. Para isso seria necessário considerar o mesmo conjunto de proteínas molde dos experimentos iniciais, mas não havia a relação desses moldes para todas as proteínas utilizadas nos experimentos. Portanto, como os experimentos iniciais e os atuais foram realizados sobre conjuntos de proteínas molde diferentes, os resultados não são comparáveis de forma direta. Sendo assim, os resultados iniciais (estrutura 3D e valor de RMSD) serviram de parâmetro para os experimentos da nova versão. Os resultados iniciais considerados referem-se à conformação inicial, isto é, referem-se à conformação obtida antes da execução da etapa de otimização das regiões de volta e alça. Isso foi necessário já que a nova versão do CReF considera a otimização dessa região um processo pós-CReF. Dessa forma, as predições geradas pelos experimentos iniciais são comparáveis apenas às conformações obtidas após o refinamento da conformação inicial resultante da execução da nova versão.

Nesse capítulo serão apresentados os resultados para algumas proteínas alvo do conjunto de teste: duas delas pertencentes ao conjunto teste dos experimentos iniciais e as outras duas pertencentes ao novo conjunto de proteínas teste. Para as duas primeiras proteínas o objetivo, além de analisar os resultados, é avaliar a diferença no desempenho de ambas as versões do CReF. Para as outras proteínas o objetivo é analisar se as características do método mantêm-se com outros alvos. O resultado obtido para as demais proteínas alvo pode ser consultado em formato resumido no Apêndice C deste trabalho.

6.3.1 Estudo de caso 1: 1ZDD

O Domínio A da mini-proteína estabilizada por pontes dissulfeto cujo código PDB é 1ZDD foi alvo dos testes de predição de ambas as versões do CReF. A sua sequência com 34 resíduos de aminoácidos foi dividida em 30 fragmentos com tamanho de $l = 5$ resíduos. Para a busca dos fragmentos molde no PDB foram excluídas 19 proteínas que apresentaram relação evolucionária com a proteína alvo. Essas são as mesmas proteínas que foram desconsideradas no teste realizado com a versão inicial do CReF: 1ZDC, 1ZDD, 1L6X, 1OQO, 1OQX, 1ZDA, 1ZDB, 2SPZ, 1LP1, 1Q2N, 1FC2, 1BDC, 1BDD, 1SS1, 1DEE, 1EDK, 1EDJ, 1EDI e 1EDL.

A Tabela 9 apresenta os fragmentos na qual a 1ZDD se dividiu e um comparativo entre a quantidade de moldes obtida com a versão inicial e com a nova versão. Em média, houve um crescimento de 137% na quantidade de moldes encontrados por fragmento.

Tabela 9 – Quantidade de moldes por fragmento da proteína alvo 1ZDD.

Fragmentos	CReF 1.0 Nº moldes	CReF 2.0 Nº moldes	Crescimento
FNMQC	34	49	144 %
NMQCQ	40	47	118 %
MQCQR	38	66	174 %
QCQRR	28	65	232 %
CQRRF	47	71	151 %
QRRFY	30	32	107 %
RRFYE	52	31	60 %
RFYEA	30	26	87 %
FYEAL	42	51	121 %
YEALH	37	64	173 %
EALHD	45	58	129 %
ALHDP	49	53	108 %
LHDPN	53	58	109 %
HDPNL	22	40	182 %
DPNLN	50	46	92 %
PNLNE	24	41	171 %
NLNEE	46	57	124 %
LNEEQ	34	26	76 %
NEEQR	37	28	76 %
EEQRN	35	45	129 %
EQRNA	57	44	77 %
QRNAK	14	39	279 %
RNAKI	29	36	124 %
NAKIK	20	55	275 %
AKIKS	28	45	161 %
KIKSI	40	42	105 %
IKSIR	74	40	54 %
KSIRD	22	44	200 %
SIRDD	16	28	175 %
IRDDC	35	33	94 %

A nova versão do CReF possibilita a escolha da quantidade de grupos a serem buscados na etapa de mineração de dados, que pode ser quatro ou seis. Foram realizados testes com ambas as quantidades de grupos, mas considerando o mesmo conjunto de proteínas molde. Desta forma, para ambos os casos, foi submetido o mesmo conjunto de dados para a mineração. O resultado da mineração indica um conjunto de grupos de ângulos *phi* e *psi* para representar cada resíduo de aminoácido da sequência alvo, excluindo os dois primeiros e os dois últimos resíduos da sequência. Os grupos candidatos a representar cada resíduo são apresentados em ordem de significância, isto é, o primeiro grupo é aquele que reuniu uma quantidade maior de moldes e, assim por diante. Cada

grupo é classificado de acordo com a região do mapa de Ramachandran a que ele se refere. O resultado da predição da estrutura secundária orienta na escolha do grupo que deve representar cada resíduo. Portanto, para cada resíduo deve ser escolhido o grupo mais significativo que tenha a mesma classificação indicada pela predição 2D. Na análise, sob o ponto de vista da região do mapa de Ramachandran, o processo de mineração teve um acerto de 67%, isto é, para 20 resíduos não houve a necessidade de selecionar outro grupo para representá-lo. Para os resíduos em que foi necessário selecionar outro grupo, encontrou-se para todos os casos um grupo com classificação conforme a indicação da predição 2D.

A partir da conformação inicial foram analisadas as estruturas secundárias da 1ZDD que são duas hélices: hélice 1 com 11 resíduos e hélice 2 com 14 resíduos. A sobreposição dessas estruturas com as estruturas secundárias da proteína experimental comprovou que as estruturas foram bem preditas, mantendo o desempenho da versão inicial, conforme demonstram os valores em RMSD:

Tabela 10 – Análise do RMSD em Å das estruturas secundárias e conformação inicial das predições da 1ZDD.

Estrutura 2D	RMSD 4 Grupos	RMSD 6 Grupos
Hélice 1	0,49	0,55
Hélice 2	0,70	0,99
Conformação inicial	9,82	6,64

Os mapas da Figura 35 também confirmam a boa predição das hélices (correspondem à concentração de pontos na região “A” em vermelho) e, de forma geral, a predição (B e C) ocupou as mesmas regiões do mapa que foram ocupadas pela estrutura experimental (A). Os resíduos da predição que ficaram mais deslocados foram os da alça, como o ASP15 que nas predições ficou na região de “~l” quando deveria ficar na região de “b”.

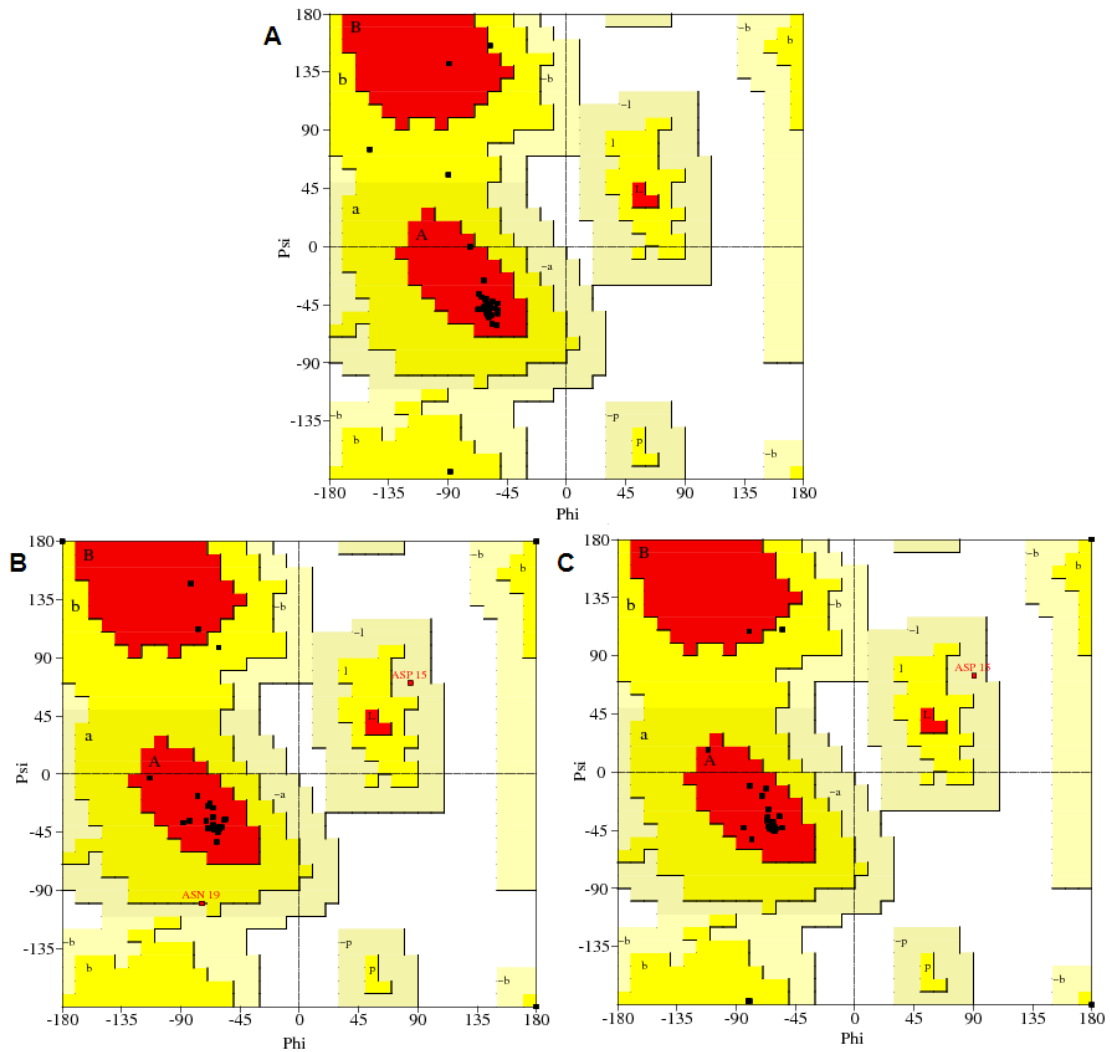


Figura 35 – Mapa de Ramachandran da proteína alvo 1ZDD: (A) mapa da estrutura experimental, (B) mapa da estrutura predita sem otimização com base em 4 grupos na etapa de mineração de dados, (C) mapa da estrutura predita sem otimização com base em 6 grupos na mineração.

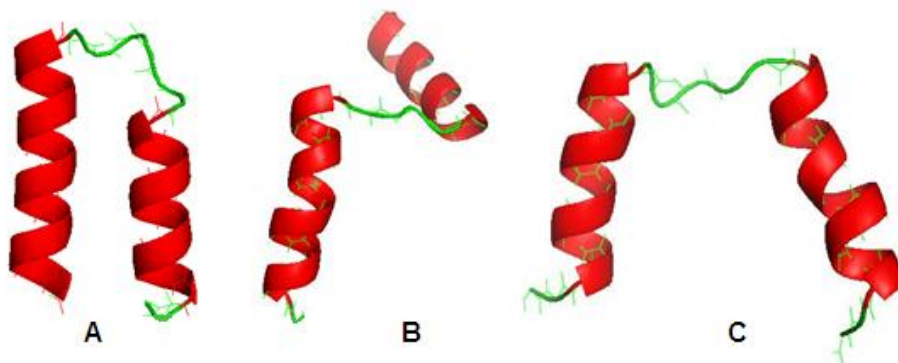


Figura 36 – Comparação da conformação inicial da proteína 1ZDD com variação de grupos na mineração dos dados: (A) estrutura experimental da 1ZDD, (B) conformação inicial com 4 grupos (RMSD: 9,82 Å) e (C) conformação inicial com 6 grupos (RMSD: 6,64 Å).

A análise das estruturas gráficas da Figura 36 ajuda a compreender a diferença indicada pelos valores RMSD. Quanto às hélices pode-se ver que elas estão um pouco desalinhadas quando comparadas à estrutura experimental, mas bem formadas conforme mostram os valores de RMSD da Tabela 10 e mapa de Ramachandran (Figura 35). A diferença principal está na conformação da alça que liga as hélices, da mesma forma como mostraram os resultados preliminares.

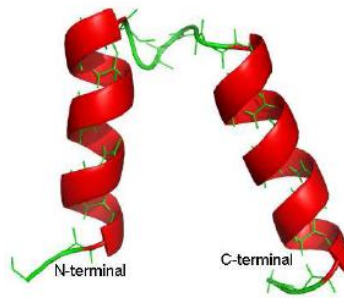


Figura 37 – Estrutura 3D predita (conformação inicial) pela versão inicial do CReF antes da otimização de voltas (RMSD: 5,51 Å) (Dorn, 2008).

Para a proteína alvo 1ZDD a nova versão do CReF obteve uma conformação inicial (Tabela 10) com valor de RMSD próximo à conformação obtida inicialmente. Devido à falta de histórico e aos problemas para execução da versão inicial do método não é possível afirmar se a versão inicial manteria esse resultado nas condições atuais do banco de dados PDB.

A 1ZDD também foi submetida para predição em dois dos principais métodos de predição de estrutura terciária. O método I-Tasser permite a exclusão de proteínas que tem relação evolucionária com a proteína alvo, desta forma a submissão foi realizada excluindo o mesmo conjunto de proteínas alvo que o CReF. O RMSD da estrutura predita foi de 1,57 Å, mas que não pode ser comparado aos valores da Figura 36 já que estas são conformações iniciais sem refinamento de voltas e alças. A 1ZDD também foi submetida ao Robetta (Kim et al., 2004) cuja estrutura predita apresentou 1,32 Å de RMSD, mas este não é comparável ao CReF. Isso se deve ao fato que a predição do Robetta foi realizada com base em moldes que foram excluídos pelo CReF, isto é, a predição foi realizada com base em moldes muito próximos à proteína alvo. Esses resultados referem-se à conformação final da predição, isto é, podem ser comparáveis ao resultado do refinamento de uma conformação inicial. Neste sentido, são resultados de referência de uma excelente predição 3D para a proteína 1ZDD.

6.3.2 Estudo de caso 2: 1GB1

O Domínio B1 da proteína G do streptococcal cujo código PDB é 1GB1 foi alvo dos testes de predição de ambas as versões do CReF. Na busca por fragmentos molde no PDB foram excluídas 62 proteínas que apresentaram relação evolucionária com a proteína alvo: 1GB1, 1PGA, 1PGB, 2GB1, 2KBT, 3GB1, 3MP9, 1IBX, 1PN5, 2GI9, 2I2Y, 2I38, 2JSV, 2JU6, 2K0P, 2KHU, 2KHW, 2KN4, 2KQ4, 2KWD, 2QMT, 1FCC, 2IGG, 2J52, 2J53, 2PLP, 1FD6, 1QKZ, 1UWX, 2KLL, 2RMM, 2CWB, 2DEN, 2RPV, 1Q10, 2ZW0, 1FCL, 1IGC, 1IGD, 1PGX, 2IGD, 2IGH, 2NMQ, 1EM7, 1P7E, 1P7F, 2OED, 1GB4, 1MPE, 1MVK, 3FIL, 2ONQ, 2ON8, 2ZW1, 1MI0, 1MHX, 2JWU, 2KDM, 2KDL, 1ZXH, 2JWS e 1LE3. Dentre elas estão as 35 proteínas desconsideradas no teste realizado com a versão inicial do CReF.

Os 56 resíduos de aminoácidos da sua sequência foram divididos em 52 fragmentos com tamanho de $l = 5$ resíduos. Estes fragmentos e um comparativo entre a quantidade de moldes obtida com a versão inicial e com a nova versão são apresentados na tabela a seguir. Em média, houve um crescimento de 353% na quantidade de moldes encontrados por fragmento.

Tabela 11 – Quantidade de moldes por fragmento da proteína alvo 1GB1.

Fragmentos	CReF 1.0 Nº moldes	CReF 2.0 Nº moldes	Crescimento
MTYKL	16	29	181 %
TYKLI	37	159	430 %
YKLIL	18	53	294 %
KLILN	12	44	367 %
LILNG	19	63	332 %
ILNGK	34	51	150 %
LNGKT	10	27	270 %
NGKTL	35	49	140 %
GKTLK	15	49	327 %
KTLKG	15	63	420 %
TLKGE	15	74	493 %
LKGET	12	55	458 %
KGETT	8	43	538 %
GETTT	19	29	153 %
ETTTE	21	43	205 %
TTTEA	3	29	967 %
TTEAV	13	70	538 %
TEAVD	18	46	256 %
EAVDA	42	45	107 %
AVDAA	18	35	194 %
VDAAT	33	48	145 %

Continua na próxima página

Fragmentos	CReF 1.0 Nº moldes	CReF 2.0 Nº moldes	Crescimento
DAATA	15	28	187 ‰
AATAE	14	57	407 ‰
ATAEK	14	38	271 ‰
TAEKV	21	64	305 ‰
AEKVF	19	29	153 ‰
EKVFK	16	32	200 ‰
KVFKQ	20	118	590 ‰
VFKQY	15	61	407 ‰
FKQYA	10	60	600 ‰
KQYAN	58	47	81 ‰
QYAND	52	60	115 ‰
YANDN	15	59	393 ‰
ANDNG	5	59	1.180 ‰
NDNGV	18	45	250 ‰
DNGVD	29	39	134 ‰
NGVDG	13	51	392 ‰
GVDGE	7	44	629 ‰
VDGEW	21	66	314 ‰
DGEWT	16	35	219 ‰
GEWTY	49	66	135 ‰
EWTYD	17	57	335 ‰
WTYDD	12	43	358 ‰
TYDDA	11	56	509 ‰
YDDAT	35	33	94 ‰
DDATK	15	36	240 ‰
DATKT	11	39	355 ‰
ATKTF	16	48	300 ‰
TKTFT	5	72	1440 ‰
KTFTV	17	49	288 ‰
TFTVT	15	44	293 ‰
FTVTE	18	42	233 ‰

O processo de mineração teve um acerto de 56% em relação aos grupos mais significativos para representar os resíduos, isto é, para 23 grupos foi necessário selecionar outro grupo para representar o resíduo. Para todos esses casos encontrou-se um grupo com classificação conforme a indicada pela predição 2D.

A proteína 1GB1 apresenta uma estrutura secundária do tipo hélice (com 13 resíduos) e uma folha β com quatro fitas. Essas estruturas foram analisadas na conformação inicial e a hélice foi bem predita conforme demonstra o valor do RMSD abaixo. Já as fitas não foram formadas, da mesma forma como aconteceu no teste da versão inicial.

Tabela 12 – Análise do RMSD em Å das estruturas secundárias e conformação inicial das predições da 1GB1.

Estrutura 2D	RMSD 4 Grupos	RMSD 6 Grupos
Hélice 1	1,58	1,60
Fita 1	0,71	0,88
Fita 2	0,89	1,04
Fita 3	0,27	0,39
Fita 4	0,43	0,33
Conformação inicial	21,05	14,85

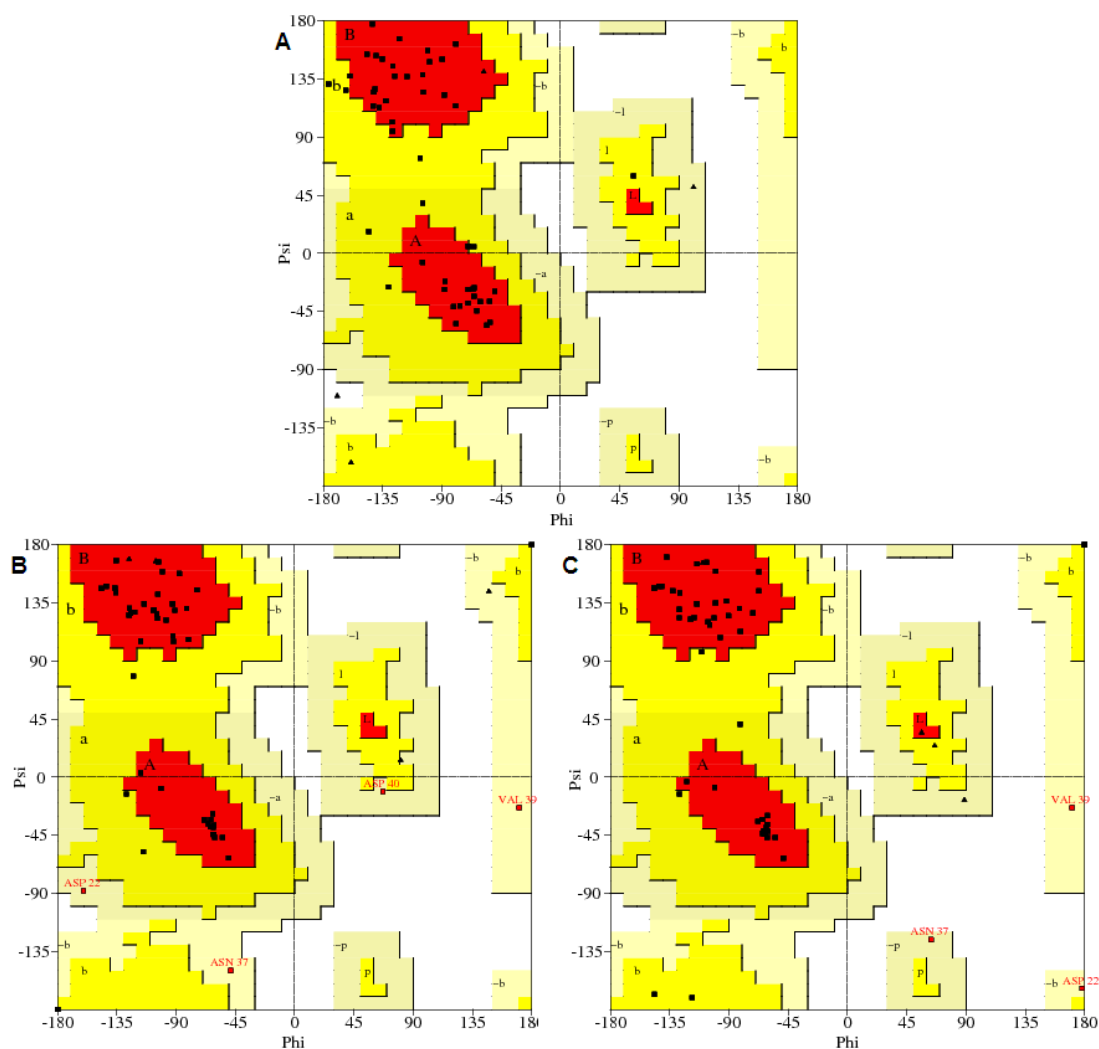


Figura 38 – Mapa de Ramachandran da proteína alvo 1GB1: (A) mapa da estrutura experimental, (B) mapa da estrutura predita sem otimização com base em 4 grupos na etapa de mineração de dados, (C) mapa da estrutura predita sem otimização com base em 6 grupos na mineração.

Os mapas apresentados na Figura 38 também mostram a boa predição da hélice, que corresponde à concentração de pontos na região “A” (mais favorável em vermelho). Em (B) e (C) verifica-se que nestas concentrações os pontos ficaram reunidos e alguns sobrepostos, já em (A) vê-se os pontos mais dispersos. Ainda em relação à hélice, vê-se que a predição com quatro e seis grupos ocuparam praticamente a mesma região. Em relação às fitas, as predições com ambos os grupos concentraram os resíduos na região de “B”. Alguns resíduos das fitas localizam-se em outras regiões do mapa na estrutura experimental (A), para estes resíduos as predições apresentaram a principal divergência mapeando-os em regiões diferentes e com maiores distorções. Nos mapas (B) e (C) observa-se alguns resíduos identificados e que ocupam outras regiões, como por exemplo: Asn37, Val39 e Asp 40. Estes resíduos são de regiões de volta e juntamente com os demais resíduos dessas regiões são os que apresentaram maior discrepância. Isso ocasionou o posicionamento incorreto das estruturas no espaço 3D e, fazendo com que as ligações de hidrogênio não se estabelecessem e a folha não se formasse.

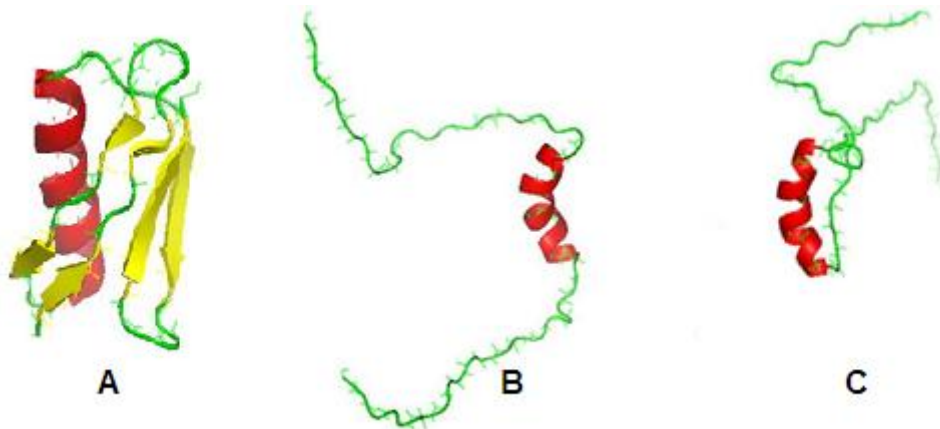


Figura 39 – Comparação da conformação inicial da proteína 1GB1 com variação de grupos na mineração dos dados: (A) estrutura experimental da 1GB1, (B) conformação inicial com 4 grupos (RMSD: 21,05 Å) e (C) conformação inicial com 6 grupos (RMSD: 14,85 Å).

As estruturas gráficas da Figura 39 comprovam as análises anteriores. Na predição com quatro grupos (B) não há possibilidade de formação da folha com o enovelamento apresentado. Já na predição com seis grupos houve um enovelamento mais aproximado da estrutura experimental, fato também expressado pelo valor do RMSD. Nas predições (B) e (C) verifica-se que a hélice apresenta uma curvatura semelhante, mas que difere da estrutura experimental (A). De forma geral, a má conformação das regiões de voltas e alças implicou em uma má predição de proteínas com folhas β expressada por um valor de RMSD mais alto.

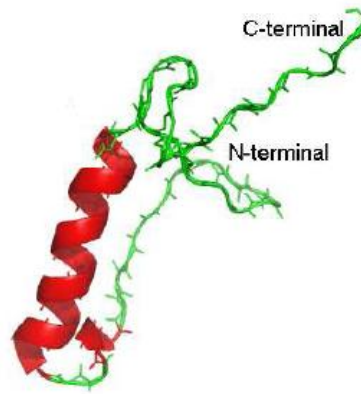


Figura 40 – Estrutura 3D predita (conformação inicial) pela versão inicial do CReF antes da otimização de voltas (RMSD: 12,31 Å) (Dorn, 2008).

A nova versão do CReF fazendo uso de seis grupos na etapa de mineração de dados (Tabela 12), obteve uma conformação inicial com RMSD próximo da predição da versão inicial. A 1GB1 também foi submetida aos métodos de predição I-Tasser e Robetta e foram obtidos os valores de RMSD 5,52 Å e 1,06 Å, respectivamente. O excelente resultado do Robetta é explicado pelo uso de moldes muito próximos e que foram excluídos na execução do CReF. Sendo assim, somente o valor do I-Tasser pode ser considerado como referência para a predição da 1GB1 após um processo de refinamento da conformação inicial.

6.3.3 Estudo de caso 3: 1C5A

A estrutura tridimensional da C5a_{desArg} de porco, cujo código PDB é 1C5A, é uma das proteínas de tamanho médio que compõe o novo conjunto de teste. Na busca por fragmentos molde no PDB foram excluídas 10 proteínas que apresentaram relação evolucionária com a proteína alvo: 1C5A, 3CU7, 3PRX, 3PVM, 3HQA, 3HQB, 1KJS, 1CFA, 2B39 e 2A73. A sua sequência com 73 resíduos de aminoácidos foi dividida em 69 fragmentos com tamanho de $l = 5$ resíduos que são apresentados na tabela a seguir junto da quantidade de moldes obtida na execução do método.

No processamento do resultado da mineração de dados, para 48% dos resíduos, foi necessário selecionar outro conjunto de ângulos para representá-los. Para todos esses resíduos encontrou-se um grupo com classificação adequada para representá-los na conformação inicial. O método Porter de predição da estrutura secundária não conseguiu atribuir um tipo de estrutura para os sete últimos resíduos da sequência, para estes resíduos optou-se pela escolha de grupos com a classificação de *coil*.

Tabela 13 – Quantidade de moldes por fragmento da proteína alvo 1C5A.

Fragmentos	CReF 2.0	Fragmentos	CReF 2.0
	Nº moldes		Nº moldes
MLQKK	66	ERAAR	45
LQKKI	41	RAARI	20
QKKIE	45	AARIK	35
KKIEE	43	ARIKI	63
KIEEE	34	RIKIG	30
IEEEA	53	IKIGP	49
EEEEA	31	KIGPK	48
EEAAK	28	IGPKC	55
EAAKY	60	GPKCV	42
AAKYK	75	PKCVK	38
AKYKY	38	KCVKA	58
KYKYA	40	CVKAF	75
YKYAM	27	VKAFK	40
KYAML	33	KAFKD	50
YAMLK	54	AFKDC	27
AMLKK	86	FKDCC	28
MLKKC	98	KDCCY	47
LKKCC	61	DCCYI	45
KKCCY	68	CCYIA	28
KCCYD	48	CYIAN	38
CCYDG	50	YIANQ	55
CYDGA	64	IANQV	47
YDGAY	44	ANQVR	44
DGAYR	83	NQVRA	67
GAYRN	40	QVRAE	68
AYRND	55	VRAEQ	39
YRNDD	65	RAEQS	51
RNDDE	43	AEQSH	21
NDDET	48	EQSHK	36
DDETC	63	QSHKN	97
DETCE	54	SHKNI	103
ETCEE	27	HKNIQ	60
TCEER	24	KNIQL	49
CEERA	25	NIQLG	55
EERAA	63		

A proteína alvo 1C5A é composta por quatro hélices: hélice1 com nove resíduos, hélice2 com 11 resíduos, hélice3 com sete resíduos e hélice4 com 20 resíduos. As hélices das estruturas preditas foram sobrepostas com as hélices da estrutura experimental, o que demonstrou uma boa predição das hélices e conforme comprovam os valores RMSD:

Tabela 14 – Análise do RMSD em Å das estruturas secundárias e conformação inicial das predições da 1C5A.

Estrutura 2D	RMSD 4 Grupos	RMSD 6 Grupos
Hélice 1	1,16	0,41
Hélice 2	2,20	1,74
Hélice 3	0,96	0,61
Hélice 4	1,76	1,57
Conformação inicial	12,29	9,56

Na submissão aos métodos I-Tasser e Robetta, a proteína 1C5A obteve RMSD de 5,11 Å e 1,49 Å, respectivamente. A predição considerando seis grupos na mineração de dados foi a que apresentou RMSD mais próximo ao da predição do I-Tasser que é considerada como referência.

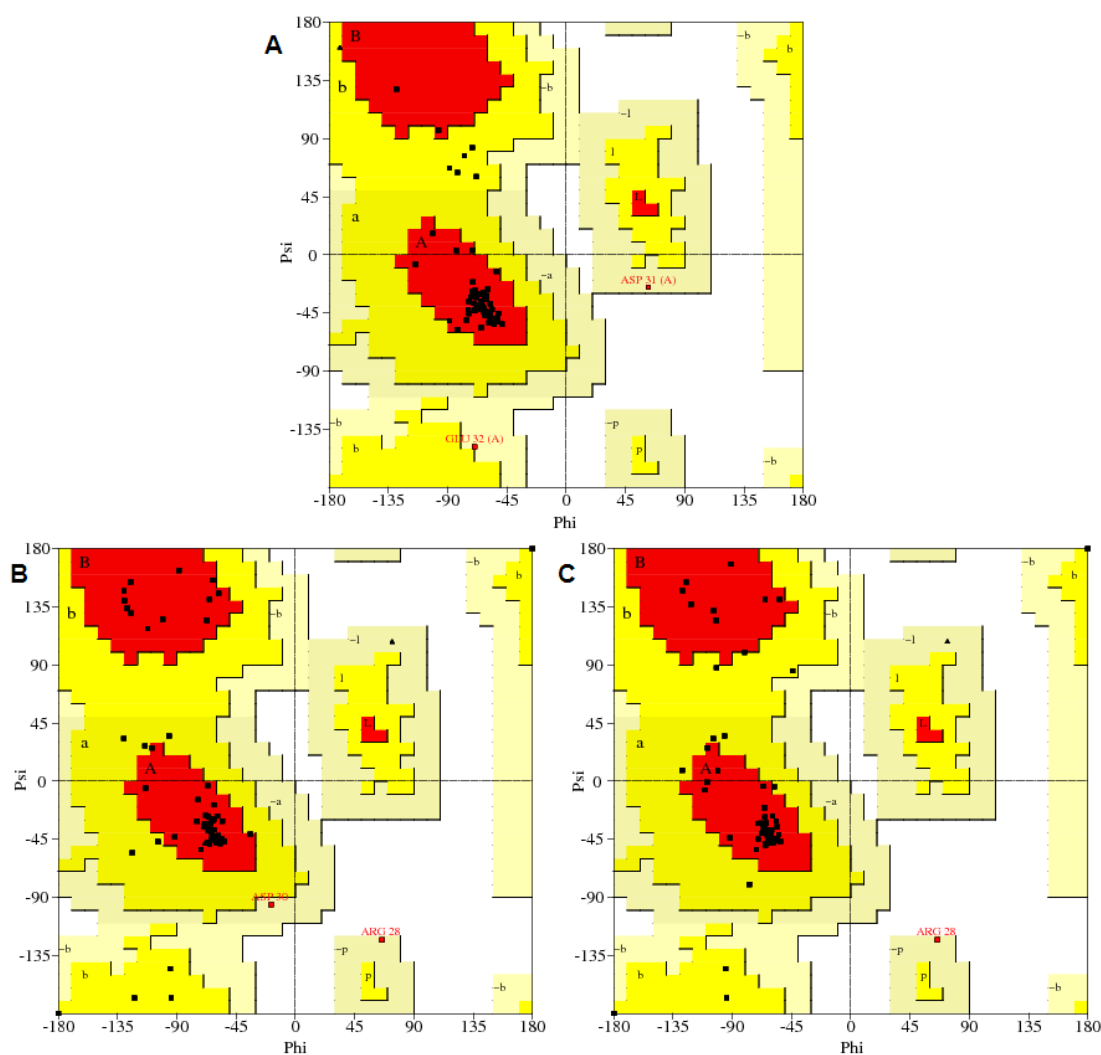


Figura 41 – Mapa de Ramachandran da proteína alvo 1C5A: (A) mapa da estrutura experimental, (B) mapa da estrutura predita sem otimização com base em 4 grupos na etapa de mineração de dados, (C) mapa da estrutura predita sem otimização com base em 6 grupos na mineração.

Na Figura 41 são apresentados os mapas de Ramachandran que analisam a disposição dos resíduos na predição. Nos três mapas verifica-se uma concentração de resíduos na região de “A” que corresponde às hélices. A análise dos mapas mostrou que, a maioria dos resíduos das hélices, se concentraram em uma mesma área da região “A”. Nas predições das hélices houve vários resíduos que ocuparam pontos muito próximos aos ocupados na estrutura experimental. Considerando os valores RMSD das hélices, essa proximidade compensou o posicionamento de alguns resíduos que ficaram em outra região do mapa ou apresentaram posição mais deslocada em relação à estrutura experimental. Os resíduos Asp31 e Glu32 destacados em (A) correspondem à região de volta e é uma demonstração de que também nesta proteína teste, a predição de voltas e alças apresentou as maiores distorções.

As estruturas gráficas da Figura 42 comprovam a boa formação das hélices e como o mau posicionamento as regiões de voltas fizeram com que as hélices não ocupassem a posição correta no espaço 3D. Em (C) é possível verificar que a hélice4 se posicionou de forma muito semelhante à estrutura experimental, o que contribuiu à obtenção de um RMSD global melhor do que o da outra predição.

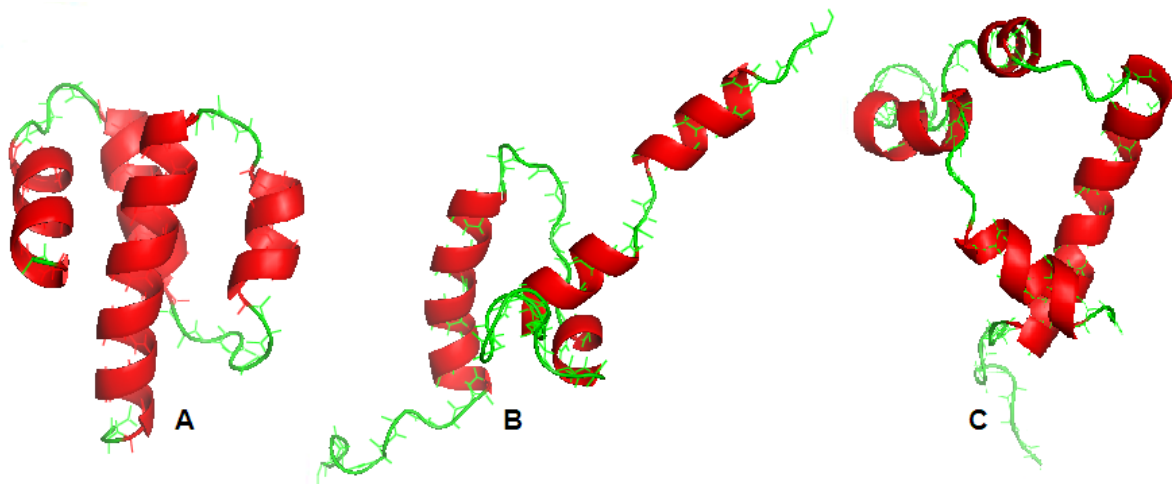


Figura 42 – Comparação da conformação inicial da proteína 1C5A com variação de grupos na mineração dos dados: (A) estrutura experimental da 1C5A, (B) conformação inicial com 4 grupos (RMSD: 12,29 Å) e (C) conformação inicial com 6 grupos (RMSD: 9,56 Å).

6.3.4 Estudo de caso 4: IOPD

A estrutura da proteína rica em histidina (HPr), mutante Ser46Asp de *Escherichia coli*, cujo código PDB é IOPD, é uma das proteínas de tamanho médio que faz parte do novo conjunto de proteínas teste. Na busca por fragmentos molde no PDB foram excluídas 50 proteínas que

apresentaram relação evolucionária com a proteína alvo: 1OPD, 2JEL, 2EZE, 2EZB, 2EZA, 1GGR, 1J6T, 1VRC, 1HDN, 1POH, 2XDF, 3CCD, 1CM2, 1CM3, 1PFH, 3LE5, 3LFG, 3LE1, 3LE3, 3LNW, 1ZVV, 2RLZ, 1MU4, 2AK7, 1MO1, 1K1C, 1RZR, 2NZU, 2NZV, 2OEN, 1SPH, 1KKL, 2FEP, 1KKM, 1PTF, 1PCH, 1QR5, 3OQM, 3OQN, 1FU0, 2HPR, 2HID, 1QFR, 3IHS, 1TXE, 1JEM, 1KA5, 1Y4Y, 1Y50 e 1Y51. Os 85 resíduos de aminoácidos da sua sequência foram divididos em 81 fragmentos com tamanho de $l = 5$ resíduos que são apresentados na tabela a seguir com a quantidade de moldes obtida.

Tabela 15 – Quantidade de moldes por fragmento da proteína alvo 1OPD.

Fragmentos	CReF 2.0	Fragmentos	CReF 2.0
	Nº moldes		Nº moldes
MFEQE	63	ASAKD	48
FEQEV	53	SAKDL	66
EQEVT	66	AKDLF	53
QEVTI	74	KDLFK	47
EVTIT	80	DLFKL	64
VTITA	57	LFKLQ	31
TITAP	41	FKLQT	20
ITAPN	51	KLQTL	77
TAPNG	73	LQTLG	71
APNGL	57	QTLGL	40
PNGLH	73	TLGLT	43
NGLHT	51	LGLTQ	31
GLHTR	45	GLTQG	48
LHTRP	61	LTQGT	70
HTRPA	40	TQGTV	63
TRPAA	46	QGTVV	45
RPAAQ	67	GTVVT	27
PAAQF	52	TVVTI	53
AAQFV	56	VVTIS	45
AQFVK	70	VTISA	57
QFVKE	57	TISAE	45
FVKEA	44	ISAEG	50
VKEAK	40	SAEGE	52
KEAKG	56	AEGED	96
EAKGF	51	EGEDE	41
AKGFT	57	GEDEQ	55
KGFTS	42	EDEQK	54
GFTSE	48	DEQKA	50
FTSEI	57	EQKAV	57
TSEIT	76	QKAVE	63
SEITV	59	KAVEH	54
EITVT	51	AVEHL	50

Continua na próxima página

Fragmentos	CReF 2.0	Fragmentos	CReF 2.0
	Nº moldes		Nº moldes
ITVTS	37	VEHLV	58
TVTSN	51	EHLVK	63
VTSNG	47	HLVKL	62
TSNGK	60	LVKLM	64
SNGKS	57	VKLMA	48
NGKSA	42	KLMAE	59
GKSAS	43	LMAEL	64
KSASA	56	MAELE	35
SASAK	37		

No processamento do resultado da mineração de dados, para 44% dos resíduos, foi necessário selecionar outro conjunto de ângulos para representá-los. Para todos esses resíduos encontrou-se um grupo com classificação adequada para representá-los na conformação inicial.

A proteína alvo 1OPD apresenta três hélices, formadas por: hélice1 com 12 resíduos, hélice2 com seis resíduos, hélice3 com 14 resíduos, e quatro fitas. As hélices das estruturas preditas foram sobrepostas com as hélices da estrutura experimental, o que demonstrou uma boa predição das hélices, também indicada pelos valores RMSD:

Tabela 16 – Análise do RMSD em Å das estruturas secundárias e conformação inicial das predições da 1OPD.

Estrutura 2D	RMSD 4 Grupos	RMSD 6 Grupos
Hélice 1	0,40	0,46
Hélice 2	0,26	0,39
Hélice 3	1,15	1,10
Fita 1	1,18	1,44
Fita 2	0,49	0,50
Fita 3	0,42	0,35
Fita 4	1,33	1,21
Conformação inicial	25,67	15,19

Nos mapas da Ramachandran da Figura 43 verifica-se uma concentração de pontos na região de “A” que correspondem às hélices. Nas predições (B) e (C) os resíduos ocuparam pontos bem semelhantes aos observados na estrutura experimental (A). O resíduo Lys27 da primeira hélice foi o que apresentou um maior deslocamento, pois nas predições ficou na região de “~I” quando deveria ter ficado em “A”. A hélice3 foi a que apresentou o RMSD mais alto, o que foi ocasionada por resíduos que ficaram mais afastados do ponto onde era esperado estarem. Assim como na

estrutura experimental, os resíduos das fitas concentraram-se na região de “B” em padrão disperso. Das fitas os resíduos mais distorcidos foram a Ala44 que ficou na região de “~p” quando deveria ter ficado em “A” e o Thr59 que deveria ficar em “B” e ficou na região de “b”. Os demais resíduos indicados em (B) e (C) (Ala10, Ser31, Gln57 e Glu68) fazem parte de regiões de volta e são indicativos de que novamente a principal distorção foi destas regiões.

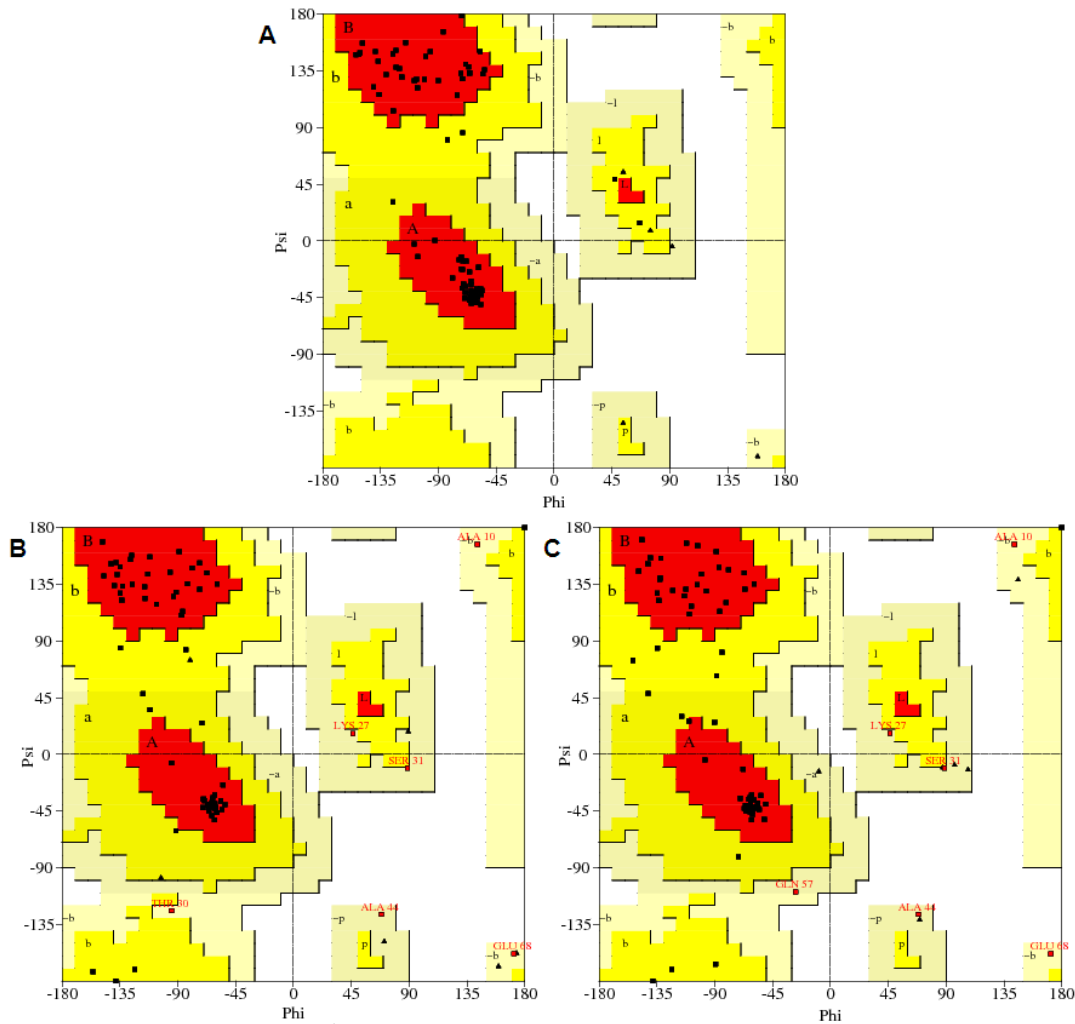


Figura 43 – Mapa de Ramachandran da proteína alvo 1OPD: (A) mapa da estrutura experimental, (B) mapa da estrutura predita sem otimização com base em 4 grupos na etapa de mineração de dados, (C) mapa da estrutura predita sem otimização com base em 6 grupos na mineração.

A Figura 44 mostra que nas predições não houve a formação das folhas, da mesma forma como aconteceu com a proteína alvo 1GB1. Na conformação inicial com quatro grupos (B) não há possibilidade de formação da folha com o enovelamento apresentado. Já na conformação com seis grupos, houve uma melhora no empacotamento, conforme demonstrado pelo valor do RMSD, mas ainda insuficiente para a formação da folha. As hélices estão próximas da estrutura experimental, apesar do posicionamento diferente que é influenciado pela má conformação das regiões de volta. Na conformação (C) a segunda hélice não foi formada na estrutura gráfica apresentada, mas

corresponderia a região de *coil* que aparece enrolada na parte debaixo da representação. Em algumas situações, uma proteína pode apresentar uma representação gráfica diferente conforme o *software* utilizado na visualização. Um exemplo disto é a representação da figura acima que foi gerada através do PyMOL, já uma representação gerada através do Swiss-PdbViewer formou a segunda hélice.

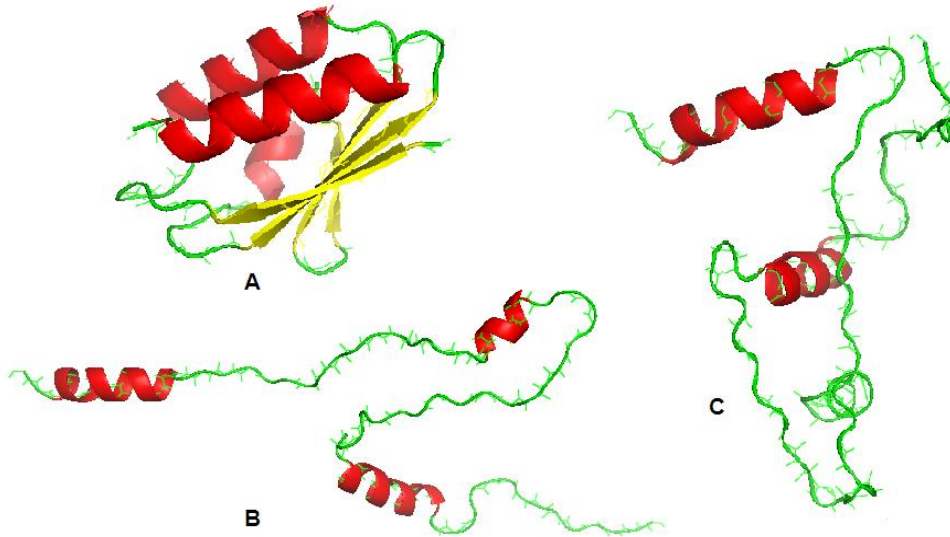


Figura 44 – Comparação da conformação inicial da proteína 1OPD com variação de grupos na mineração dos dados: (A) estrutura experimental da 1OPD, (B) conformação inicial com 4 grupos (RMSD: 25,67 Å) e (C) conformação inicial com 6 grupos (RMSD: 15,19 Å).

A proteína 1OPD foi submetida aos métodos I-Tasser e Robetta e obteve RMSD de 6,35 Å e 1,04 Å, respectivamente. Os valores de RMSD obtidos pelas conformações iniciais das predições estão bem distantes do valor de referência (I-Tasser). Essa diferença é influenciada pelo tamanho da sequência e por suas estruturas secundárias, especialmente as do tipo de folha que o CReF não consegue formar devido à má conformação das voltas e alças.

6.3.5 Demais estudos de caso

Além das quatro proteínas descritas nas seções anteriores, foram testadas as outras quatro proteínas do conjunto inicial e as demais oito proteínas do novo conjunto de teste. O resultado dos testes é apresentado, em formato resumido, na Tabela 17 com as seguintes informações: identificação da proteína alvo, tamanho da sequência, valor do RMSD obtido na execução com quatro e seis grupos na mineração de dados, valor do RMSD obtido na versão inicial do CReF (somente para o conjunto inicial de teste), valor do RMSD da ferramenta I-Tasser e informações

sobre as estruturas secundárias que compõem a proteína alvo. Para cada proteína alvo foram identificados os atributos das hélices e fitas que as compõem, como: quantidade de aminoácidos e valor do RMSD das conformações obtidas com quatro e seis grupos na mineração de dados em relação ao mesmo elemento na estrutura experimental.

Tabela 17 – Resumo dos demais experimentos realizados com a nova versão do CReF com valores de RMSD calculados em Å.

<i>Classe Alfa</i>									
Proteína	Tamanho (aminoácidos)	RMSD 4 Grupos	RMSD 6 Grupos	RMSD CReF 1.0	RMSD I-Tasser	Estruturas secundárias			
						Id	Tamanho	RMSD - 4 gr.	RMSD - 6 gr.
2ERL	40	9,08	12,40	ND	3,34	H1	8	0,26	0,21
						H2	6	0,52	0,57
						H3	12	1,18	0,80
1GAB	53	10,79	10,35	13,57	3,87	H1	2	0,18	0,19
						H2	4	0,37	0,24
						H3	2	0,18	0,17
						H4	7	0,36	0,17
						H5	14	0,86	1,02
1ROP	54	6,48	22,08	5,15	3,10	H1	25	0,84	0,69
						H2	25	3,30	2,15
1UTG	70	19,55	15,41	26,23	12,15	H1	11	0,29	0,95
						H2	10	1,60	1,44
						H3	15	1,33	1,28
						H4	14	0,76	0,91
						H5	3	0,10	0,10
2EZK	99	12,82	15,39	ND	8,92	H1	7	0,25	0,48
						H2	9	0,46	0,67
						H3	11	0,47	0,78
						H4	11	0,62	0,73
						H5	9	0,22	0,58
						H6	4	0,05	0,09
<i>Classe Beta</i>									
Proteína	Tamanho (aminoácidos)	RMSD 4 Grupos	RMSD 6 Grupos	RMSD CReF 1.0	RMSD I-Tasser	Estruturas secundárias			
						Id	Tamanho	RMSD - 4 gr.	RMSD - 6 gr.
1K43	14	9,93	9,91	1,37	ND	F1	3	0,36	0,04
						F2	3	0,25	0,05
1YWJ	41	8,64	10,85	ND	1,53	F1	7	1,77	1,42
						F2	7	0,60	0,72
						F3	3	0,47	0,43
1CSP	67	29,49	22,49	ND	1,62	F1	9	1,67	2,34
						F2	6	1,42	1,51
						F3	5	0,19	0,26
						F4	8	1,98	1,83
						F5	3	0,05	0,04
						F6	3	0,25	0,21

Continua na próxima página

1KSR	100	40,85	23,72	ND	2,64	F1	5	1,47	1,39
						F2	8	1,88	1,88
						F3	5	0,65	0,71
						F4	3	0,26	0,22
						F5	7	1,47	1,13
						F6	8	0,68	0,69
						F7	5	0,87	0,96
Classe Alfa Beta									
Proteína	Tamanho (aminoácidos)	RMSD 4 Grupos	RMSD 6 Grupos	RMSD CReF 1.0	RMSD I-Tasser	Estruturas secundárias			
						Id	Tamanho	RMSD - 4 gr.	RMSD - 6 gr.
1GPT	47	20,76	14,67	ND	1,74	H1	11	0,81	0,98
						F1	7	3,32	3,47
						F2	5	1,39	1,16
						F3	8	2,37	1,53
1CTF	74	17,71	11,45	ND	0,81	H1	11	0,48	0,87
						H2	9	0,41	2,68
						H3	14	0,42	0,71
						F1	6	0,54	0,87
						F2	6	1,61	0,48
						F3	5	0,73	0,65
1ERV	105	17,49	22,10	ND	1,67	H1	10	0,30	0,41
						H2	16	0,70	1,34
						H3	7	2,00	0,65
						H4	11	0,32	0,37
						F1	3	0,41	0,29
						F2	7	3,30	0,73
						F3	7	1,15	1,20
						F4	7	0,79	0,28
						F5	8	2,60	2,09

Legenda: ND = Não Disponível; gr. = grupos; Hn = Hélice n, onde n varia de 1 a 6; Fn = Fita n, onde n varia de 1 a 7.

Na tabela alguns valores estão apresentados em negrito que indicam o melhor RMSD dentre os valores de predição da conformação inicial com o CReF. São considerados os valores da conformação inicial obtida com quatro e seis grupos com a nova versão do CReF e os valores obtidos com a versão inicial (Dorn, 2008).

Os resultados obtidos até a etapa de construção da conformação inicial confirmaram que o problema principal está na conformação de voltas e alças. Isso se comprova pelos bons resultados da predição das estruturas secundárias apresentados acima, transferindo para a região de voltas e alças a responsabilidade pelo altos valores de RMSD das conformações iniciais. Por isso optou-se por pesquisar métodos mais atuais que contribuam para a otimização dessas regiões, em vez de investir tempo para reorganizar a execução do processo usado na versão inicial do CReF (etapa 9 – seção 4.9).

6.4 Desempenho do CReF

O método CReF tem duas tarefas que demandam mais tempo para a sua execução: a execução do BLASTp para cada fragmento e a busca das proteínas molde. A execução do BLASTp é tarefa essencial ao método e é importante que seja executada para toda a proteína alvo submetida a fim de refletir atualizações ocorridas no PDB. Quanto ao download dos arquivos PDB, é possível reduzir esse tempo de execução através da manutenção de um repositório para estes arquivos. Esse repositório tem o objetivo de armazenar os PDBs utilizados na execução do método e pode ser atualizado conforme as proteínas alvo forem submetidas ao CReF ou pode ser criado um processo que atualize o repositório com os PDBs existentes no banco de dados PDB. A manutenção de um repositório local de arquivos PDBs contribui significativamente na redução do tempo de execução do CReF. Em contrapartida, para esta solução, é necessário equipamento capaz de armazenar o grande volume de informações e a existência de um processo que atualize de forma periódica este repositório com novos PDBs.

Podem também causar impactos na performance do método as tarefas com dependência manual que estão associadas à predição das estrutura 2D. Proteínas com 20 ou mais aminoácidos são submetidas à ferramenta Porter para predição, e as demais são submetidas à ferramenta SAM-T08. A submissão ao Porter é feita de forma automática pelo método, já a submissão ao SAM-T08 precisa ser feita manualmente pelo usuário. Isso pode ser evitado se já existir no repositório de predições de estrutura 2D um arquivo com a predição buscada. Caso contrário, será necessária a intervenção manual para submissão ao SAM-T08 e para disponibilizar o resultado de ambas as predições, que são enviadas por e-mail, para uso do método.

A implementação da execução automática de todas as etapas do método contribuiu significativamente para uma execução mais eficiente e rápida do método. Para uma proteína grande e considerando a necessidade do download de um conjunto de PDBs, a conformação inicial foi obtida através da execução da nova versão após 1 hora e 40 minutos. Desta forma é possível afirmar que a nova versão do CReF mantém a boa performance da versão inicial inclusive em plataforma de baixo custo e com facilidade na instalação.

6.5 Refinamento de conformações iniciais

Em substituição à etapa 9, que existia na versão inicial, criou-se um processamento pós-CReF que será responsável pelo refinamento das estruturas preditas (conformações iniciais). A estratégia adotada será ilustrada pelo refinamento da estrutura predita da proteína 1ZDD, referenciada aqui por 1ZDD_P. Esta proteína foi escolhida por ser a mais simples entre todas aquelas investigadas neste trabalho.

Na seção 5.2 discutiu-se sobre o problema da otimização de voltas e alças em proteínas e do grande desafio que este processo representa. Como visto na seção 6.3.1, que trata do estudo de caso da 1ZDD, o maior problema na predição foi a conformação da alça que conecta as hélices 1 e 2 (Figura 36B e 36C). Os valores de RMSD mostraram que a estrutura da cadeia principal das hélices, individualmente, eram muito similares às da proteína experimental 1ZDD (Tabela 10). Entretanto, o RMSD global, da proteína inteira, apresentou para a conformação inicial com quatro grupos na mineração um valor de 9,82 Å (Figura 36B) e para a conformação com seis grupos apresentou um RMSD de 6,64 Å, o que representa uma melhoria em relação à anterior. Como as hélices foram bem formadas, a conformação da alça que as conecta foi responsável pela disposição inadequada dessas hélices no espaço 3D, o que faz com que a estrutural global da 1ZDD_P seja tão diferente da experimental 1ZDD (Figura 36B e 36A, respectivamente). Apesar do valor elevado do RMSD, na visualização ambas as conformações iniciais aproximam-se bem da estrutura experimental. Por isso, são denominadas estruturas 3D aproximadas preditas.

A ideia subjacente ao método CReF é que as estruturas aproximadas preditas sejam boas o suficiente para serem submetidas a protocolos de refinamento pelas técnicas de simulação pela dinâmica molecular (DM) (Dorn e Norberto de Souza, 2008). Em uma simulação da DM, as equações clássicas de movimento que governam a evolução temporal e microscópica de um sistema de muitos corpos (átomos em uma macromolécula, por exemplo) são resolvidas numericamente e sujeitas a condições periódicas apropriadas à geometria e simetria do sistema (van Gunsteren e Berendsen, 1990). Portanto, a metodologia da DM é fundamentada nos princípios da Mecânica Clássica e pode fornecer uma visão microscópica do comportamento dinâmico de átomos individuais que constituem um sistema, como uma proteína. Como resultado, uma simulação por DM produz um conjunto de conformações (*ensemble*) da proteína em função do tempo. A partir do *ensemble* em equilíbrio, o valor médio de parâmetros termodinâmicos como pressão, temperatura, volume, calor específico, pode ser calculado, assim como parâmetros estruturais, incluindo o raio de giro e a estrutura média da proteína (Norberto de Souza e Ornstein, 1999).

6.5.1 Detalhes da simulação por DM da 1ZDD_P

Apesar da estrutura predita da 1ZDD_P, na Figura 36C, representar o melhor resultado para este estudo de caso, o modelo na Figura 36B foi escolhido como conformação inicial por estar mais distante da estrutura desejada. Essa conformação apresenta-se, a princípio, como um maior desafio para os protocolos de refinamento pelo método de simulação pela DM, os quais ainda precisam ser desenvolvidos.

O programa usado para a realização das simulações por DM da 1ZDD_P foi o módulo SANDER do pacote AMBER 9 fazendo uso do campo de força ff99SB (Roe et al., 2007). Nas simulações da DM para o refinamento da estrutura predita com quatro grupos na mineração 1ZDD_P (Figura 36B), o ambiente aquoso (solvente) é representado de forma implícita, como um dielétrico contínuo, utilizando o formalismo denominado *Generalized Born* (GB) (Bashford e Case, 2000; Jayaram et al., 1998). Se a adição do solvente ocorresse de forma explícita, isso aumentaria de maneira considerável o tempo de simulação. Com o solvente representado implicitamente, o efeito hidrofóbico do sistema é calculado atribuindo-se penalidades para a exposição de resíduos hidrofóbicos e compensações quando os resíduos expostos são polares. A energia de solvatação é calculada através da mudança de área total do sistema acessível ao solvente. A estabilidade da simulação, para o propósito deste trabalho, foi avaliada pelo monitoramento do RMSD da cadeia principal das proteínas (apenas os átomos C α) com relação à estrutura experimental inicial desejada.

Simulações por DM analisam milhares de átomos e calculam seus movimentos, exigindo recursos computacionais de alto desempenho. As simulações deste trabalho foram realizadas em um PC QuadCore de 2,4 GHz e 4 GB RAM. Cada uma delas demanda dias, semanas ou até meses, dependendo do tamanho da proteína e do tempo de duração. Nos testes realizados o tempo médio para cada simulação de 1.000,0 ps (1,0 picossegundo = 10^{-12} segundos), para a proteína 1ZDD_P com 569 átomos, foi de 12 horas.

Sete simulações por DM foram realizadas a diferentes temperaturas, incluindo 281 K, 298,16 K e 325 K, e com diferentes escalas de tempo, todas começando da estrutura predita 1ZDD_P (Figura 36B). A estrutura experimental, determinada pelo método de NMR (código PDB 1ZDD), foi adotada como estrutura de referência e, portanto, parâmetro de comparação para os resultados. A conformação inicial para a simulação da DM foi gerada com o módulo tleap do programa AMBER 9. Como as hélices foram previstas com alta acurácia (RMSDs muito pequenos, próximos de zero), suas conformações foram mantidas, por meio da introdução de restrições harmônicas às coordenadas internas (ângulos diedrais) que as definem. Assim, durante as simulações todo o sistema pode se mover, mas as conformações das hélices ficaram restritas

àquelas previstas pelo CReF. Isso fez com que, neste protocolo, o espaço conformacional acessível à alça que conecta as hélices fosse explorado de forma mais eficiente.

6.5.2 Resultados do refinamento da 1ZDD_P

Dentre os diferentes protocolos das sete simulações por DM testados, apresenta-se aqui aquele que forneceu o melhor resultado de refinamento da 1ZDD_P. Os principais parâmetros deste protocolo foram: temperatura final de 325 K, restrições das conformações das hélices α (H1 e H2) previstas. Ainda, o resíduo de histidina (HIS14) teve liberdade para se mover como se fosse parte da alça que conecta as hélices. O tempo total de simulação da 1ZDD_P foi de 1.000 ps, com conformações instantâneas salvas a cada 0,5 ps. Assim, 2.000 conformações foram empregadas na análise do refinamento da 1ZDD_P.

Neste refinamento o objetivo foi fazer com que a estrutura prevista 1ZDD_P se aproximasse o máximo possível da estrutura desejada, a estrutura experimental 1ZDD. Uma métrica bastante comum para medir essa similaridade estrutural é o RMSD (Fiser et al., 2000), o qual pode ser local ou global. A Figura 45 mostra o RMSD global e local, medido em Å, da trajetória dinâmica da 1ZDD_P com relação a estrutura experimental, como função do tempo de simulação. O gráfico menor inserido nesta figura mostra apenas os primeiros 100,0 ps para que detalhes das transições conformacionais ocorridas neste intervalo de tempo sejam visualizadas mais claramente. Como se pode observar, neste intervalo ocorreram as principais transições conformacionais que permitiram à 1ZDD_P, durante a simulação, se aproximar da 1ZDD, atingindo um RMSD global de 1,63 Å (curva G) em 100,0 ps. Quando se analisa apenas a sobreposição das hélices α , o RMSD é bem inferior e igual a 1,13 Å (curva H). O RMSD local (curva A) foi calculado apenas para a alça que conecta as hélices H1 e H2.

Pode-se notar que a curva A flutua durante quase toda a simulação em torno de um valor um pouco acima de 2,00 Å. Porém, o gráfico inserido mostra que há uma mudança conformacional na alça que vai de 3,0 ps a 15,0 ps. Nestes dois instantes o RMSD é de 2,50 Å aproximadamente. No interior desta região, em 5,5 ps e em 11,5 ps, o RMSD da alça atinge os seus menores valores, 0,68 Å e 0,65 Å, respectivamente. Isso significa que a conformação da alça da 1ZDD_P é, nestes instantes, praticamente idêntica a da estrutura alvo, à da 1ZDD (Figuras 46B e 46C). É possível observar que nesta faixa de intervalo em que ocorre a transição conformacional na alça, os RMSDs globais (curvas G e H) atingem seus maiores valores de 11,02 Å em 10,5 ps. Isso é claramente ilustrado quando se observa as posições relativas das hélices nas Figuras 46B e 46C.

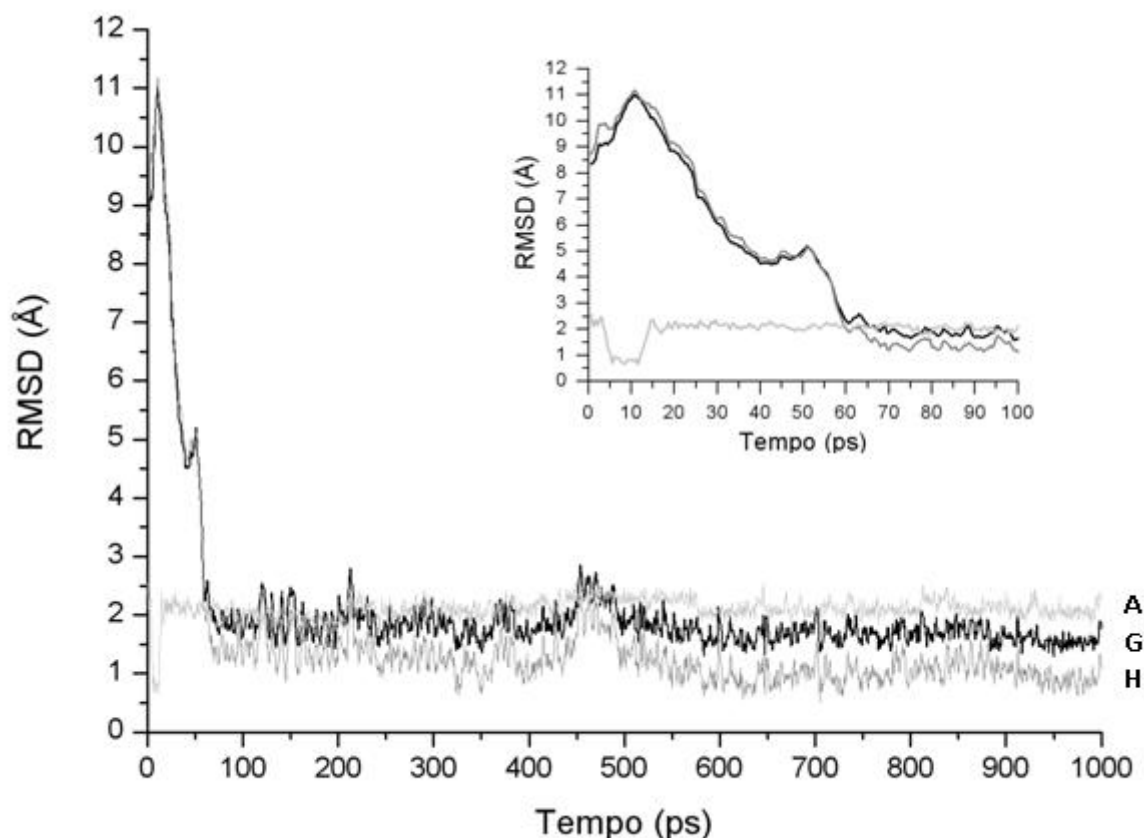


Figura 45 – RMSD da trajetória dinâmica, durante o refinamento da estrutura predita 1ZDD_P em relação à estrutura experimental 1ZDD, como função do tempo de simulação. Apenas os átomos da cadeia principal ($C\alpha$ átomos) foram utilizados nos três cálculos de RMSD mostrados na figura. A curva G (linha preta) representa o RMSD da sobreposição dos resíduos de aminoácidos MET3 a GLU32 e trata-se de uma medida de RMSD global. A curva H (linha cinza escuro) representa o RMSD da sobreposição dos resíduos das hélices α (MET3 a LEU13 e GLU20 a ASP32). A curva A (linha cinza claro) corresponde ao RMSD da alça que conecta as duas hélices, e formada pelos resíduos HIS14 a ASN19. O gráfico menor inserido nesta figura ilustra, de forma detalhada, os eventos dinâmicos que ocorreram nos 100,0 ps iniciais da simulação por DM e que não podem ser visualizados de forma clara no gráfico da trajetória dinâmica completa. A transição conformacional espontânea que ocorre na alça no intervalo de 3,0 ps a 15,0 ps é fundamental para o sucesso deste protocolo de refinamento. O comportamento dos RMSDs globais (curvas G e H) mostra que a simulação da 1ZDD_P começa a convergir para a estrutura desejada, a 1ZDD, já antes dos primeiros 100,0 ps.

A Figura 46 mostra uma parte representativa da trajetória dinâmica da 1ZDD_P durante o seu refinamento. Verifica-se, a partir da Figura 46D à Figura 46H, como o protocolo de refinamento foi efetivo em promover as mudanças conformacionais para que em 950,0 ps (Figura 46H) a estrutura predita 1ZDD_P atingisse um RMSD de apenas 1,29 Å e 0,64 Å para as curvas G e H, respectivamente, na Figura 45. Estes valores significam que o enovelamento da cadeia principal da 1ZDD_P é praticamente idêntico ao da proteína alvo.

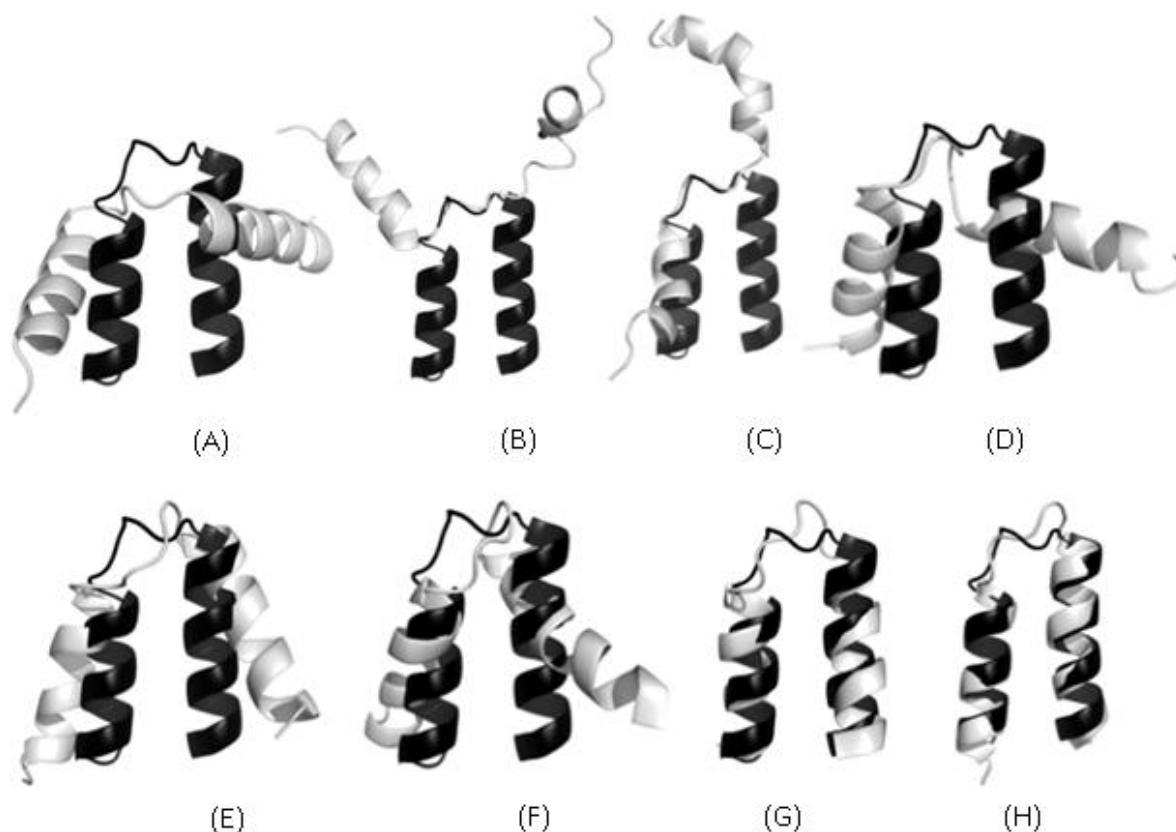


Figura 46 – Representação do tipo *Ribbons* da cadeia principal ($C\alpha$ átomos) da estrutura experimental da proteína 1ZDD (preto) e de conformações representativas da trajetória dinâmica durante o refinamento da estrutura predita 1ZDD_P (cinza). A orientação das cadeias é da esquerda para a direita, indo da região N-terminal à C-terminal. A sobreposição da trajetória dinâmica das estruturas inclui os resíduos de aminoácidos MET3 a ASP32, exceto para as figuras (B) e (C), onde apenas a alça (HIS14 a ASN19) foi sobreposta. A trajetória dinâmica (cinza) é composta pelas seguintes conformações instantâneas: (A) 0,5 ps, (B) 5,5 ps, (C) 11,5 ps, (D) 25,0 ps, (E) 40,0 ps, (F) 50,0 ps, (G) 60,0 ps e (H) 950,0 ps de um total de 1.000 ps (Figura 45). Figura gerada com PyMOL.

A análise das estruturas gráficas justifica o baixo valor de RMSD obtido pela conformação refinada resultante do protocolo de refinamento. Na Figura 47 observa-se a conformação refinada da 1ZDD sem sobreposição com a experimental e que corresponde à representação (H) na Figura 46.

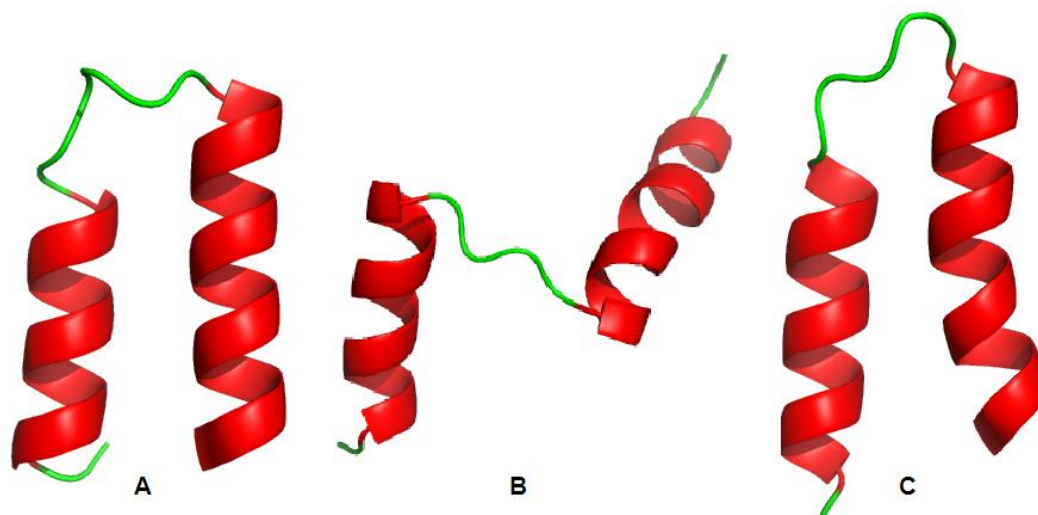


Figura 47 – Comparação de conformações predita e refinada da proteína 1ZDD: (A) estrutura experimental da 1ZDD, (B) conformação inicial com 4 grupos na mineração de dados (RMSD: 9,82 Å) e (C) conformação predita refinada (RMSD: 1,29 Å).

6.6 Resumo do capítulo

O capítulo 6 apresentou o conjunto de proteínas teste que foi submetido ao método. Esse conjunto foi composto pelas seis proteínas testadas por Dorn (2008) e pelas dez proteínas selecionadas para compor um novo grupo, as quais representam diferentes classes e tamanhos de proteínas. Também foram listadas as ferramentas e os recursos usados para compor o método e as ferramentas empregadas nas análises e na representação gráfica das proteínas e suas conformações preditas.

Para o conjunto de proteínas teste foram realizados experimentos considerando quatro e seis grupos na mineração de dados para comparação de resultados. Por meio de representações no mapa de Ramachandran, de representações gráficas das conformações e dos valores de RMSD local (por estrutura secundária) e global das predições para a conformação inicial foram analisados os experimentos com as proteínas alvo 1ZDD, 1GB1, 1C5A e 1OPD. Os resultados dos testes com as 12 proteínas restantes foram apresentados de forma resumida e com a indicação do melhor valor do RMSD obtido para cada uma.

Além dos experimentos, analisou-se o desempenho da nova versão do CReF. O uso de repositórios para os arquivos PDBs e predições de estrutura secundária contribuiu para um desempenho ainda melhor. O único entrave está associado à atividade manual associada à predição de estrutura secundária quando a proteína alvo não for encontrada no repositório. Em contrapartida,

a execução automática traz alertas sobre as tarefas em andamento ou em falta e elimina a dependência do usuário para uma execução rápida e eficiente.

Este capítulo também tratou sobre o refinamento de conformações iniciais ilustrando sua aplicação à conformação inicial da 1ZDD obtida com quatro grupos na mineração de dados (1ZDD_P). Como as hélices foram bem preditas optou-se por manter sua conformação a partir de restrições de coordenadas internas, procurando permitir que todo o sistema pudesse se mover, inclusive os resíduos que precedem e sucedem a alça. Foram realizados diversos testes com variação dos parâmetros até chegar-se ao protocolo que obteve o melhor resultado. O resultado obtido apresentou-se promissor para um método de refinamento de estruturas 3D de proteínas.

7. Considerações Finais

A importância da Bioinformática é indiscutível e comprovada pelo aumento exponencial das bases de dados biológicos. Diversos bancos de dados e inúmeras ferramentas para análise bioinformática estão disponíveis para domínio público com o objetivo principal de permitir a geração de conhecimento a partir da análise e transformação dos dados (Galperin e Fernández-Suárez, 2011). Este tem sido um desafio bastante instigante, pois a partir do conhecimento se realizam descobertas que permitem entender essa máquina excepcional chamada vida. Uma das importantes fontes de informação para o entendimento da vida são as proteínas e os processos relacionados a elas. Para compreendê-los, uma etapa fundamental é conhecer as funções das proteínas, o que é possível fazer a partir de sua estrutura terciária. Para auxiliar na construção deste conhecimento existem diferentes métodos de predição que buscam traduzir a informação contida na sequência de uma proteína para construir uma possível estrutura tridimensional. Confirmando a importância destes métodos, as diferentes metodologias de predição reúnem-se no CASP (Mariani et al., 2011) para serem avaliadas e comparadas entre si. Através do CASP é possível acompanhar os resultados de uma série de métodos que tem realizado boas predições 3D de proteínas atualmente. A maior parte destes métodos tem a sua disposição infraestrutura e recursos de alta tecnologia.

Juntamente com o intuito de aprender sobre o problema da predição e de contribuir nesta área, surgiu a possibilidade de trabalhar com o método CReF de predição aproximada. Na contramão dos métodos mais expressivos, CReF se diferenciava fazendo uso de uma plataforma de baixo custo, entretanto apresentava dificuldades na usabilidade e na predição. Este trabalho de dissertação propôs-se, então, a implementar melhorias no método, a partir do conhecimento mais aprofundado da ferramenta. O primeiro passo para isso foi a simulação dos experimentos da versão inicial, o que já se apresentou como uma dificuldade. Além das alterações previamente identificadas, outras alterações mostraram-se necessárias para que o CReF pudesse ser executado fazendo uso de versões mais atuais das ferramentas acessórias (descritas na seção 6.2), o que demandou mais tempo do que o planejado para esta etapa.

O conjunto de proteínas teste foi composto por proteínas de diferentes tamanhos e de diferentes classes estruturais a fim de permitir uma melhor avaliação do método. Os resultados dos experimentos realizados com a nova versão do CReF são resumidos na Tabela 18.

Tabela 18 – Resultados dos experimentos com a nova versão do CReF. Em negrito estão indicados os melhores valores de RMSD para cada proteína alvo.

Proteína	Tamanho (aminoácidos)	RMSD (Å) 4 Grupos	RMSD (Å) 6 Grupos	RMSD (Å) CReF 1.0	RMSD (Å) I-Tasser
<i>Classe Alfa</i>					
1ZDD	34	9,82	6,64	5,51	1,57
2ERL	40	9,08	12,40	ND	3,34
1GAB	53	10,79	10,35	13,57	3,87
1ROP	54	6,48	22,08	5,15	3,10
1UTG	70	19,55	15,41	26,23	12,15
1C5A	73	12,29	9,56	ND	5,11
2EZK	99	12,82	15,39	ND	8,92
<i>Classe Beta</i>					
1K43	14	9,93	9,91	1,37	ND
1YWJ	41	8,64	10,85	ND	1,53
1CSP	67	29,49	22,49	ND	1,62
1KSR	100	40,85	23,72	ND	2,64
<i>Classe Alfa Beta</i>					
1GPT	47	20,76	14,67	ND	1,74
1GB1	56	21,05	14,85	12,31	5,52
1CTF	74	17,71	11,45	ND	0,81
1OPD	85	25,67	15,19	ND	6,35
1ERV	105	17,49	22,10	ND	1,67

Não é possível estabelecer uma relação direta entre os resultados da versão inicial do CReF (coluna “RMSD CReF 1.0”) e a atual versão (colunas “RMSD 4 Grupos” e “RMSD 6 Grupos”). Uma diferença importante está no conjunto de proteínas molde considerado na predição por cada versão. Como não existia um histórico do conjunto de moldes usado pela versão inicial, a nova versão não pode simular uma predição sob o mesmo conjunto de moldes. Como consequência do aumento crescente da quantidade de proteínas armazenadas no PDB, o conjunto de moldes encontrado na nova versão foi bem maior. Na Tabela 18 observa-se que para quatro das seis proteínas do conjunto inicial de teste, o RMSD da versão 1.0 foi melhor do que o da nova versão. A partir desses dados pode-se levantar a hipótese de que uma maior diversidade de proteínas pode causar uma maior dificuldade na predição. Em suma, conforme o número de proteínas aumenta, a predição a partir de moldes também se tornará mais complexa e precisará evoluir suas estratégias ou estabelecer critérios de exclusão mais rígidos.

Para as proteínas 1GAB e 1UTG, a nova versão do CReF apresentou resultados melhores do que a versão 1.0, sendo que para a 1UTG a melhora foi significativa. Possivelmente, isso deve ter

sido proporcionado pelo conjunto de moldes considerado, já que as outras proteínas da classe alfa ou da mesma categoria de tamanho não apresentaram o mesmo comportamento. Considerando os resultados das proteínas alvo pertencentes ao conjunto inicial de testes (1ZDD, 1GAB, 1ROP, 1UTG, 1K43 e 1GB1) e que foram submetidas a ambas as versões, é possível afirmar que a nova versão mantém o mesmo desempenho da anterior. As diferenças encontradas entre as predições da conformação inicial foram ocasionadas por influência dos moldes (seja para melhor ou pior).

A análise das predições de todos os experimentos realizados aponta que para vários casos houve uma diferença importante entre a predição com quatro grupos e seis grupos. Para a maioria (11 proteínas), os melhores valores de RMSD foram de predições realizadas com seis grupos na etapa de mineração de dados (Tabela 18). Apenas para proteínas grandes seria interessante considerar outros alvos para avaliar se melhores valores de RMSD com quatro grupos são uma tendência para esta categoria ou não. Com os resultados obtidos e com a amostragem considerada, não é possível afirmar que existam tendências que se reflitam em valores de RMSD e que estejam associadas a categorias de proteínas (classe ou tamanho).

Como esperado, as predições da nova versão apresentaram uma grande diferença em relação aos valores obtidos pelo I-Tasser, isso porque as predições são de conformações iniciais. Serão comparáveis ao I-Tasser os valores obtidos pelas predições após o processo de refinamento, por isso não é possível avaliar compatibilidade nessas condições. De forma geral, os valores de RMSD do I-Tasser também não apresentaram tendências relacionadas à classe ou tamanho da proteína. Por exemplo, na Tabela 18 observa-se que o melhor resultado do I-Tasser foi da 1CTF (proteína média da classe alfa beta) e o pior, o da 1UTG (proteína média da classe alfa). Na comparação com o I-Tasser, o CReF apresenta uma vantagem, pois consegue prever proteínas (polipeptídeo) muito pequenas, como a 1K43.

Os resultados dos experimentos demonstraram que as alterações realizadas no método não implicaram em melhoria da conformação, mas mantiveram a boa predição de estruturas secundárias. Além disso, comprovaram a dificuldade de se prever as conformações de regiões de voltas e alças. Conforme já indicado pelos experimentos iniciais, um processo de refinamento dessas regiões é essencial para implicar em melhoria da qualidade das predições. Em contrapartida, houve grandes avanços em relação a usabilidade da ferramenta por meio da sua automatização. Nessa nova versão o usuário poderá: parametrizar os diretórios a serem usados, criar repositórios de arquivos PDBs e predição 2D, executar o processo com apenas uma linha de comando, encontrar apoio em um manual para a instalação e para a execução da ferramenta. Além disso, na forma em que foi implementada, a nova versão do CReF possibilita a proveniência dos experimentos realizados e um acompanhamento por tarefa executada pela ferramenta com mensagens amigáveis ao usuário. O CReF mantém-se uma ferramenta de boa performance e executável em plataforma de

baixo custo que tornou-se parametrizável, de execução simples e que ao final permite ao usuário uma visualização gráfica da conformação inicial (estrutura 3D aproximada predita). As melhorias técnicas implementadas permitem ao CReF estar bem mais próximo de uma disponibilização para acesso público. Esta disponibilização poderá ser sustentável a partir do momento que um processo de refinamento tenha sido claramente definido e implementado.

O fator surpresa deste trabalho foi a dificuldade em definir uma boa estratégia para melhorar a conformação das regiões de voltas e alças. Uma das estratégias consideradas foi o uso de bibliotecas de alças, mas eram limitadas a estas regiões não permitindo que trechos próximos se movimentassem em busca da conformação mais adequada no espaço 3D. Então, optou-se pela simulação por Dinâmica Molecular para a modelagem dessas regiões. Em virtude da limitação de tempo, este trabalho contou com apoio do conhecimento de um especialista no desenho de um protocolo de refinamento. Uma série de fatores químicos e físicos precisaram ser combinados e ajustados para se obter o resultado esperado que, neste caso, era a conformação mais próxima o possível da estrutura experimental. O resultado obtido com o refinamento da conformação inicial que usou quatro grupos na mineração de dados para a 1ZDD foi excelente (1,29 Å). Este resultado também ficou muito próximo do RMSD do I-Tasser (1,57 Å). A conformação refinada foi obtida aos 950,0 picossegundos (ps) da simulação e para o cálculo do RMSD foram considerados apenas os átomos Ca dos resíduos da posição 3 a 32. Acredita-se que um bom resultado também seja obtido no refinamento da conformação inicial com seis grupos da 1ZDD, para este caso espera-se que o posicionamento correto seja atingido em menos tempo.

Para proteínas com apenas uma estrutura irregular (como: 1ZDD, 1ROP, 1K43), a versão inicial do CReF, precisou de, aproximadamente, 24 a 48 horas para que o algoritmo de otimização de voltas atingisse o critério de parada (Dorn, 2008) para, então, compor a conformação final da estrutura predita. No refinamento por DM, o tempo médio para cada simulação foi de 12 horas. Portanto, após a definição do protocolo mais adequado a ser usado, o refinamento de uma conformação inicial é executado em torno desse tempo, o que representa uma melhora expressiva do tempo de execução em relação à versão inicial. A melhora também é perceptível em relação à conformação final predita, pois a estrutura 3D final predita da versão inicial obteve um RMSD de 5,00 Å (Dorn, 2008), enquanto a estrutura resultante do processo de refinamento obteve RMSD de 1,29 Å, a figura a seguir apresenta estas estruturas.

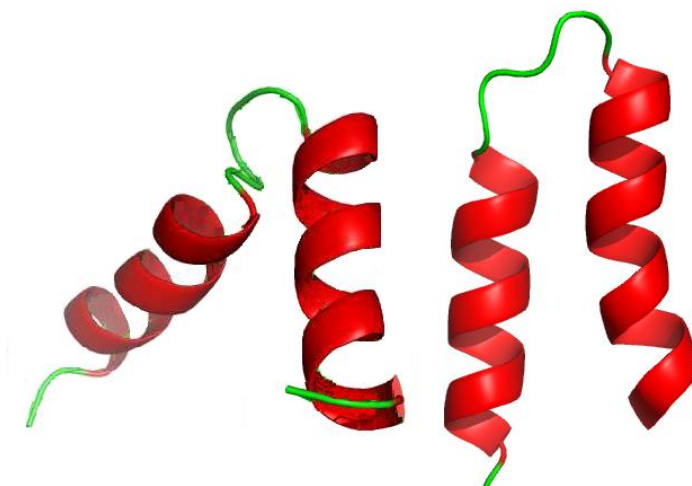


Figura 48 – Estrutura 3D final predita pelo CReF: (A) estrutura 3D após a etapa de otimização da versão inicial (RMSD: 5,00 Å), (B) estrutura 3D refinada da nova versão (RMSD: 1,29 Å).

Esse bom resultado é fator motivador para ampliar o teste do protocolo às demais proteínas alvo consideradas por este trabalho. Após um teste mais amplo será possível definir um protocolo capaz de obter boas conformações refinadas para diferentes classes de proteínas, como as testadas nesta dissertação. Um bom protocolo precisará obter o melhor resultado possível para os mais diferentes tipos de proteína, o que claramente define-se como um trabalho extenso. Para a proteína 1ZDD, o protocolo demonstrou ser possível obter uma conformação de alça excelente (0,68 Å e 0,65 Å observadas nas Figuras 46B e 46C, respectivamente), isto é, muito próxima da estrutura experimental. Em contraposição a isto, a conformação das hélices obteve o pior valor de RMSD no mesmo ponto da simulação (gráfico menor na Figura 45). Isso indica que há uma barreira a ser rompida para que as hélices possam assumir a conformação adequada quando a região de alça está na conformação esperada. Esta barreira energética foi vencida por meio do uso de uma temperatura de 325 K, bem acima da temperatura (281 K) na qual a estrutura experimental foi resolvida.

A teoria diz, mas a prática comprovou que prever estrutura tridimensional de proteínas é uma tarefa complexa pela influência de um conjunto muito grande de fatores. Contudo, ao lembrar-se da instigante máquina que é a vida, que nos surpreende pela sua perfeição e que nos aflige por seus mistérios, é razoável que o segredo a ser descoberto a partir das proteínas não fosse revelado de forma tão fácil. Entender a vida é uma das missões do ser humano, assim como vencer obstáculos não deve ser uma opção, mas deve ser um caminho para chegar às respostas.

7.1 Principais contribuições

As principais contribuições desta dissertação foram:

- Automatização da ferramenta e possibilidade de parametrizações.
- Interação automática com o sítio de predição de estrutura secundária do método Porter.
- Análise de diferentes métodos de predição de estrutura secundária.
- Agrupamento das tuplas em quatro ou seis grupos na mineração de dados.
- Protocolo de refinamento com excelente resultado para a estrutura 3D aproximada predita da 1ZDD.

7.2 Trabalhos futuros

Apesar dos resultados encorajadores desta dissertação, trabalhos futuros poderão abordar os seguintes problemas:

- Investigação do impacto que um número maior de moldes causa na predição e estudo de novos critérios para seleção das proteínas moldes.
- Aplicação de estratégias mais elaboradas de mineração de dados.
- Extensão do protocolo de refinamento para as outras classes de proteínas investigadas neste trabalho.
- Automatização do protocolo de refinamento.
- Integração do protocolo de refinamento ao CReF.
- Disponibilização da ferramenta via webApp.

Referências

- Al-Karadaghi, S. Protein Structures and Structural Bioinformatics Guide. Centre for Molecular Protein Science of Lund University, Lund – Sweden, Capturado em: <http://www.proteinstructures.com/index.html>, Novembro 2011.
- Altschul, S. F.; Madden, T. L.; Schäffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17): 3389–3402, 1997.
- Bashford, D.; Case, D. A. Generalized Born models of macromolecular solvation effects. *Annual Review of Physical Chemistry*, 51: 129–152, 2000.
- Answers.com. Amino acid: Peptide bond formation. Capturado em: <http://www.answers.com/topic/amino-acid>, Novembro 2011.
- Banner, D. W.; Kokkinidis, M.; Tsernoglou, D. Structure of the ColE1 rop protein at 1.7 Å resolution. *Journal of Molecular Biology*, 196(3): 657–675, 1987.
- Baxevanis, A. D.; Ouellette, B. F. F. Bioinformatics: a practical guide to the analysis of genes and proteins. John Wiley and Sons, Inc., New Jersey, EUA, 3ª edição, 2005.
- Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bath, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The protein data bank. *Nucleic Acids Research*, 8(1): 235–242, 2000.
- Berman, H.; Henrick, K.; Nakamura, H.; Markley, J. L. The worldwide protein data bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Research*, 35(Database Issue): 301–303, 2006.
- Biochemistry 462a. Proteins: Secondary Structure and Fibrous Proteins. Department of Biochemistry and Molecular Biophysics, The University of Arizona. Capturado em: http://www.biochem.arizona.edu/classes/bioc462/462a/NOTES/Protein_Structure/secondary_structure.htm, Novembro 2011.
- Branden, C.; Tooze, J. Introduction to protein structure. Garland Publishing Inc., New York, EUA, 2ª edição, 1998.
- Cahill, S.; Cahill, M.; Cahill, K. On the kinematics of protein folding. *Journal for Computational Chemistry*, 24(11): 1364–1370, 2003.

- Canutescu, A.; Dunbrack, R. Cyclic coordinate descent: A robotics algorithm for protein loop closure. *Protein Science*, 12: 963–972, 2003.
- Case, D.A.; Darden, T.A.; Cheatham III, T.E.; Simmerling, C.L.; Wang, J.; Duke, R.E.; Luo, R. e et al. AMBER 9. University of California, San Francisco, 2006.
- Champe, P. C.; Harvey, R. A.; Ferrier, D. R. *Bioquímica Ilustrada*. Tradução Carla Dalmaz et al., 3ª Edição, Porto Alegre: Artmed, 544p, 2006.
- Chapman, B.; Chang, J. Biopython: python tools for computational biology. *ACM SIGBIO Newsletter*, 20(2): 15–19, 2000.
- Cheng, J.; Randall, A.; Sweredoski, P.; Baldi, M. SCRATCH: a protein structure and structural feature prediction server. *Nucleic Acids Research*, 33: 72–76, 2005.
- Chou, P. Y.; Fasman, G. D. Empirical predictions of protein conformation. *Annual Review of Biochemistry*, 47: 251-76, 1978.
- Cochrane, G. R.; Galperin, M. Y. The 2010 Nucleic Acids Research database issue and online database collection: a community of data resources. *Nucleic Acids Research*, 38: D1–D4, 2010.
- Cohen, J. *Bioinformatics - An Introduction for Computer Scientists*. ACM Computing Surveys, 36(2): 122-158, 2004.
- Creighton, T. E. *Proteins: Structures and Molecular Properties*. W. H. Freeman, New York, 2ª edição, 1993.
- da Silveira, N. F. *Bioinformática Estrutural Aplicada ao Estudo de Proteínas Alvo do Genoma do Mycobacterium tuberculosis*. (Tese de Doutorado em Biofísica Molecular – Universidade Estadual Paulista - Júlio de Mesquita Filho/UNESP), 2005.
- Dayhoff, M. O.; Schwartz, R. M.; Orcutt, B. C. A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure*, 5(3): 345–352, 1978.
- Delano, W. L. *The PyMOL molecular graphics system*. Delano Scientific, San Carlos, CA, USA, 2002.
- Dong, Q.; Zhou, S.; Guan, J. A new taxonomy-based protein fold recognition approach based on autocross-covariance transformation. *Bioinformatics*, 25(20): 2655–2662, 2009.
- Dorn, Márcio. Uma proposta para a predição computacional da estrutura 3D aproximada de polipeptídeos com redução do espaço conformacional utilizando análise de intervalos.

- Biblioteca Central da PUCRS - Irmão José Otão (Dissertação de Mestrado em Ciência da Computação - FACIN/PUCRS), Porto Alegre, 2008.
- Dorn, Márcio; Norberto de Souza, O. CReF: a central-residue-fragment-based method for predicting approximate 3-D polypeptides structures. In: 23th ACM Symposium on Applied Computing, Fortaleza/CE. Proceedings of the 23th ACM Symposium on Applied Computing. New York: ACM, Inc., 2: 1261-1267, 2008.
- Dorn, Márcio; Norberto de Souza, O. Mining the Protein Data Bank with CReF to predict approximate 3-D structures of polypeptides. *International Journal of Data Mining and Bioinformatics*, 4(3): 281-299, 2010.
- Dunbrack, R. L.; Karplus, M. Backbone-dependent rotamer library for proteins: application to side-chain prediction. *Journal of Molecular Biology*, 230(2): 543–574, 1993.
- Efimov, A. V. Standard structures in proteins. *Progress in biophysics and molecular biology*, 60(3): 201–239, 1993.
- Fiser, A.; Do, R. K.; Sali, A. Modeling of loops in protein structures. *Protein Science*, 9(9): 1753–1773, 2000.
- Frishman, D.; Frishman, P. Incorporation of non-local interactions in protein secondary structure prediction from the amino acid sequence. *Protein Engineering*, 9(2): 133–142, 1996.
- Galperin, M. Y.; Fernández-Suárez, X. M. The 2012 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection. *Nucleic Acids Research*, 40(D1): D1–D8, 2011.
- Gibas, G.; Jambeck, P. *Desenvolvendo bioinformática*. Editora Campus - O'Reilly, Rio de Janeiro, 1ª Edição, 2001.
- Gopalakrishnan, K.; Sheik, S. S.; Sekar, K. Ramachandran Plot 2.0. Centre of Excellence in Structural Biology and Bio-computing, Indian Institute of Science, India. Capturado em: <http://dicsoft1.physics.iisc.ernet.in/rp/index.html>, Setembro 2010.
- Gronenborn, A. M.; Filpula, D. R.; Essig, N. Z.; Achari, A.; Whitlow, M.; Wingfield, P. T.; Clore, G. M. A novel, highly stable fold of the immunoglobulin binding domain of streptococcal protein G. *Science*, 253(5020): 657–661, 1991.
- Guex, N.; Peitsch, M. C. SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling. *Electrophoresis*, 18: 2714–2723, 1997.

- Henikoff, S.; Henikoff, J. G. Performance evaluation of amino acid substitution matrices. *Proteins*, 17(1): 49–61, 1993.
- Hu, X.; Wang, H.; Ke, H.; Kuhlman, B. High-resolution design of a protein loop. *Proceedings National Academy Sciences*, 104: 17668–17673, 2007.
- Jayaram, B.; Sprous, D.; Beveridge, D. L. Solvation free energy of biomacromolecules: Parameters for a modified Generalized Born model consistent with the AMBER force field. *Journal of Physical Chemistry B*, 102: 9571–9756, 1998.
- Johansson, M. U.; de Chateau, M.; Wikstrom, M.; Forsen, S.; Drakenberg, T.; Bjorck, L. Solution structure of the albumin-binding GA module: a versatile bacterial protein domain. *Journal of Molecular Biology*. 266: 859–865, 1997.
- Jones, D. T.; Taylor, W. R.; Thornton, J. M. A new approach to protein fold recognition. *Nature*, 358(6381): 86–89, 1992.
- Jones, D. T. Predicting novel protein folds by using FRAGFOLD. *Proteins*, 5(Suppl): 127–132, 2001.
- Karplus, M. The Levinthal paradox: yesterday and today. *Folding and Design*, 2(1): S69–S75, 1997.
- Kim, D. E.; Chivian, D.; Baker, D. Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Research*, 32(suppl 2): W526–W531, 2004.
- King, R. D.; Sternberg, M. J. Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. *Protein Science*, 5(11): 2298–2310, 1996.
- Kolodny, R.; Guibas, L.; Levitt, M.; Koehl, P. Inverse Kinematics in Biology: The Protein Loop Closure Problem. *International Journal of Robotics Research*, 24: 151–163, 2005.
- Kundrotasa, P. J.; Lensinkb, M. F.; Alexova, E. Homology-based modeling of 3D structures of protein–protein complexes using alignments of modified sequence profiles. *International Journal of Biological Macromolecules*, 43: 198–208, 2008.
- Lai, P.; Kaplan, W.; Church, W. B.; Wong, R. K. Informative 3D visualization of multiple protein structures. In *Proceedings of the Second Conference on Asia-Pacific Bioinformatics - Volume 29 (Dunedin, New Zealand)*. Y. P. Chen, Ed. ACM International Conference Proceeding Series. Australian Computer Society, Darlinghurst, Australia, 55: 201–208, 2004.

- Laskowski, R. A.; MacArthur, M. W.; Moss, D. S.; Thornton, J. M. PROCHECK: a program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography*, 26(2): 283–291, 1993.
- Lehninger, A. L.; Nelson, D. L.; Cox, M. M. *Princípios de bioquímica*. Sarvier, 3ª Edição, São Paulo, 2002.
- Lesk, A. M. *Introduction to Protein Architecture*. New York: Oxford University Press, 347p, 2001.
- Lesk, A. M. *Introdução à Bioinformática*. Tradução de Ardala Elisa Breda Andrade et al. (LABIO/FACIN/PUCRS), 2ª Edição, Porto Alegre: Artmed, 384p, 2008.
- Liu, P.; Zhu, F.; Rassokhin, D. N.; Agrafiotis, D. K. A Self-Organizing Algorithm for Modeling Protein Loops. *PLoS Computational Biology*, 5(8): e1000478, 2009.
- Luscombe, N. M.; Greenbaum, D., Gerstein, M. What is Bioinformatics? A Proposed Definition and Overview of the Field. *Methods of Information in Medicine*, 40(4): 346–358, 2001.
- MacCallum, J. L.; Pérez, A.; Schmierders, M. J.; Hua, L.; Jacobson, M. P.; Dill, K. A. Assessment of protein structure refinement in CASP9. *Refinement Assessment, Proteins*, 79(S10): 74–90, 2011.
- Machine, Flashcard. Create, study and share online flash cards. Capturado em: <http://www.flashcardmachine.com/bmb461-structures.html>, Novembro 2011.
- MacKerell, Jr. A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D. e et al. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *Journal of Physical Chemistry*, 102(18): 3586–3616, 1998.
- MacQueen, J. 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*, University of California Press, 281–297, 1967.
- Mandal, S.; Moudgil, M.; Mandal, S. K. Rational drug design. *European Journal of Pharmacology, New Vistas in Anti-Cancer Therapy*, 625: 90-100, 2009.
- Mariani, V.; Kiefer, F.; Schmidt, T.; Haas, J.; Schwede, T. Assessment of template based protein structure predictions in CASP9. *Templated Based Assessment, Proteins*, 79(S10): 37–58, 2011.

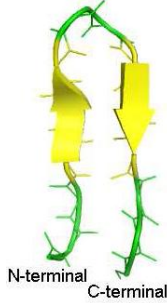
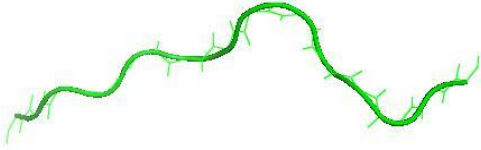
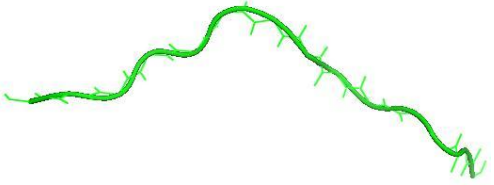
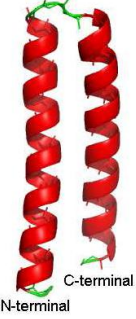

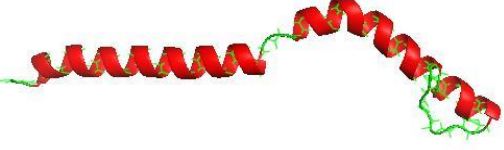
- Martí-Renom, M. A.; Stuart, A.; Fiser, A.; Sanchez, R.; Mello, F.; Sali, A. Comparative protein structure modelling of genes and genomes. *Annual Review of Biophysics and Biomolecular Structure*, 29(16): 291–235, 2000.
- Martin, A. C.; MacArthur, M. W.; Thornton, J. M. Assessment of comparative modeling in CASP2. *Proteins – Structure, Function and Bioinformatics*, 1(29): 14–28, 1997.
- Morize, I.; Surcouf, E.; Vaney, M. C.; Epelboin, Y.; Buehner, M.; Fridlansky, F.; Milgrom, E.; Mornon, J. P. Refinement of the C222(1) crystal form of oxidized uteroglobin at 1.34 Å resolution. *Journal of Molecular Biology*, 194(4): 725–739, 1987.
- Morris, A. L.; MacArthur, M. W.; Hutchinson, E. G.; Thornton, J. M. Stereochemical quality of protein structure coordinates. *Proteins*, 12(4): 345–364, 1992.
- Murzin, A. G.; Brenner, S. E.; Hubbard, T.; Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247(1): 536–540, 1995.
- Norberto de Souza, O.; Ornstein, R. L. Molecular dynamics simulations of a protein-protein dimer: particle-mesh Ewald electrostatic model yields far superior results to standard cutoff model. *Journal of Biomolecular Structure & Dynamics*, 16: 1205–1217, 1999.
- Orengo, C. A.; Michie, A. D.; Jones, S.; Jones, D. T.; Swindells, M. B.; Thornton, J. M. CATH - A hierarchic classification of protein domain structures. *Structure*, 5(8): 1093–1108, 1997.
- Pastor, M. T.; Lopez de la Paz, M.; Lacroix, E.; Serrano, L.; Perez-Paya, E. 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.
- Pauling, L.; Corey, R. B. 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.
- Pollastri, G.; McLysaght, A. Porter: a new, accurate server for protein secondary structure prediction. *Bioinformatics*, 21(8): 1719-1720, 2005.
- Pressman, R. S. *Engenharia de Software*. 3ª Edição, São Paulo: Makron Books, 1995.
- Prosdocimi, F.; Cerqueira, G. C.; Binneck, E.; Silva, A. F.; Reis, A. N.; Junqueira, A. C. M.; Santos, A. C. F.; Nbani, A.; Wust, C. I.; Filho, F. C.; Kessedjian, J. L.; Petretski, J. H.; Camargo, L. P.; Ferreira, R. G. M.; Lima, R. P.; Pereira, R. M.; Jardim, S.; Sampaio, V. S.;


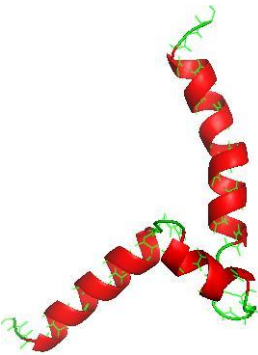

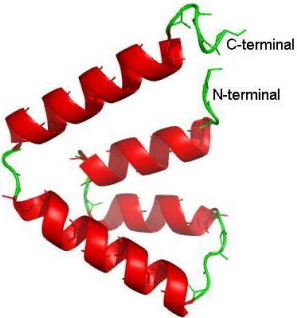
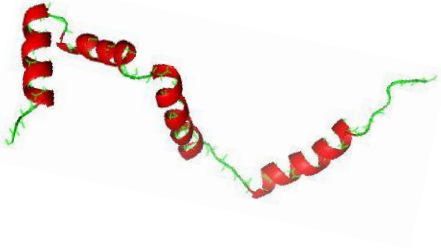
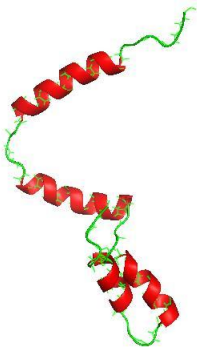
- Flatschart, A. V. F. *Bioinformática: manual do usuário*. Biotecnologia Ciência e Desenvolvimento, (29): 12-25, 2002.
- Rajgaria, R.; Wei, Y.; Floudas, C. A. Contact prediction for beta and alpha-beta proteins using integer linear optimization and its impact on the first principles 3d structure prediction method ASTRO-FOLD. *Proteins – Structure, Function and Bioinformatics*, 78: 1825-1846, 2010.
- Ramachandran, G. N.; Sasisekharan, V. Conformation of polypeptides and proteins. *Advances in Protein Chemistry*, 23: 238–437, 1968.
- Roe, D. R.; Okur, A.; Wickstrom, L.; Hornak, V.; Simmerling, C. Secondary structure bias in Generalized Born Solvent Models: Comparison of conformational ensembles and free energy of solvent polarization from explicit and implicit salvation. *Journal of Physical Chemistry B*, 111: 1846–1857, 2007.
- Rohl, C. A.; Strauss, C. E. M.; Misura, K. M. S.; Baker, D. Protein structure prediction using Rosetta. *Methods in Enzymology*, 383: 66–93, 2004.
- Rost, B.; Sander, C. Prediction of protein secondary structure at better than 70 percent accuracy. *Journal of Molecular Biology*, 232(2): 584–599, 1993.
- Simons, K. T.; Bonneau, R.; Ruczinski, I.; Baker, D. *Ab initio* protein structure prediction of CASP III targets using Rosetta. *Proteins, Suppl 3(6)*: 171–176, 1999.
- Srinivasan, R.; Rose, G. D. LINUS - A hierarchic procedure to predict the fold of a protein. *Proteins*, 22: 81–99, 1995.
- Srinivasan, R.; Rose, G. D. *Ab initio* prediction of protein structure using LINUS. *Proteins*, 47: 489–495, 2002.
- Srinivasan, R.; Fleming, P. J.; Rose, G. D. *Ab initio* protein folding using LINUS. *Methods in Enzymology*, 383: 48–66, 2004.
- Starovasnick, M. A.; Brasisted, A. C.; Wells, J. A. 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.
- Velankar, S.; McNeil, P.; Mittard-Runte, V.; Suarez, A.; Barrell, D.; Apweiler, R.; Henrick, K. E-MSD: an integrated data resource for bioinformatics. *Nucleic Acids Research*, 33(Database Issue), 2005.

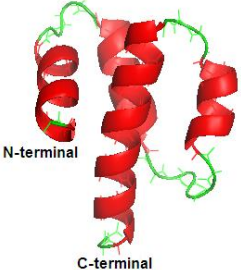
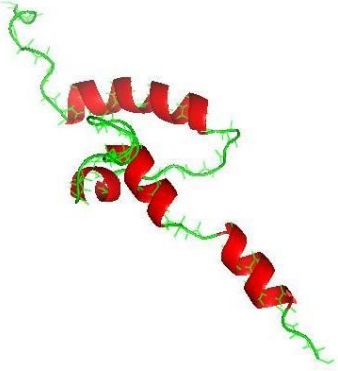

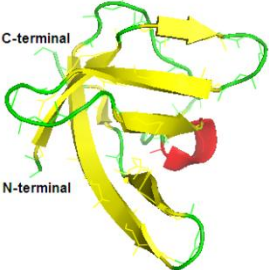
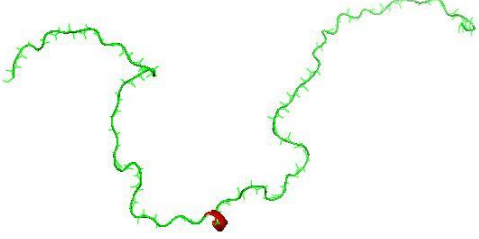

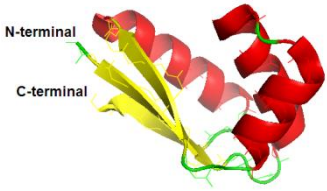
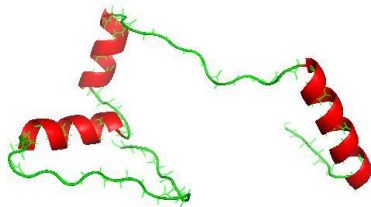

- Venkatachalam, C. M. Stereochemical criteria for polypeptides and proteins. V. Conformation of a system of three linked peptide units. *Biopolymers*, 6: 1425-1436, 1968.
- van Gunsteren, W. F.; Berendsen, H. J. C. Computer simulation of molecular dynamics: methodology, applications, and perspectives in chemistry. *Angewandte Chemie International Edition in English*, 29: 992–1023, 1990.
- van Vlijmen, H. W. T.; Karplus, M. PDB-based protein loop prediction: Parameters for selection and methods for optimization. *Journal of Molecular Biology*, 267: 975–1001, 1997.
- Voet, D.; Voet, J. G. *Bioquímica*. 3ª Edição, Porto Alegre: Artmed, 1616p, 2006.
- Wang, C.; Bradley, P.; Baker, D. Protein-protein docking with backbone flexibility. *Journal of Molecular Biology* 373: 503–519, 2007.
- Witten, I. H.; Frank, E. *Data mining: practical machine learning tools and techniques*. Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, Oxford, UK, 2ª edição, 2005.
- Wu, S.; Skolnick, J.; Zhang, Y. Ab initio modeling of small proteins by iterative tasser simulations. *BMC Biology*, 5: 17, 2007.
- Zhang, Y. I-TASSER server for protein 3D structure prediction. *BCM Bioinformatics*, 9: 40, 2008.

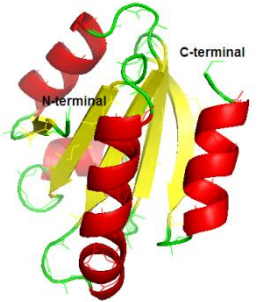

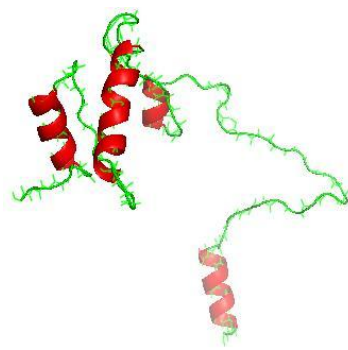
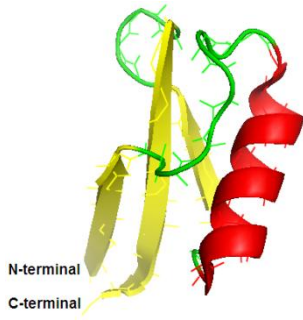
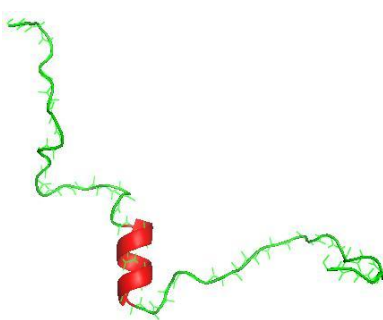
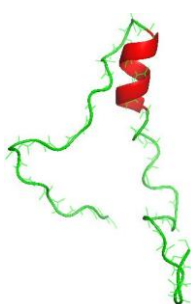
Apêndice A – Comparação da conformação inicial predita pelo CReF

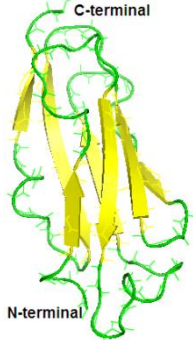
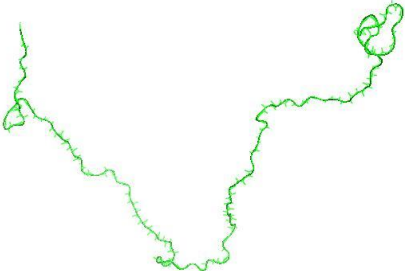
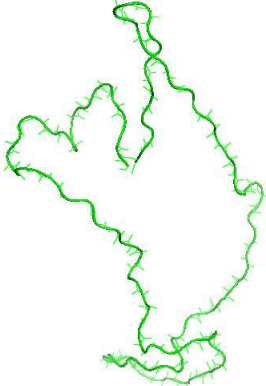
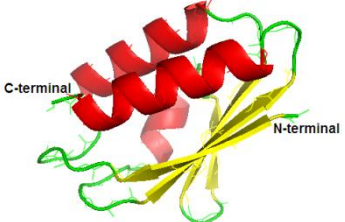
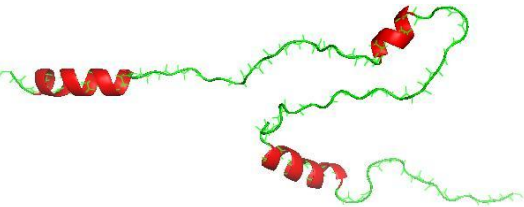
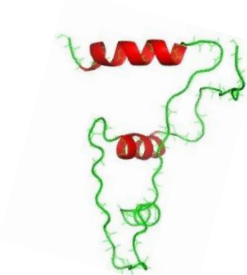
Tabela 19 – Comparação entre a estrutura experimental e as estruturas preditas considerando quatro e seis grupos na mineração de dados para o conjunto de proteínas teste da nova versão do CReF.

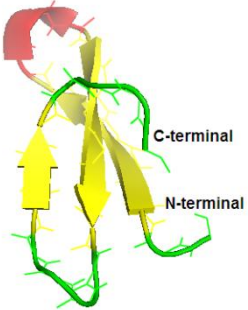


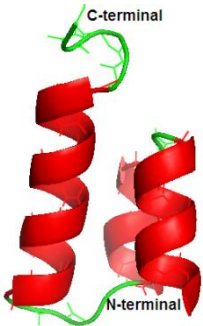

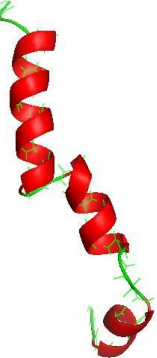
Proteína	Experimental	4 Grupos	6 Grupos
1K43			
		RMSD: 9,93 Å	RMSD: 9,91 Å
1ROP			
		RMSD: 6,48 Å	RMSD: 22,08 Å

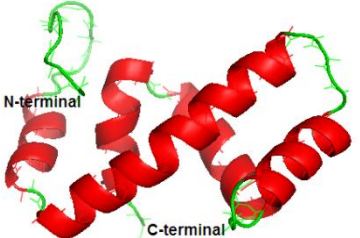
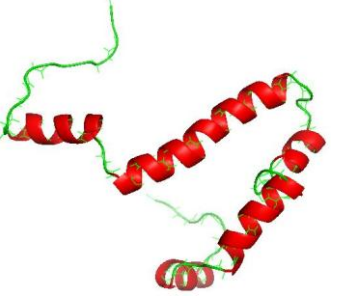

Proteína	Experimental	4 Grupos	6 Grupos
1GAB	 <p data-bbox="555 300 645 323">C-terminal</p> <p data-bbox="450 655 533 679">N-terminal</p>	 <p data-bbox="1048 770 1234 802">RMSD: 10,79 Å</p>	 <p data-bbox="1599 770 1785 802">RMSD: 10,35 Å</p>
1UTG	 <p data-bbox="629 895 712 919">C-terminal</p> <p data-bbox="622 951 705 975">N-terminal</p>	 <p data-bbox="1048 1321 1234 1353">RMSD: 19,55 Å</p>	 <p data-bbox="1599 1321 1785 1353">RMSD: 15,41 Å</p>

Proteína	Experimental	4 Grupos	6 Grupos
1C5A	 <p>N-terminal C-terminal</p>	 <p>RMSD: 12,29 Å</p>	 <p>RMSD: 9,56 Å</p>
1CSP	 <p>C-terminal N-terminal</p>	 <p>RMSD: 29,49 Å</p>	 <p>RMSD: 22,49 Å</p>
1CTF	 <p>N-terminal C-terminal</p>	 <p>RMSD: 17,71 Å</p>	 <p>RMSD: 11,45 Å</p>

Proteína	Experimental	4 Grupos	6 Grupos
1ERV	 The experimental structure of protein 1ERV is shown as a ribbon model. It features a complex fold with several alpha-helices and beta-strands. The N-terminal region is highlighted in yellow, and the C-terminal region is highlighted in red. Labels "N-terminal" and "C-terminal" are present.	 The prediction for protein 1ERV using 4 groups is shown as a ribbon model. It captures the general shape of the protein but lacks the detailed helical and strand structure seen in the experimental model.	 The prediction for protein 1ERV using 6 groups is shown as a ribbon model. It provides a more detailed representation of the protein's structure, including the helices and strands, compared to the 4-group prediction.
1GPT	 The experimental structure of protein 1GPT is shown as a ribbon model. It features a complex fold with several alpha-helices and beta-strands. The N-terminal region is highlighted in yellow, and the C-terminal region is highlighted in red. Labels "N-terminal" and "C-terminal" are present.	 The prediction for protein 1GPT using 4 groups is shown as a ribbon model. It captures the general shape of the protein but lacks the detailed helical and strand structure seen in the experimental model.	 The prediction for protein 1GPT using 6 groups is shown as a ribbon model. It provides a more detailed representation of the protein's structure, including the helices and strands, compared to the 4-group prediction.
		RMSD: 17,49 Å	RMSD: 22,10 Å
		RMSD: 20,76 Å	RMSD: 14,67 Å

Proteína	Experimental	4 Grupos	6 Grupos
1KSR	 The experimental structure of protein 1KSR is shown as a ribbon model. The N-terminal region is colored yellow, and the C-terminal region is colored green. Labels "C-terminal" and "N-terminal" are present.	 A green wireframe approximation of the protein 1KSR structure, representing the 4 Grupos model.	 A green wireframe approximation of the protein 1KSR structure, representing the 6 Grupos model.
1OPD	 The experimental structure of protein 1OPD is shown as a ribbon model. The N-terminal region is colored red, and the C-terminal region is colored yellow. Labels "C-terminal" and "N-terminal" are present.	 A green wireframe approximation of the protein 1OPD structure, representing the 4 Grupos model. Red helical segments are visible.	 A green wireframe approximation of the protein 1OPD structure, representing the 6 Grupos model. Red helical segments are visible.
		RMSD: 40,85 Å	RMSD: 23,72 Å
		RMSD: 25,67 Å	RMSD: 15,19 Å

Proteína	Experimental	4 Grupos	6 Grupos
1YWJ	 <p>The experimental structure of protein 1YWJ is shown as a ribbon model. The N-terminal region is colored yellow and the C-terminal region is colored green. Labels 'N-terminal' and 'C-terminal' are present.</p>	 <p>The prediction for protein 1YWJ using 4 groups is shown as a green ribbon model.</p> <p data-bbox="1055 751 1223 783">RMSD: 8,64 Å</p>	 <p>The prediction for protein 1YWJ using 6 groups is shown as a green ribbon model.</p> <p data-bbox="1603 751 1783 783">RMSD: 10,85 Å</p>
2ERL	 <p>The experimental structure of protein 2ERL is shown as a ribbon model. The N-terminal region is colored green and the C-terminal region is colored red. Labels 'N-terminal' and 'C-terminal' are present.</p>	 <p>The prediction for protein 2ERL using 4 groups is shown as a ribbon model with red and green segments.</p> <p data-bbox="1055 1342 1223 1374">RMSD: 9,08 Å</p>	 <p>The prediction for protein 2ERL using 6 groups is shown as a ribbon model with red and green segments.</p> <p data-bbox="1603 1342 1783 1374">RMSD: 12,40 Å</p>

Proteína	Experimental	4 Grupos	6 Grupos
2EZK	 The experimental protein structure 2EZK is shown as a red ribbon model. It features a complex fold with several alpha-helices and beta-strands. The N-terminal and C-terminal regions are highlighted in green and labeled "N-terminal" and "C-terminal" respectively.	 The protein structure with 4 groups is shown as a red ribbon model. It is a simplified representation of the experimental structure, with the N-terminal and C-terminal regions highlighted in green.	 The protein structure with 6 groups is shown as a red ribbon model. It is a further simplified representation of the experimental structure, with the N-terminal and C-terminal regions highlighted in green.
		RMSD: 12,82 Å	RMSD: 15,39 Å

Proteína	Método	Erro	Sequência																																																	
			M	T	Y	K	L	I	L	N	G	K	T	L	K	G	E	T	T	T	E	A	V	D	A	A	T	A	E	K	V	F	K	Q	Y	A	N	D	N	G	V	D	G	E	W	T	Y					
1GB1	Experimental		c	e	e	e	c	c	e	e	c	c	c	c	e	e	c	c	e	e	e	c	c	c	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	c	c	c	c	c	c	e	e	e	e
	DPM	28	c	c	e	e	e	e	e	t	c	c	c	t	c	c	c	c	c	c	h	h	h	h	h	h	h	h	h	h	e	h	h	h	h	t	t	t	c	t	t	t	c	c	c	c	c	c				
	DSC	16	c	c	e	e	e	e	e	c	c	c	c	c	c	c	c	e	e	e	e	h	h	h	h	h	h	h	h	h	h	h	h	e	e	c	c	c	c	c	c	c	c	e	e	e	e	e	e			
	PHD	11	c	e	e	e	e	e	e	c	c	c	c	c	c	e	e	e	e	e	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	c	c	c	c	c	c	c	e	e	e	e	
	Predator	5	c	e	e	e	e	e	e	e	c	c	c	c	e	e	e	e	e	e	e	c	c	c	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	c	c	c	c	c	e	e	e	e
	SCRATCH	23	c	c	e	e	e	e	e	e	e	e	e	e	e	c	c	c	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	t	t	t	c	c	e	e	e	e	e		
	SOPM	21	e	e	e	e	e	e	e	t	t	c	c	c	c	c	c	c	c	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	c	t	t	t	t	c	c	c	e	e	e	e	e		
	NetSurfP	20	c	e	e	e	e	e	e	c	c	c	e	e	c	c	c	c	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	c	c	c	c	c	c	c	e	c	c			
	GorV	16	c	c	c	e	e	e	e	c	c	c	c	c	c	c	c	c	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	c	c	c	c	c	c	e	e	e	e	
	CSSP	13	c	e	e	e	e	e	e	c	c	c	c	c	c	c	c	c	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	c	c	c	c	c	c	e	e	e	e	
	Sable	13	c	e	e	e	e	e	e	c	c	c	c	c	c	c	c	c	h	h	c	c	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	c	c	c	c	c	c	c	e	e	c			
	SAM-T08	16	e	e	e	e	e	e	e	c	c	c	e	e	e	c	e	e	e	e	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	c	c	c	c	c	c	c	e	e	e	e		
	SSpro4	5	c	e	e	e	e	e	e	e	c	c	c	c	e	e	e	e	e	e	e	c	c	c	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	c	c	c	c	c	e	e	e	e	
	Porter2	5	c	e	e	e	e	e	e	e	c	c	c	c	e	e	e	e	e	e	e	c	c	c	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h	c	c	c	c	c	e	e	e	e

	D	D	A	T	K	T	F	T	V	T	E
Experimental	e	c	c	c	c	e	e	e	e	e	c
DPM	c	c	h	c	h	e	e	e	e	c	c
DSC	c	c	c	c	c	e	e	e	e	e	c
PHD	e	c	c	c	c	e	e	e	e	e	c
Predator	e	c	c	c	c	e	e	e	e	e	c
SCRATCH	c	h	t	c	e	e	e	e	e	c	c
SOPM	e	t	t	t	c	e	e	e	e	e	c
NetSurfP	c	c	c	c	e	e	e	e	e	c	c
GorV	c	c	c	c	c	e	e	e	e	c	c
CSSP	c	c	c	c	c	e	e	e	e	e	c
Sable	c	c	c	c	e	e	e	e	e	e	c
SAM-T08	c	c	c	c	c	e	e	e	e	c	c
SSpro4	e	c	c	c	c	e	e	e	e	e	c
Porter2	e	c	c	c	c	e	e	e	e	e	c

