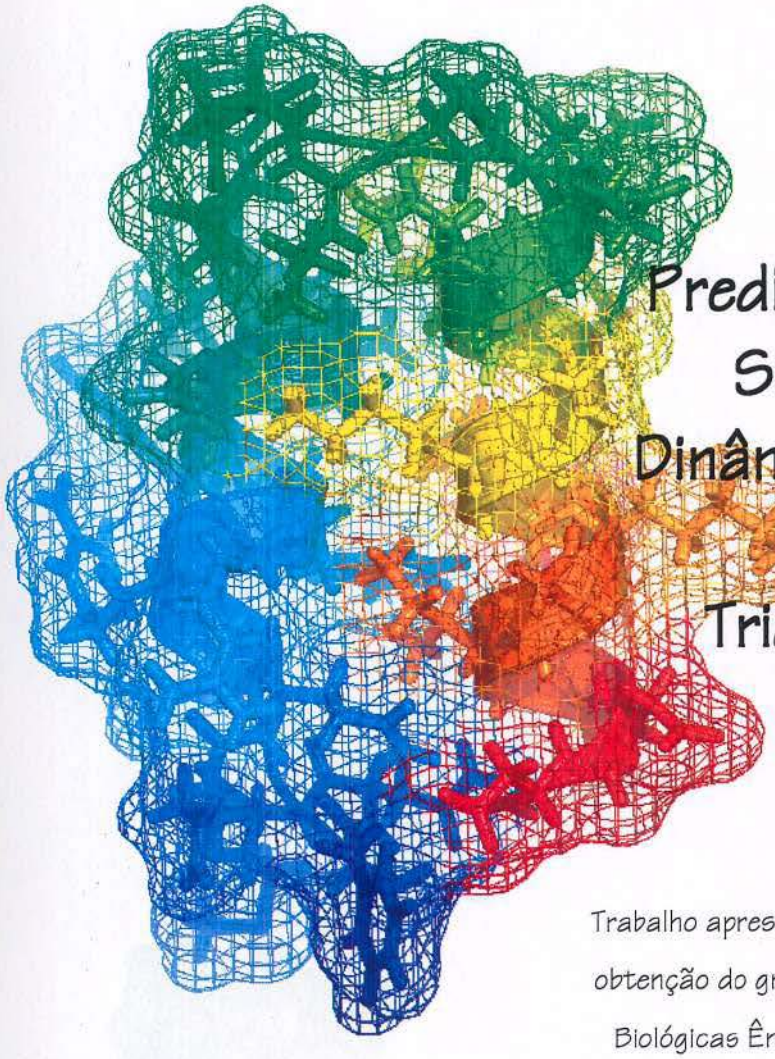
Universidade Federal do Rio Grande do Sul

Instituto de Biociências

Centro de Biotecnologia do Estado do Rio Grande do Sul



Predição *Ab Initio* por Simulação pela Dinâmica Molecular da Estrutura Tridimensional de Proteínas

Trabalho apresentado como um dos requisitos para obtenção do grau de Bacharel no Curso de Ciências Biológicas Ênfase Molecular, Celular e Funcional.

Autora: Ardala Breda

Orientador: Prof. Dr. Diógenes Santiago Santos

Co-orientador: Prof. Dr. Osmar Norberto de Souza

Porto Alegre, janeiro de 2004.

A ilustração da capa é o polipeptídeo PA_Z, um dos modelos de estudo apresentado neste trabalho. A figura representa a estrutura atômica de PA_Z e sua superfície molecular. A estrutura está colorida da extremidade N-terminal (azul) para C-terminal (vermelho).

(...) foi no formigueiro que começaram as dignas formigas e no formigueiro terminarão, o que lhes honra a perseverança e o senso prático. Mas o homem é um ser versátil, e é possível que, como jogador de xadrez ele ame o processo de atingir o objetivo, e não o objetivo em si mesmo (...)

Fiodor Dostoievski

Índice

1. Introdução

1.1. Bioinformática Estrutural

Durante a década de 90 ocorreu a chamada 'explosão de dados biológicos' derivados principalmente dos Projetos Genoma, executados em diferentes países tanto na iniciativa pública quanto privada (Figura 1). Como resultado destes projetos, um grande número de novos genes foi identificado e são tidos como potenciais alvos de estudo, porém, seus produtos protéicos ainda não estão caracterizados [1]. Um dos grandes desafios da era pós-genômica é a determinação e validação de novos genes e suas proteínas correspondentes; ou seja, a conversão dos dados obtidos a partir dos seqüenciamentos em informação útil. Um dos componentes cruciais deste desafio reside no estudo destas proteínas não anotadas e na análise de suas estruturas, sejam elas determinadas experimentalmente ou modeladas.

Os métodos clássicos de determinação de estruturas de proteínas, cristalografia por difração de raios-X e ressonância magnética nuclear em solução (NMR) não são suficientes para a caracterização de toda e qualquer proteína, pois em muitos casos não é e não será possível a superexpressão e purificação das mesmas para a preparação de cristais para cristalografia, ou das soluções para uso em NMR. A Bioinformática Estrutural tem impacto significante nas caracterizações bem sucedidas destas proteínas, pois desenvolve estratégias para a determinação de estruturas tridimensionais (3D) e pode prover hipóteses também acerca das suas funções.

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 a matemática, ciência da computação e estatística) para entender, organizar e explorar a informação estrutural associada a essas moléculas em uma grande escala [2].

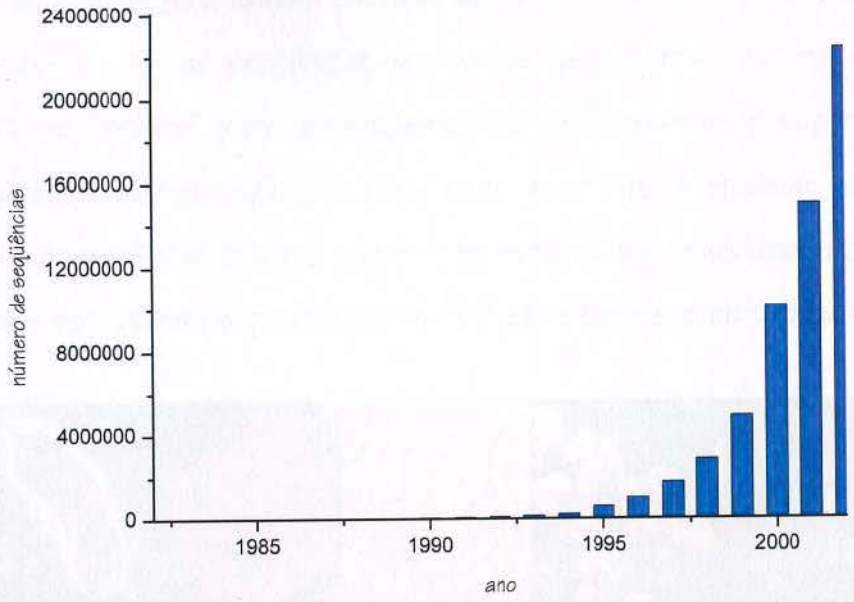


Figura 1. *O gráfico mostra o número de seqüências depositados do GenBank, e seu aumento acentuado a partir da segunda metade da década de 90, período conhecido como 'explosão de dados biológicos'. Fonte: www.ncbi.nlm.nih.gov/Genbank, última atualização em fevereiro de 2003.*

Das seqüências já completas derivadas dos genomas de diferentes organismos (Figura 2), poucas são as proteínas que podem ser identificadas pelos métodos clássicos. Estima-se que cerca de 50% das proteínas de um organismo podem ter sua função inferida com uma confiança razoável a partir de sua comparação com homólogas [3, 4]. Para as demais, informações sobre suas funções podem ser obtidas de seqüências padrões ou motivos que são característicos de uma superfamília. Possivelmente, de 30% a 40% das proteínas não poderão ser identificadas, pois são membros de famílias com funções ainda não conhecidas [5]; e, por não possuírem similares nos genomas seqüenciados e anotados até o momento, dificilmente serão reconhecidas somente com base sua na seqüência.

Além do problema de determinação de estrutura deste terceiro grupo de proteínas, cabe ressaltar que os motivos estruturais 3D disponíveis atualmente em bancos de dados como PDB (*Protein Data Bank* — Research Collaboratory for

Structural Bioinformatics), não são suficientes para a anotação completa de genomas. Vitkup *et al.* [6], usando estruturas não redundantes depositadas no PDB e comparando-as com as seqüências completas de genomas, estima que existam apenas 30% de "moldes" para as proteínas destes genomas; e sugere ainda que aproximadamente outros 16.000 seriam necessários para modelar um proteoma completo. A discrepância entre o número de estruturas resolvidas e o número de novos motivos estruturais depositados no PDB pode ser visto na tabela 1 e Figura 3.

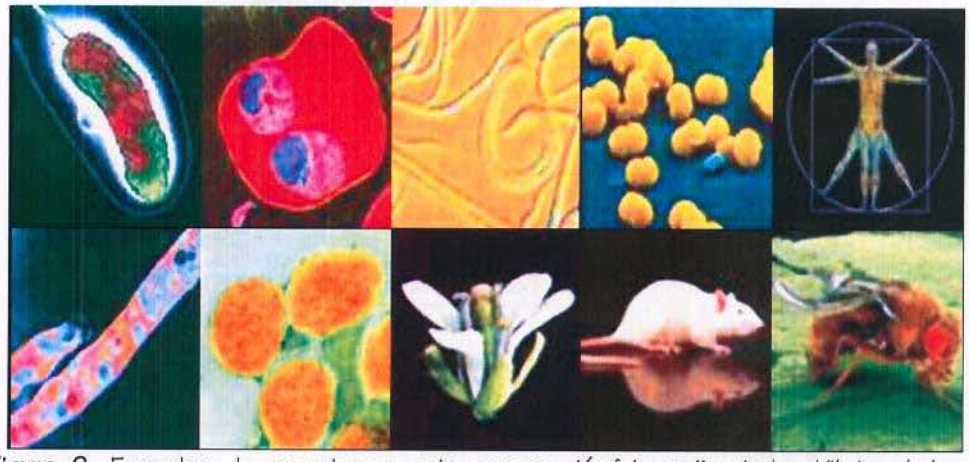


Figura 2. Exemplos de organismos cujo genoma já foi seqüenciado: *Vibrio cholerae, Plasmodium falciparum, Caenorhabditis elegans, Neisseria meningitidis, Homo sapiens, Mycobacterium tuberculosis, Mycobacterium leprae, Arabdopsis thaliana, Mus musculus* e *Drosophila melanogaster*.

Esta problemática requer uma nova tecnologia de estudo e caracterização de estruturas protéicas, que seja capaz de inferir corretamente (com pequena margem de erro) a conformação 3D nativa de uma proteína tendo como base apenas sua seqüência de aminoácidos, pois esta é a única informação disponível; e que possa ainda permitir a descoberta de novas formas de enovelamento ou dobramento (motivos estruturais). A técnica que permite esta abordagem é a predição por 'primeiros princípios' ou *ab initio*.

Tabela 1. Número de motivos estruturais não redundantes (*folds*) já classificados. Adaptado de SCOP em http://scop.mrc-lmc.cam.ac.uk/scop, última atualização 1 de agosto de 2003.

Classe	Nº de folds
Proteínas alpha	179
Proteínas beta	126
Proteínas alpha/beta	121
Proteínas alpha + beta	234
Proteínas multi-domínio	38
Proteínas de membrana e de superfície celular	36
Miniproteínas	66
Total	800

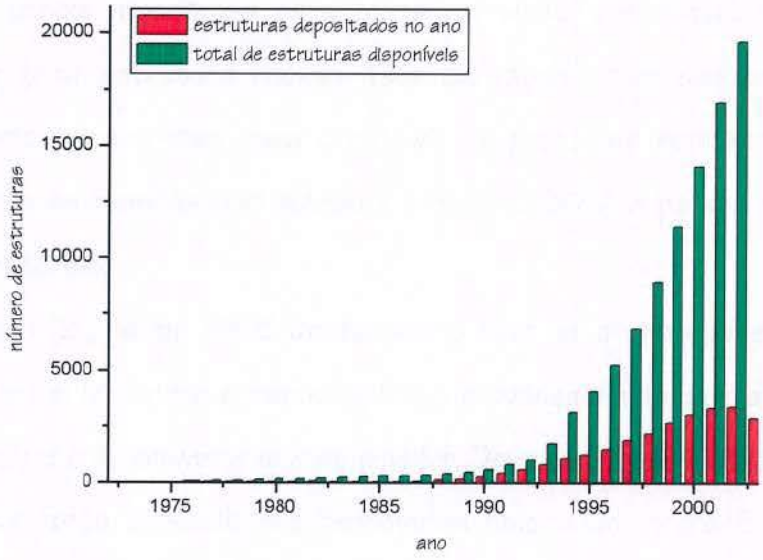


Figura 3. Número de estruturas protéicas depositadas no banco de dados Protein Data Bank - PDB. A grande discrepância entre o número de estruturas depositadas e o número de motivos (Tabela 1) ocorre uma vez que os motivos se repetem nas estruturas resolvidas. Adaptado de www.rcsb.org/pdb, última atualização 6 de agosto de 2003.

O método de predição *ab initio* usa a seqüência linear de aminoácidos como ponto de partida para a construção do modelo 3D, baseado em métodos de dinâmica molecular (DM), mecânica molecular e mecânica quântica para descrever as interações entre os átomos que constituem a molécula, de acordo com leis químicas

e físicas [7]. A predição *ab initio*, mesmo que apenas parcialmente correta, provê meios de entendimento dos princípios de formação da estrutura secundária e terciária de proteínas [8], e tem papel central no entendimento das relações entre seqüência e estrutura.

Outros métodos de predição de estruturas, como modelagem comparativa por homologia e métodos de reconhecimento de motivos (*folds*) via *threading*, também são utilizados para a construção de modelos 3D hipotéticos; porém ambos o fazem tendo como molde outras proteínas cuja estrutura já foi resolvida. No caso da modelagem comparativa, são criados modelos com base na homologia (acima de 30%) entre as seqüências do molde e da proteína em estudo. O método de *threading* é aplicado em seqüências relacionadas evolutivamente (também com homologia acima de 30%), e utiliza algoritmos para mensurar a compatibilidade entre uma dada seqüência e uma estrutura molde. Tais características descartam estes dois métodos como ferramentas para o estudo de proteínas sem homólogos (ou cuja homologia com as famílias conhecidas é inferior a 30%) e para a procura de novos motivos estruturais.

Anfinsen [9], já em 1960 destacava o fato de o enovelamento de proteínas globulares ser um fenômeno puramente físico e conseqüência direta dos aminoácidos que as compõem e do solvente que as envolve. Deve ser possível então, a definição de um campo de força baseado nos fenômenos físicos de interações entre átomos, incluindo o solvente, e o uso de processos estatísticos e/ou mecânicos para a determinação da estrutura mais estável de uma proteína, a uma dada temperatura e condições de solvatação. De acordo com estes princípios, o método de predição *ab initio* requer três elementos principais: a representação geométrica de sistemas protéicos, campo de força adequado e a mensuração da variação da energia do sistema [10].

Um dos requisitos para o estudo do comportamento de um sistema protéico é uma técnica que permita a varredura do espaço conformacional que pode ser ocupado

ao longo de uma simulação. Diferentes técnicas podem ser empregadas, como algoritmos genéticos e simulações de Monte Carlo e Dinâmica Molecular [11], neste trabalho optamos pela DM.

Geometricamente, proteínas podem ser modelos onde todos seus átomos constituintes são representados explicitamente e individualizados (*all-atom model*); com alguns átomos unidos, onde apenas os átomos pesados e os átomos de hidrogênio polares de cada resíduo são representados (*united-atom model*); e finalmente considerando um único átomo "virtual" como todo o resíduo de aminoácido. Freqüentemente alguma forma de simplificação geométrica é utilizada para aceleração dos cálculos de DM; neste trabalho todos os átomos que constituem cada resíduo de aminoácido foram representados explicitamente.

O estado nativo de uma proteína é definido como aquele em que a energia global do sistema é mínima, de acordo com princípios químicos de que quanto mais estável um sistema menor a sua energia livre. A estabilidade termodinâmica e valores mínimos de energia são procurados ao longo de uma simulação como prováveis candidatos a conterem a proteína em sua conformação nativa.

Nos últimos anos, tem havido estudos extensivos sobre como, e quais são os princípios que regem o enovelamento de proteínas, em busca de soluções para o paradoxo de Levinthal* [12]; no entanto, poucos são os grupos de pesquisa que se dedicam ao método de predição de estruturas *ab initio* por simulações pela DM, e mesmo dentre eles, a atenção volta-se principalmente aos processos de enovelamento e não à obtenção de estruturas 3D corretas, por exemplo, os trabalhos de Carlos Simmerling [13], Michael Levitt [14], Yong Duan [15], David Bashford [16] e Eugene Shakhnovich [17].

* **Cyrus Levinthal**, professor da universidade de Illinois, em um encontro onde se discutiam sistemas biológicos, levantou a questão de que uma proteína durante seu processo de enovelamento não percorre todas as conformações que seriam possíveis a ela adotar, pois o tempo que seria consumido neste método de tentativa e erro seria infinitamente maior que o tempo de enovelamento observado experimentalmente. Cada aminoácido poderia teoricamente assumir 10 diferentes conformações, de acordo com seus ângulos de rotação, portanto, uma proteína deveria "experimentar" 10^n conformações, sendo n seu número de aminoácidos. Assumindo que esta proteína possa adotar diferentes conformações, na ordem de 10^{14} estruturas por segundo, ela levaria aproximadamente 10^{18} anos para testar todas suas possibilidades, e este valor ultrapassa em muitas vezes a idade do universo; como o enovelamento de proteínas ocorre na escala de milisegundos a segundos, este é o paradoxo de Levinthal.

1.2. Estrutura Protéica

As proteínas exercem papéis crucias em virtualmente todos os processos biológicos, por exemplo: na catálise enzimática; carreando transporte e armazenamento intracelular; possibilitando o movimento muscular coordenado e a sustentação mecânica; promovendo a proteção imunológica, a geração e transmissão de impulsos nervosos, além do controle do crescimento e da diferenciação celular. Proteínas essenciais para microorganismos patogênicos como, por exemplo, *Mycobacterium tuberculosis*, podem ser definidas e utilizadas como alvos terapêuticos.

Os aminoácidos são as unidades estruturais de qualquer proteína; um **α**-aminoácido é constituído de um grupamento amina, uma carboxila, um átomo de hidrogênio e um radical R diferenciado ou cadeia lateral, todos ligados a um carbono **α** (este carbono é chamado **α** por ser adjacente à carboxila ácida — Figura 4).

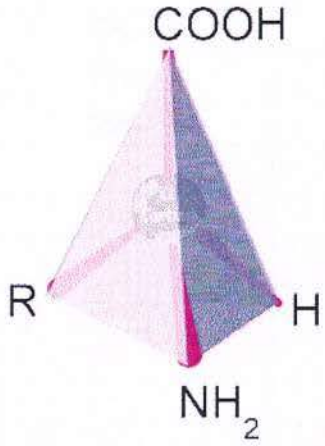


Figura 4. Estrutura tetraédrica de um **α** aminoácido e seus diferentes ligantes: grupamento amina (NH_2), hidrogênio (H), carboxila (COOH) e radical R ou cadeia lateral. A constituição dos vinte aminoácidos naturais difere apenas quanto a sua cadeia lateral.

Os aminoácidos naturais são divididos em três classes distintas, definidas pela natureza química da cadeia lateral (tabela 2). A primeira classe compreende os aminoácidos cujas cadeias laterais são hidrofóbicas, ou apolares: alanina, valina, leucina, isoleucina, fenilalanina, prolina e metionina. A segunda classe compreende os

resíduos polares: serina, treonina, cisteína, asparagina, glutamina, histidina, tirosina, triptofano e glicina. Os quatro resíduos carregados formam as demais classes: arginina e lisina, de carga positiva; e aspartato e glutamato, de carga negativa. Estes vinte aminoácidos diferem quanto a sua cadeia lateral, que varia em tamanho, forma, carga, capacidade de formação de pontes de hidrogênio e reatividade química (Figura 5).

Tabela 2. Lista dos vinte aminoácidos de ocorrência natural, seus nomes e código referente a cada um, de três letras e uma letra, coloridos em função de seu caráter. Em cinza, aminoácidos apolares, amarelo, aminoácidos polares, em **azul** os positivos e em **vermelho** os negativos.

Aminoácido	Código de três letras	Código de uma letra
Alanina	Ala	A
Valina	Val	V
Leucina	Leu	L
Isoleucina	Ile	I
Fenilalanina	Phe	F
Prolina	Pro	P
Metionina	Met	M
Serina	Ser	S
Treonina	Thr	T
Cisteína	Cis	C
Asparagina	Asn	N
Glutamina	Gln	Q
Histidina	His	H
Tirosina	Tir	Y
Triptofano	Trp	W
Glicina	Gli	G
Arginina	**Arg**	**R**
Lisina	**Lis**	**K**
Aspartato	**Asp**	**D**
Glutamato	**Glu**	**E**

Resíduos cujas cadeias laterais são apolares tendem a estar mais internalizados nos motivos estruturais protéicos, isto porque sofrem efeito hidrofóbico e não ficam expostos a solventes polares; suas cadeias laterais são constituídas basicamente por carbono e hidrogênio. Já os resíduos carregados apresentam seu grupamento amina ou carboxila na forma ionizada e ficam expostos ao solvente, pois são hidrofílicos, assim como os resíduos polares, que apesar de não possuírem carga efetiva, possuem polaridade pela presença de grupamentos OH, SH ou anéis aromáticos e anéis nitrogenados em sua estrutura.

As cadeias polipeptídicas possuem um sentido, uma vez que seus componentes têm extremidades diferentes; por convenção, a ponta amídica é considerada o início da cadeia, portanto, uma seqüência de aminoácidos é escrita a partir da sua porção amino terminal (N-terminal).

Uma cadeia peptídica distendida ou disposta ao acaso é isenta de atividade biológica, pois sua função surge da conformação, isto é, do arranjo tridimensional dos átomos em uma estrutura. No final da década de 30, Linus Pauling e Robert Corey iniciaram estudos cristalográficos por difração de raios-X da estrutura de aminoácidos e peptídeos cuja meta era a obtenção das distâncias e ângulos padrões de ligação entre os átomos de um aminoácido, e a utilização desta informação na predição de conformações protéicas [18, 19].

O mais importante de seus achados foi que a unidade de ligação peptídica é rígida e plana. A ligação entre o carbono da carboxila e o nitrogênio da amina não é livre para rotar porque essa ligação tem um caráter parcial de ligação dupla. O comprimento desta ligação é de 1,32 Å, valor entre o comprimento de uma ligação simples (1,49Å) e o de uma ligação dupla (1,27Å) [20].

A ligação entre o carbono α e o carbono carboxílico e entre o carbono α e o nitrogênio peptídico são ambas simples e portanto, possuem considerável grau de liberdade de rotação em torno da ligação peptídica rígida (ângulos diedros phi —φ e psi —ψ da cadeia principal — Figura 6).

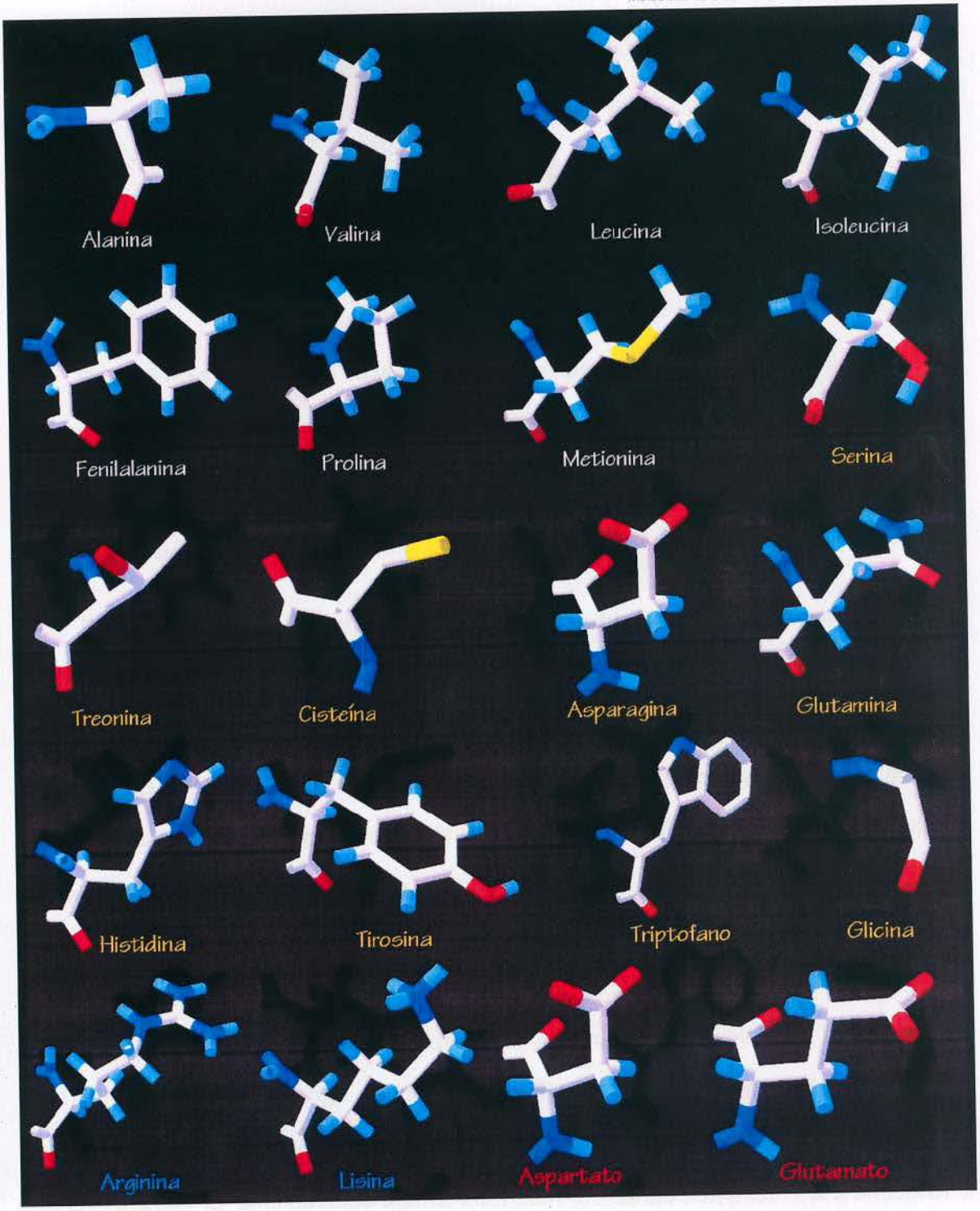


Figura 5. Os mesmos aminoácidos listados na tabela 2 são aqui mostrados com destaque para a estrutura atômica de suas cadeias laterais. Os nomes da cada aminoácido estão coloridos conforme a tabela 2, em cinza os resíduos apolares, em amarelo resíduos polares, em **azul** os positivos e em **vermelho** os negativos. Átomos de carbono em branco, átomos de nitrogênio em **azul**, oxigênio em **vermelho**, enxofre em amarelo e hidrogênio em **azul claro**.

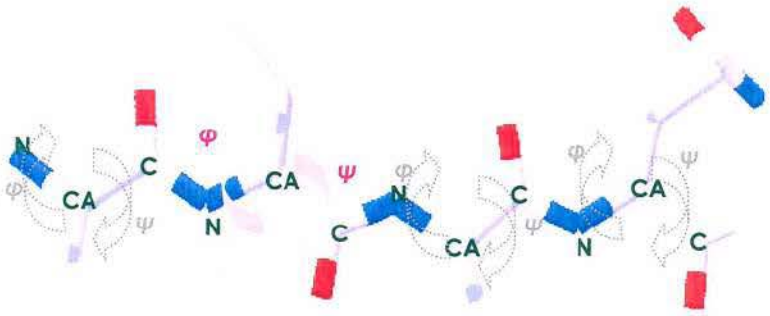


Figura 6. Representação da cadeia principal de um polipeptídeo, com destaque para os ângulos diedros φ e ψ e suas possíveis rotações em torno da ligação peptídica rígida (entre C e os nitrogênios, representados em azul). Átomos de carbono em branco, átomos de nitrogênio em azul e oxigênio em vermelho. Os átomos hidrogênio não foram representados para maior clareza.

Desta maneira, cada aminoácido possui dois ângulos 'conformacionais', φ e ψ, e uma vez que estes dois graus de liberdade estejam definidos para cada componente da cadeia peptídica, sua conformação pode ser inferida. A maior parte das combinações possíveis de φ e ψ não são plausíveis simplesmente por impedimentos estéricos, pois ocorrem colisões entre as cadeias laterais e a cadeia principal. As conformações possíveis para cada aminoácido podem ser analisadas no gráfico de pares de ângulos diedros desenvolvido pelo biofísico indiano G. N. Ramachandran [21] (Figura 7).

As proteínas que podemos observar na natureza evoluíram, por pressão seletiva, para desempenhar funções específicas. Suas propriedades funcionais dependem de suas estruturas 3D, que são uma conseqüência direta da sua estrutura linear de aminoácidos que se enovelam formando domínios. Para entender a função biológica de proteínas, devemos ser capazes de deduzir ou predizer sua estrutura 3D a partir da sua seqüência de aminoácidos. Mesmo com todos os esforços feitos nos últimos 25 anos, ainda não conseguimos resolver a questão de como se dá este enovelamento, e esta continua sendo um dos grandes desafios da biologia molecular [22].

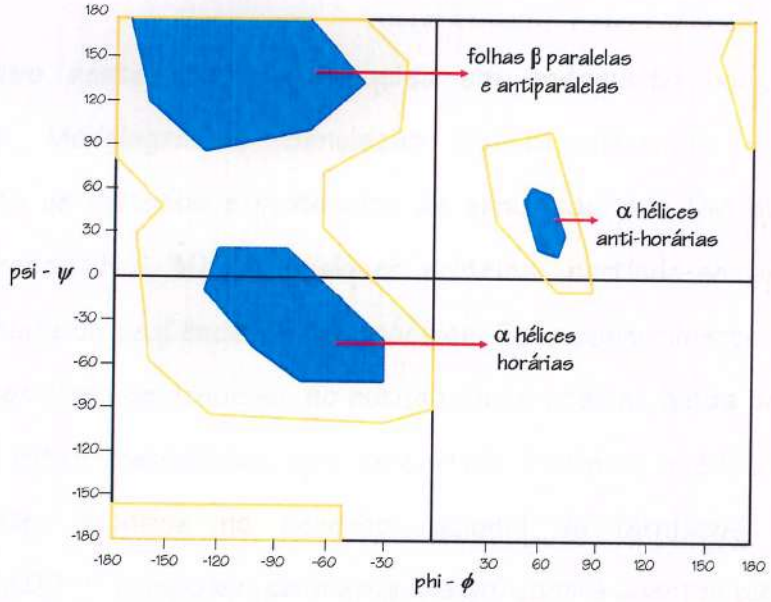


Figura 7. *Gráfico de Ramachandran. As áreas azuis e amarelas representam as combinações de ângulos* ϕ *e* ψ *possíveis para os aminoácidos naturais e as conformações encontradas em cada uma destas áreas. As áreas azuis correspondem às totalmente permitidas (mais favoráveis) e as amarelas correspondem a conformações parcialmente permitidas.*

2. Objetivos

O objetivo desta linha de pesquisa em andamento no Laboratório de Bioinformática, Modelagem e Simulação de Biossistemas - LABIO é o desenvolvimento de métodos e protocolos de simulação pela DM que permitam a predição de estruturas 3D de qualquer proteína, partindo-se apenas da sua estrutura primária ou seqüência de aminoácidos. Esse conhecimento poderá auxiliar na anotação completa de genomas, no estudo das proteínas ainda desconhecidas e envolvidas em rotas metabólicas que despertam interesse médico, e, por fim, a utilização destes modelos no desenho racional de fármacos assistido por computador (CADD — *computer aided drug design*) contra agentes patogênicos.

3. Metodologia

Há duas variações básicas do método de predição *ab initio*, a primeira desenvolve *scoring functions* capazes de distinguir as estruturas protéicas corretas (estado nativo, funcional) das incorretas (estado não-nativo), e a segunda explora o espaço conformacional das proteínas [23, 24]. Nessa segunda categoria se enquadram as simulações de Monte Carlo e da Dinâmica Molecular. A predição *ab initio* baseia-se na hipótese termodinâmica do dobramento ou enovelamento da proteína (*protein folding*), e sugere que a estrutura nativa de uma seqüência protéica corresponde ao estado de energia mínima global da sua energia livre.

Em uma simulação pela DM, as equações clássicas de movimento que governam a evolução temporal, 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 [25]. 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.

Nas simulações pela DM para predição *ab initio* da estrutura de proteínas, o ambiente aquoso (solvente) é representado de forma implícita, como um dielétrico contínuo, utilizando o formalismo denominado Generalized Born (GB) [26, 27]. A adição explícita do solvente 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 (SASA — *surface area solvent access*).

Como resultado, uma simulação pela 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 a pressão, temperatura, volume,

calor específico, pode ser calculado, assim como parâmetros estruturais, incluindo o raio de giro e a estrutura média da proteína [28].

Os programas empregados para a realização das simulações foram os pacotes **AMBER6** [29] e **AMBER7** [30], baseados na mecânica molecular, cujos campos de força incluem as interações de longa distância (van der Waals e eletrostática), ângulos diedros, ângulos de ligação e comprimentos de ligação [31]. Os parâmetros para as interações de van der Waals são determinados a partir de estruturas resolvidas experimentalmente por cristalografia de difração de raios-X, as cargas parciais e ângulos diedros são derivados da teoria quanto-mecânica de distribuição de elétrons [32].

As interações de van der Waals são calculadas de acordo com o potencial de Lennard-Jones [29, 30], que incluem as forças de atração e de repulsão; particularmente úteis durante a simulação para evitar que ocorra sobreposição de átomos.

As interações eletrostáticas, como pontes de hidrogênio e pontes salinas, são extremamente importantes na definição da estrutura e função de uma proteína, porém são difíceis de modelar. Um dos motivos para esta dificuldade reside no fato de estas interações serem dependentes do meio em que a proteína se encontra, ou seja, a eletrostática e as condições de solvatação estão intrinsecamente relacionadas. A lei de Coulomb de interação entre cargas é utilizada para avaliar essas interações.

A acurácia do campo de força a que é submetida uma proteína é determinante, em última análise, da capacidade de predição de estruturas de qualquer método aplicado; pois é o campo de força, através de seus parâmetros, funções estatísticas e potencias, que irá reger o comportamento e evolução temporal do sistema representado.

Simulações pela DM analisam milhares de átomos e calculam seus movimentos, exigindo recursos computacionais de alto desempenho. As simulações

não apenas deste trabalho, mas também de todos os projetos executados em nosso laboratório, foram realizadas nos clusters Amazônia e Ombrófila (localizados no Centro de Pesquisa em Alto Desempenho — CPAD, da Pontifícia Universidade Católica do Rio Grande do Sul, PUCRS), que rodam os programas AMBER6 e AMBER7 em paralelo. Para simulações menores, também foram utilizados uma SGI R14000 de 600MHz e PCs individuais de 2,5GHz. Cada simulação demanda dias, semanas ou até meses dependendo do tamanho da proteína e do tempo de simulação.

4. Resultados e Discussão

4.1. *Three-helix-bundle* — Predição de Estrutura Terciária

Um *three-helix-bundle* é um motivo estrutural composto por duas hélices paralelas e uma terceira antiparalela em relação às demais, comumente encontrado em proteínas citoplasmáticas, transmembranas e extracelulares e em proteínas de ligação ao DNA. O modelo inicial destes experimentos é um *three-helix-bundle* de 65 aminoácidos (RVKALEEKVKALEEKVKALGGGGRIEELKKKWEELKKKIEELGGGGEVKKC EEEVKKLEEEIKKL), sintetizado artificialmente para assumir tal topologia [33], denominado aqui A3.

O modelo 3D canônico do *three-helix-bundle* (A3-1, Figura 8) foi gerado manualmente utilizando-se o programa SPDBviewer [34], baseado no diagrama de empacotamento das cadeias laterais apresentado por Johansson *et al* [33], responsáveis pelo desenho e síntese artificial de A3 (Figura 9). A conformação inicial para a simulação pela dinâmica molecular, com ângulos diedros φ e ψ iguais a 180° para todos os 65 aminoácidos de A3-1, foi gerada com o módulo **tleap** do programa **AMBER 6.0** [29] (Figura 10). Os parâmetros do campo de força Cornell *et al.* [31] foram utilizados para avaliar a energia potencial do sistema. O solvente foi incluído implicitamente através do formalismo GB [26]. As formas neutras dos resíduos de Lisina e Glutamato foram utilizadas para reduzir a carga elétrica total do sistema.

A seqüência linear de A3, constituída de 65 aminoácidos e um total de 1.087 átomos, foi submetida inicialmente a uma simulação pela DM de 10 ns a uma temperatura de 298.16 K (25 °C) e um raio de corte de 10Å para a avaliação das interações eletrostáticas de longo alcance e van der Waals.

Figura 8. Modelo canônico A3-1, gerado manualmente com o programa SPDBviewer, a partir da seqüência linear de aminoácidos e diagrama de empacotamento de cadeias laterais utilizados por Johansson. As hélices foram coloridas esquematicamente, do azul para o vermelho, no sentido N-terminal para C-terminal e representam o motivo estrutural característico de um *three-helix-bundle*, sendo as hélices I (azul) e III (vermelha) paralelas entre si e antiparalelas em relação à hélice II (verde).

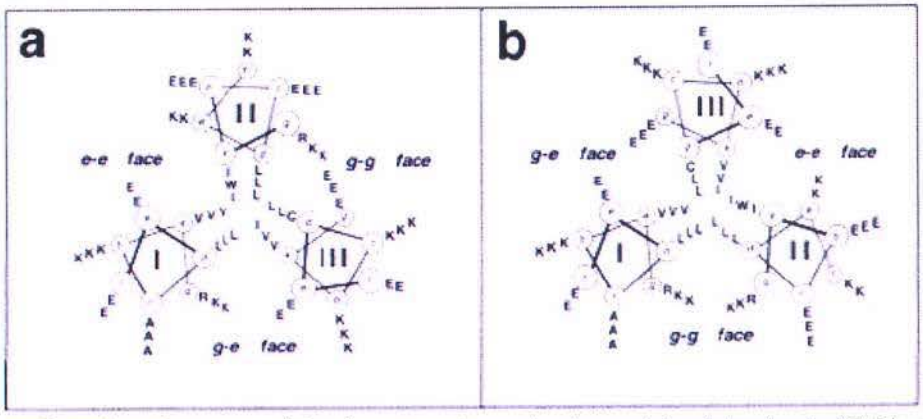


Figura 9. a: desenho esquemático do empacotamento das cadeias laterais de A3 (de onde baseia-se a construção do modelo A3-1). **b**: um *three-helix-bundle* similar, porém de orientação anti-horária. Fonte: Johansson *et al.*, 1998.

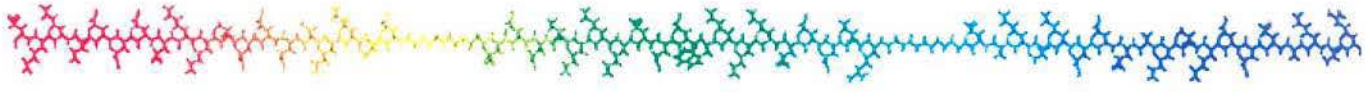


Figura 10. Conformação inicial estendida de A3 para a simulação pela dinâmica molecular, com todos os seus ângulos diedros φ e ψ iguais a 180°.

Nos primeiros **2 ns** pode-se observar a formação da estrutura secundária de A3, constituída de três hélices, separadas por duas alças ricas em Glicina e bastante flexíveis; após **4 ns** de simulação observa-se que as hélices começam a se empacotar, e A3, terminada a simulação inicial de **10 ns**, assume a conformação característica de um *three-helix bundle*, porém com orientação anti-horária, diferente da esperada baseada na sua estrutura canônica horária, A3-1 (Figuras 11 e 12).

Como resultado desta simulação pela DM, obtivemos a estrutura secundária correta (formação das três hélices), e também pudemos observar seu caráter anfifílico, com resíduos de aminoácidos polares voltados todos para um lado, e os apolares para o oposto, como pode ser observado no empacotamento das cadeias laterais hidrofóbicas das hélices I e III (Figura 12), cujos contatos formados condizem com os propostos no modelo canônico (Figura 9). O caráter anfifílico das hélices corrobora a hipótese de que elas tenderão a se "empacotar" em uma estrutura terciária coesa. Resultados de simulações de outras proteínas com estrutura característica de *three-helix bundle* indicam que este motivo estrutural de fato se enovela com a formação de intermediários, onde as hélices II e III (C-terminal) encontram-se empacotadas como no estado nativo antes que se formem contatos com a hélice I (N-terminal) [17, 35].

Outro resultado positivo foi o empacotamento parcial da estrutura terciária e sua convergência para uma conformação de *three-helix-bundle*, o que pode ser observado através do gráfico de RMSD (*root mean square deviation* ou desvio médio quadrático). O desvio médio quadrático é a medida da diferença, ou da similaridade, entre duas conformações, e ao ser medido em função do tempo ele aponta a evolução da estrutura de um estado estendido (mais distante da conformação de um *three-helix-bundle*, valores maiores de RMSD) para um estado enovelado e condizente ao seu modelo (menores valores de RMSD — Figura 13).

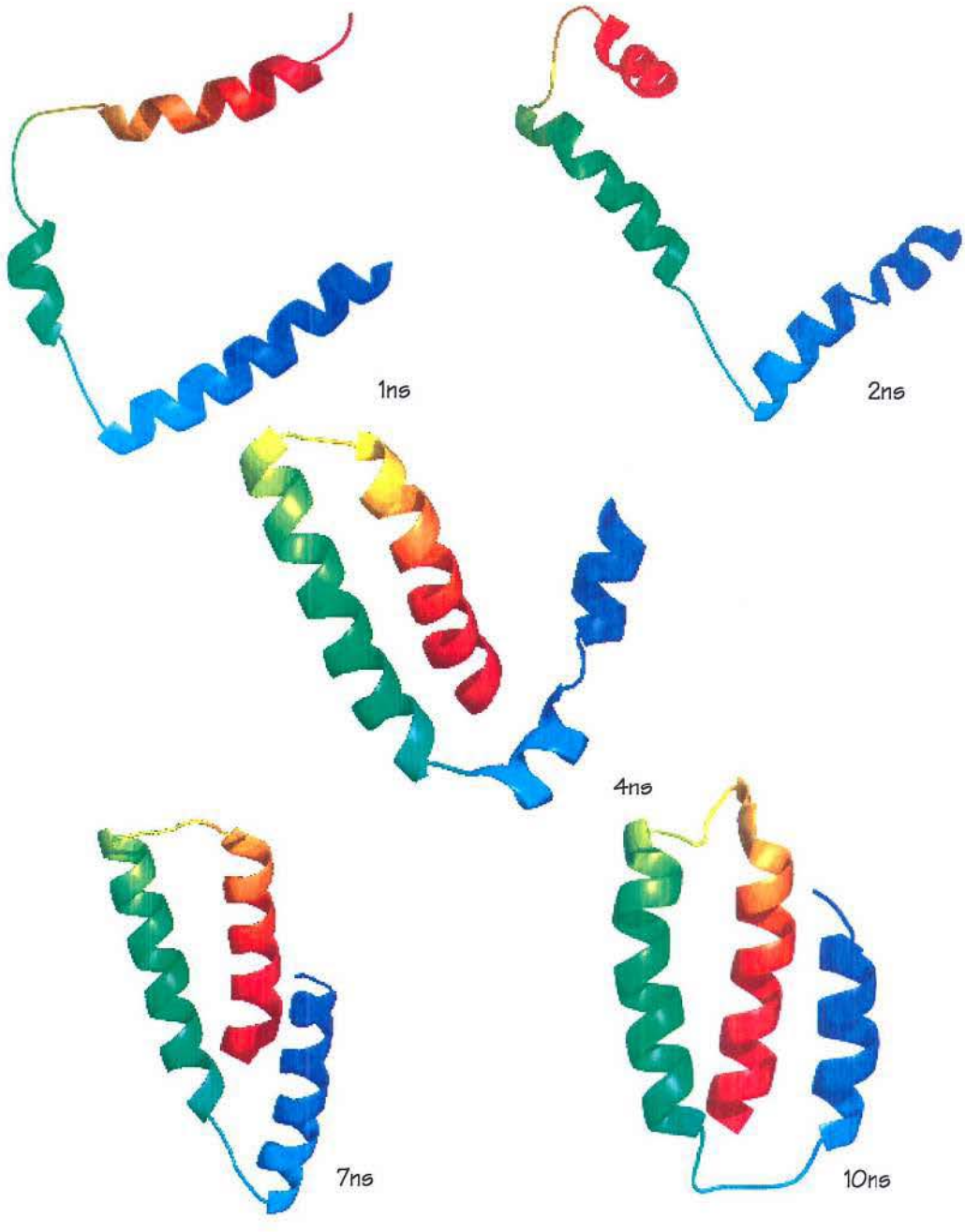


Figura 11. Seqüência de *snapshots* da simulação pela DM de A3. **1 e 2 ns:** Início da formação da estrutura secundária de A3: pode-se observar as três hélices já diferenciadas e separadas pelas alças de glicina. **4 ns:** O empacotamento das hélices começa a ocorrer em função de interações hidrofóbicas e eletrostáticas. Neste snapshot observa-se o empacotamento entre a hélice II (verde) e III (vermelha). **7 ns:** A hélice I (azul) se empacota contra as demais, e A3 começa a adotar uma conformação característica de *three-helix-bundle*. **10 ns:** Snapshot final, terminada a simulação a proteína assume uma conformação de *three-helix-bundle*, porém de orientação anti-horária, diferente da esperada conforme seu modelo A3-1.

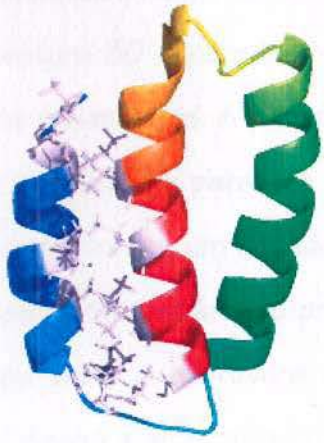


Figura 12. Empacotamento das cadeias laterais hidrofóbicas (representadas em cinza) das hélices I (**azul**) e III (**vermelha**). Mesmo que a estrutura obtida ao término da simulação não seja condizente com seu modelo canônico quanto à orientação das hélices, os contatos destacados nesta figura são condizentes aos contatos descritos no diagrama de empacotamentos de A3 (ver Figura 9).

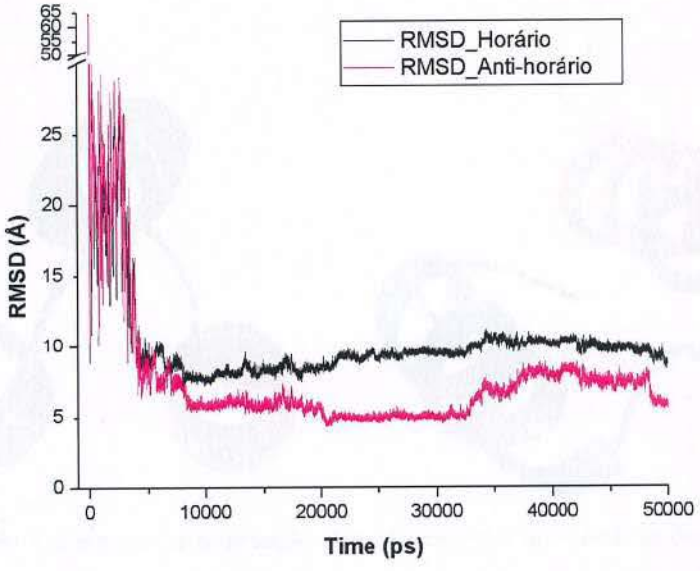


Figura 13. Gráfico de RMSD em função do tempo. A trajetória dinâmica de A3 é comparada ao seu modelo canônico A3 (em **preto**) e também a outro modelo desenhado manualmente (Figura 9b), de orientação anti-horária (em **rosa**). No início da simulação as conformações são bastante distintas, com altos valores de RMSD. Com o passar do tempo este valor se reduz a aproximadamente 7.5Å e 4.5Å, para os modelos horário e anti-horário respectivamente. Os valores de RMSD menores para o modelo anti-horário de A3 são condizentes com a sua conformação final.

Estes resultados foram obtidos num intervalo de tempo de simulação de 10 ns, no qual não se chegou à estrutura 3D esperada. Foram levantadas duas hipóteses para explicar estes resultados preliminares. A primeira é que o intervalo de tempo da simulação pode não ter sido o suficiente para que a proteína atingisse seu estado nativo; é possível que ela se encontre em um estado meta-estável, mínimo local, que pode corresponder a um estado intermediário no processo de seu enovelamento. Na segunda hipótese, a diferença entre a estrutura esperada (modelo canônico) e a obtida em 10 ns (Figura 14) é devida a artefatos do protocolo de simulação utilizado.

O intervalo de tempo da simulação de A3 foi aumentado então para 50 ns, a fim de se verificar a hipótese de a estrutura final obtida em 10 ns ser apenas um intermediário meta-estável no seu processo de enovelamento, pois sabidamente o processo de enovelamento de A3 passa por dois intermediários [36], e a hipótese de a estrutura final obtida em 10 ns de simulação ser um intermediário meta-estável foi levantada.

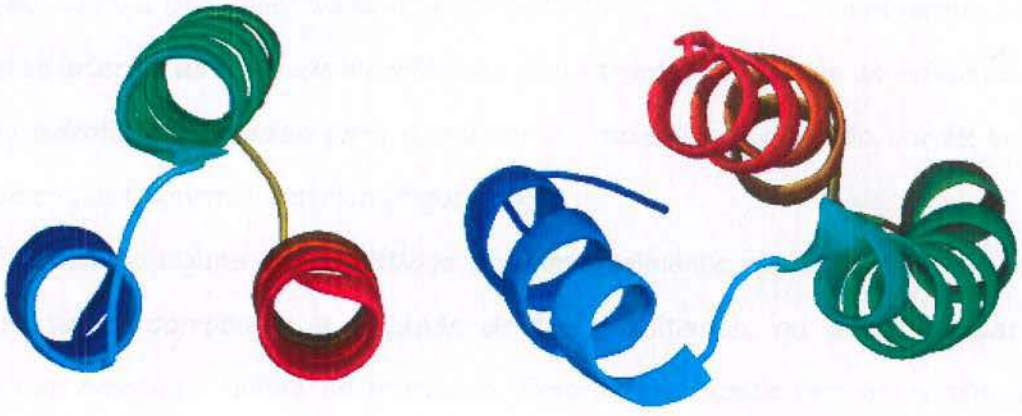


Figura 14. A diferença na orientação das hélices entre o modelo canônico A3-1, horário (à esquerda) e a estrutura obtida após 10 ns de simulação, anti-horária (à direita), pode ser melhor observada neste ângulo. Sendo a hélice I azul, a II verde e a III vermelha, o sentido de I para III percorre um caminho horário no modelo desenhado, diferente do que ocorre em A3 aos 10 ns.

Porém, ao término de 50 ns de simulação, e mesmo ao longo da trajetória, a topologia do modelo não apresentou variações significativas e permaneceu com

orientação anti-horária (Figura 15). A hipótese de termos encontrado um intermediário no processo de enovelamento de A3 foi, portanto afastada.

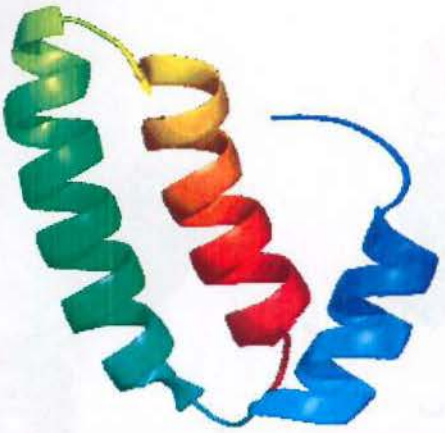


Figura 15. Estrutura de A3 ao final de 50 ns de simulação pela DM. Não houve mudanças significativas em sua conformação a partir de 10 ns, e suas hélices ainda apresentam orientação anti-horária.

Os resultados encontrados nesta nova simulação apontam para problemas na simulação, como a falha observada no empacotamento hidrofóbico do sistema, aonde as cadeias laterais de resíduos hidrofóbicos encontram-se expostas ao solvente e de resíduos hidrofílicos voltadas para o interior da molécula, o que não condiz com a teoria de empacotamento protéico (Figura 16).

Entretanto alguns dados obtidos experimentalmente por Chapeaurouge *et al.* poderiam talvez corroborar a validade destes resultados, ou ainda, ajudar na correção de eventuais falhas de protocolo. Experimentalmente tem-se evidência de que, no sentido de enovelamento, a reação segue um caminho que sai de U (*unfolded*), passa por I_2 (*intermediate 2*), depois por I_1 (*intermediate 1*) e finalmente chega à forma enovelada N (*native*). Os dados de fluorescência do resíduo de triptofano (Trp) sugerem que já ocorre uma certa proteção da exposição deste resíduo ao solvente no estado I_2 em relação à situação totalmente exposta ao solvente em U. O resíduo de Trp está na hélice II, que se encontra empacotada contra a hélice III no snapshot referente aos **4 ns**. Isto protege o TRP da exposição ao solvente, o que poderia sugerir

que o estado parcialmente enovelado mostrado aos 7 ns possa se relacionar a I_2 (Figura 17).

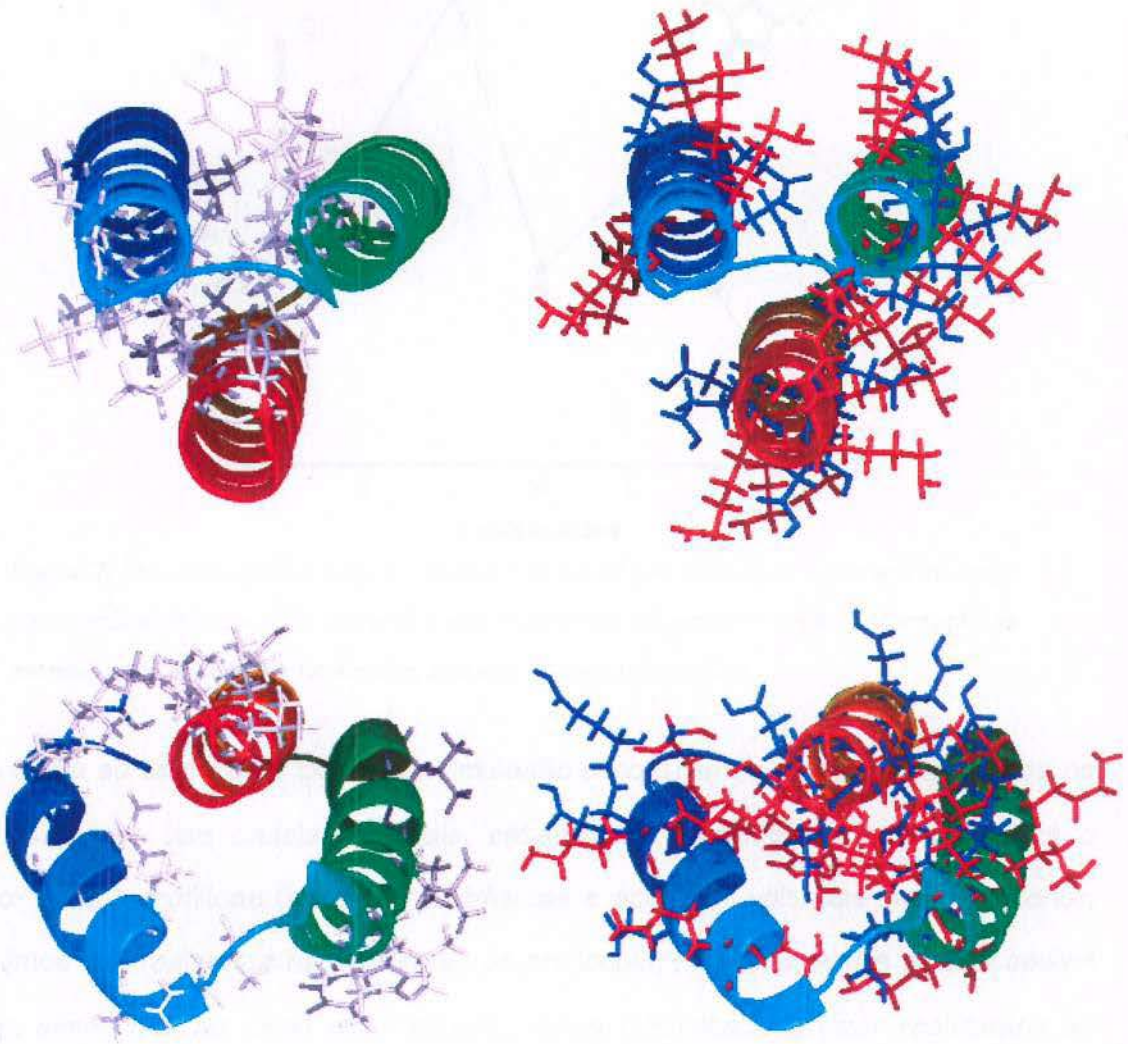


Figura 16. Acima o empacotamento ideal das cadeias laterais de A3 de acordo com seu modelo canônico; e abaixo o empacotamento observado aos 50 ns de simulação. Em cinza estão representadas as cadeias laterais dos resíduos hidrofóbicos, que no empacotamento ideal estão no interior na molécula, protegidos da exposição ao solvente, diferente do que ocorre aos 50 ns, onde apesar de haver um empacotamento parcial das hélices I (**azul**) e III (**vermelha**), há muitos resíduos voltados para fora. Em **azul** e **vermelho** estão representadas as cadeias laterais de resíduos com cargas positivas e negativas respectivamente e, portanto hidrofílicos. Também aqui são observados problemas de empacotamento, com muitos destes resíduos no interior da molécula fazendo interações salinas desfavoráveis que não são observadas no modelo canônico.

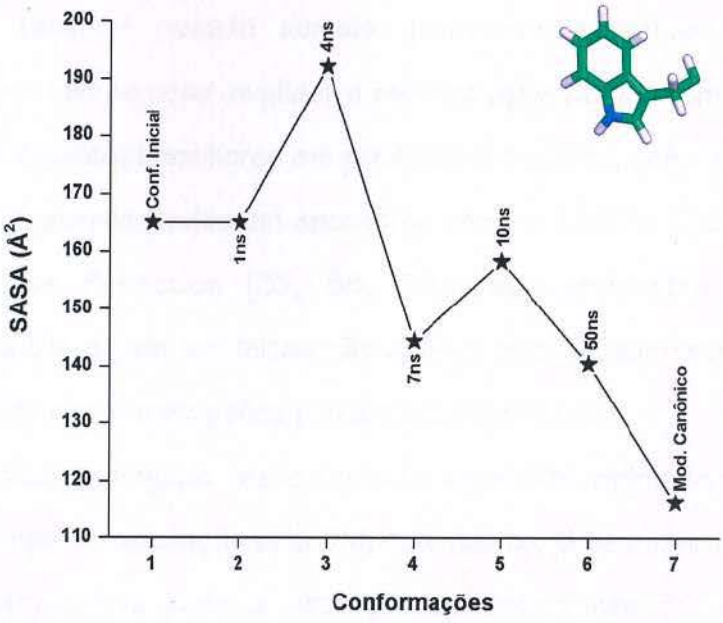


Figura 17. *Área da superfície acessível ao solvente (SASA) do resíduo de Triptofano. Ao longo da trajetória dinâmica o Trp torna-se menos exposto ao solvente em função da formação da estrutura terciária de A3. No detalhe, a cadeia lateral deste resíduo.*

Como ao término de 50 ns de simulação encontramos um grande problema no empacotamento das cadeias laterais, estando as hidrofóbicas voltadas para o exterior e as hidrofílicas (incluindo as básicas e acídicas) voltadas para o interior, concluímos que realmente há problemas de protocolo; e mesmo com a área acessível do Trp, diminuindo ao longo da simulação, seria precipitado utilizar resultados de análises experimentais para confirmação dos resultados por nós obtidos.

Os resultados encontrados com 50 ns de simulação não corroboraram a hipótese levantada de que a estrutura obtida em 10 ns pudesse ser um intermediário do processo de enovelamento de A3. Porém, nesta segunda simulação pela DM, alguns parâmetros foram alterados (temperatura e padrões da geometria do sistema) e efetivamente observamos uma melhora nos resultados obtidos. O sistema colapsa como na simulação anterior de 10 ns, mas o empacotamento hidrofóbico final apresenta-se melhor.

A confrontação de nossos resultados com os dados experimentais levantam várias dúvidas sobre o paralelismo que pode, ou não, ser empregado nestes tipos de comparações. Tanto é ousado demais questionar a metodologia experimental "clássica", como não se pode invalidar o método *ab initio* que vem progressivamente apresentando resultados melhores em predição estrutural, como se pode comprovar pelos resultados apresentados em encontros como o CASP - Critical Assesment of protein Structure Prediction [37, 38, 39]. Isto demonstra que análises de simulações pela DM devem ser feitas não apenas meticulosamente, mas também de modo crítico para evitar conclusões precipitadas e errôneas.

O protocolo empregado neste trabalho necessita refinamentos, como talvez aumento no tempo de simulação pela DM; no entanto, já se mostra promissor como uma técnica alternativa para a obtenção de estruturas 3D de proteínas, ou enzimas. A modelagem de proteínas está evoluindo até o patamar de uma tecnologia prática e seus produtos poderão ser utilizados na terapia, tratamento e mesmo num maior entendimento de inúmeras doenças, e até dos próprios princípios que regem o enovelamento e o comportamento de biossistemas.

O modelo A3 foi temporariamente abandonado em dezembro de 2002, e a partir de então passamos a estudar um sistema natural menor (miniproteína) com estrutura 3D experimentalmente determinada, o que elimina dúvidas sobre a validade do modelo canônico contra o qual são comparados os resultados obtidos. As atuais simulações em andamento no LABIO têm apresentado resultados extremamente positivos e achamos no momento ser melhor insistir nestes modelos, que também têm a vantagem de exigirem simulações mais curtas. Com base nos resultados que estamos encontrando com miniproteínas e a futura padronização de protocolo, sistemas maiores, como A3, serão retomados.

4.2. *Alpha-helical-hairpin* — Predição de Estrutura Secundária e Supersecundária

Um *alpha-helical-hairpin* é um domínio protéico constituído de duas hélices antiparalelas conectadas por uma volta (*turn*). O *helical-hairpin* aqui estudado, denominado PA_Z, corresponde ao domínio Z obtido através da minimização do domínio B da proteína A de *Staphylococcus aureus*. O processo de minimização [40] reduziu a estrutura original (um *three-alpha-helix-bundle* de 59 resíduos) a um *hairpin* helicoidal de 33 resíduos: FNMQQQRRFYEALHDPNLNEEQRNAKIKSIRDD (Figura 18).

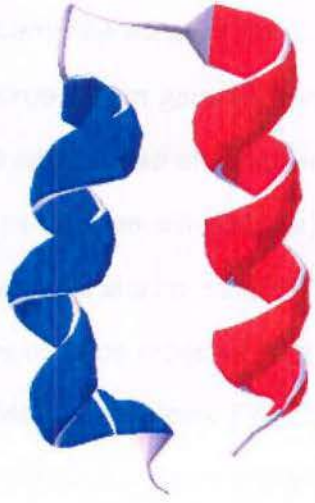


Figura 18. Estrutura do *alpha-helical-hairpin* 1zdb, determinada experimentalmente por NMR, Starovasnik *et al.* [41]. As duas hélices que o constituem estão coloridas esquematicamente, do azul para o vermelho, da porção N-terminal para C-terminal.

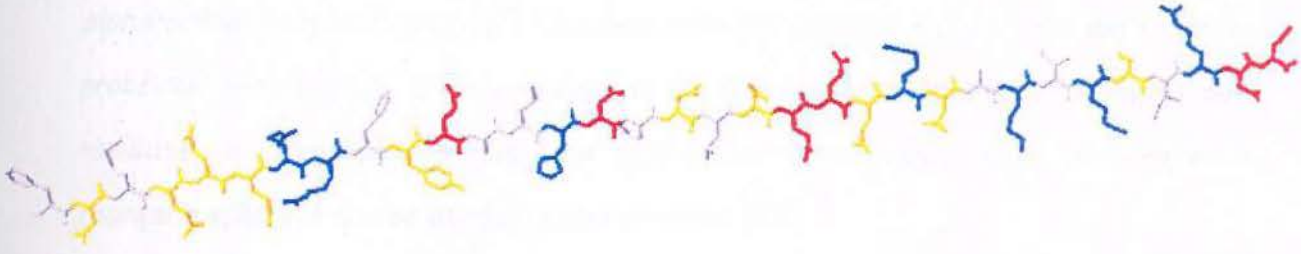


Figura 19. Conformação inicial estendida de PA_Z, com ângulos diedros ϕ e ψ iguais a 180° para todos seus aminoácidos. A cadeia polipetídica é orientada da esquerda para a direita, da porção N-terminal para C-terminal. Resíduos coloridos conforme o tipo de aminoácido, em cinza aminoácidos apolares, em amarelo polares, em azul os aminoácidos de carga positiva e em vermelho de carga negativa.

Diferentes simulações pela DM foram geradas, a 281K, todas começando com uma conformação inicial totalmente estendida de PA_Z, com os ângulos diedros da cadeia principal φ e ψ iguais a 180° (Figura 19). A estrutura experimental de PA_Z, determinada pelo método de NMR (código PDB 1zdb), foi adotada como nossa estrutura modelo, e parâmetro de comparação para nossos resultados. A conformação inicial para a simulação pela DM, com φ e ψ iguais a 180° foi gerada com o módulo **tleap** do programa **AMBER 7.0** [30]. Os parâmetros do campo de força de Cornell [31] foram utilizados para avaliar a energia potencial do sistema, e o solvente foi incluído implicitamente através do formalismo de **GB** [26]. Modelos com resíduos carregados e neutros foram gerados para confrontação dos dados obtidos e análises da influência das cargas nas simulações pela DM.

Uma vez que foram muitas as simulações geradas tendo PA_Z como modelo, os dados apresentados neste trabalho estarão focados em uma única simulação, PA_Z-10.4, cujos resultados obtidos representam satisfatoriamente o conjunto total de simulações, e será referido apenas como PA_Z.

Nossos protocolos permitem a visualização do início de formação de núcleos de hélices já nos primeiros nanosegundos de simulação pela DM. A formação das hélices é um dos primeiros fenômenos a ser observado no processo de enovelamento, seguida pelo colapso das mesmas e a formação da estrutura terciária característica de um *alpha-helical-hairpin* (Figura 20). O mesmo comportamento é observado em modelos protéicos semelhantes, onde a trajetória de enovelamento envolve um colapso dos resíduos e formação parcial da estrutura secundária, com subseqüente reorganização até que se atinja o estado nativo [35].

A evolução temporal da trajetória dinâmica de PA_Z pode ser analisada numericamente pelo gráfico de RMSD para os átomos de carbono α de todos os resíduos de aminoácidos do polipeptídeo (Figura 21), onde valores iniciais elevados refletem a estrutura desenovelada e valores gradualmente mais baixos mostram a formação da sua estrutura supersecundária até a estabilização da estrutura de

PA_Z, com valores em torno de 3.0Å. Uma vez estável, a estrutura obtida por simulação pela DM foi sobreposta a sua equivalente experimental (Figura 22), onde se pode observar a similaridade topológica entre ambas.

O mesmo cálculo de valores de RMSD, ou seja, da similaridade topológica entre as estruturas, foi feito para os resíduos de aminoácidos constituintes de cada uma das hélices individualmente. Os valores observados tanto para hélice I (Figura 23) quanto para a hélice II (Figura 24) são de aproximadamente 1.5 Å, significativamente inferior ao valor final de RMSD observado para toda estrutura de PA_Z.

Diferente do que é observado na Figura 21, os valores de RMSD para ambas as hélices já são baixos desde os primeiros instantes da simulação pela DM de PA_Z (Figuras 23 e 24), resultado condizente com o que é mostrado na Figura 20, pois a formação da estrutura secundária deste polipeptídeo ocorre muito cedo na sua trajetória de enovelamento.

A discrepância observada entre os valores de RMSD para todos resíduos de aminoácidos da estrutura e para suas hélices individualmente fez surgirem duas hipóteses para justificar a má formação da estrutura supersecundária de PA_Z: falha no empacotamento de suas cadeias laterais, afetando a orientação das hélices, ou falha na formação da alça (*turn*) de ligação entre as hélices.

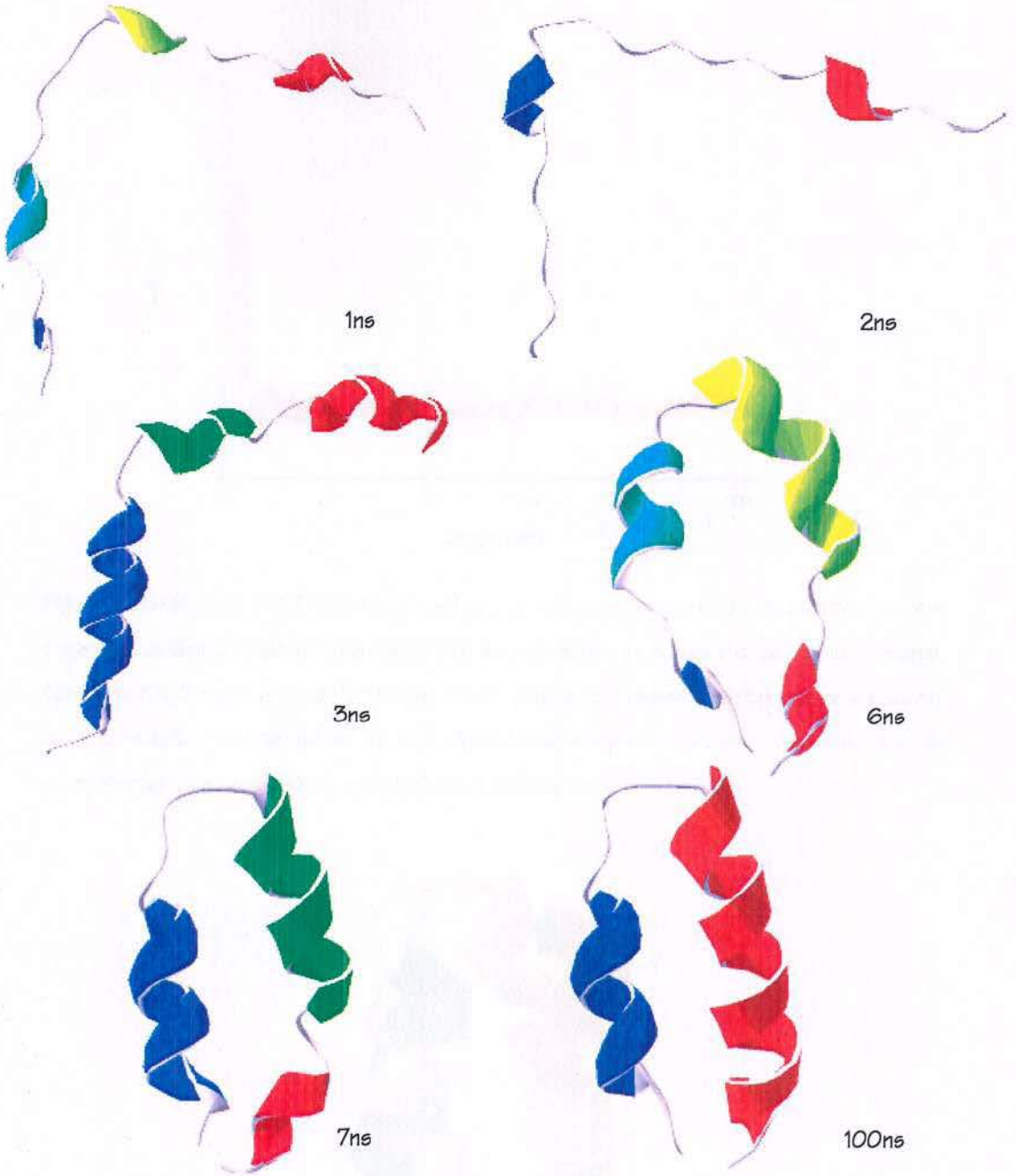


Figura 20. *Snapshots* de uma das simulações de PA_Z. Da mesma forma que o ocorrido nas simulações de A3, os núcleos de estruturas secundárias surgem nos primeiros nanosegundos de simulação (**1 e 2 ns**), seguidos da formação das hélices (**3ns**) e colapso hidrofóbico do sistema, que começa a adquirir a conformação de um *alpha-helical-hairpin* (**6 e 7ns**). A partir de então PA_Z se mantém estável até o fim desta simulação pela DM, aos **100 ns**, onde pode-se observar as duas hélices bem formadas e individualizadas (hélice I em **azul** e II em **vermelho**), separadas por uma pequena volta. Como na Figura 18, as estruturas estão coloridas do **azul** para o **vermelho**, da porção N-terminal para C-terminal.

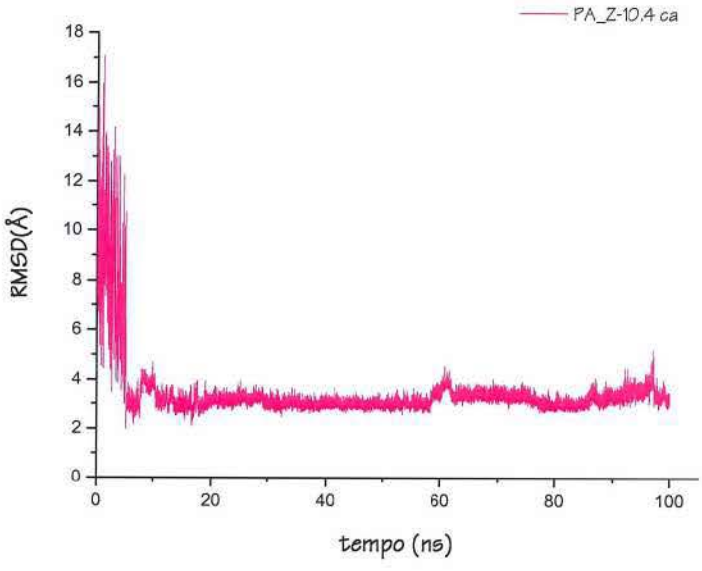


Figura 21. Gráfico de RMSD (*root mean square deviation* — desvio médio quadrático) para a trajetória dinâmica mostrada na Figura 20, apenas para os carbonos **α** da cadeia principal. Conforme PA_Z evolui ao longo do tempo no seu processo de enovelamento, passa a assumir a conformação característica de um *alpha-helical-hairpin* congruente com seu modelo experimental 1zdb, com valores de RMSD estabilizados em torno de 3.0Å.

Figura 22. Sobreposição da estrutura determinada experimentalmente (colorida como na Figura 18) e de PA_Z (em laranja), aos 100 ns da simulação apresentado na Figura 20. A sobreposição mostra a similaridade topológica que existe entre a estrutura modelada e a determinada por NMR.

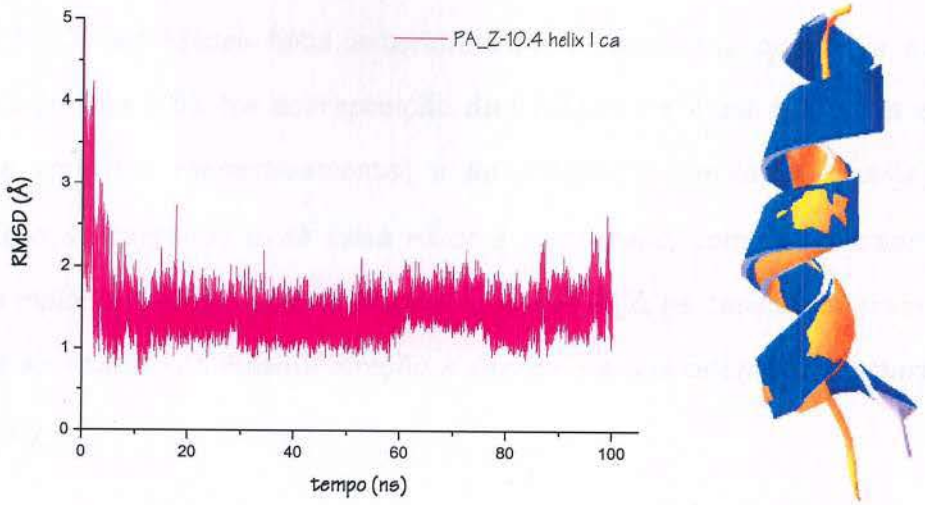


Figura 23. *Gráfico de RMSD para os átomos de carbono α dos resíduos de aminoácidos constutuintes da hélice I, com valores em torno de 1.5 Å. À direita, a sobreposição da hélice I da estrutura resolvida por NMR, colorida como na Figura 18, e da hélice I da estrutura gerada por simulação (laranja), no instante final da sua trajetória, aos 100ns.*

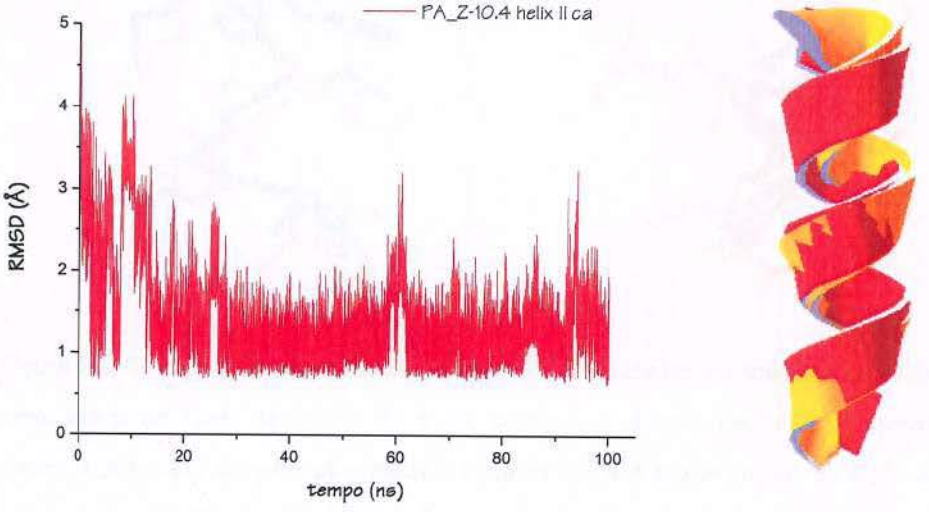


Figura 24. *Como na figura anterior, gráfico de RMSD para os átomos de carbono α dos resíduos de aminoácidos constutuintes da hélice II, com valores em torno de 1.5 Å. À direita, a sobreposição da hélice II da estrutura resolvida por NMR, colorida como na Figura 18, e da hélice II da estrutura gerada por simulação (laranja), aos 100ns da simulação. Os baixos valores de RMSD (~1.5 Å) mostrados na Figura 23, e nesta, se refletem na sobreposição quase perfeita de ambas as hélices.*

Para analisar a orientação das cadeias laterias de PA_Z, a mesma sobreposição de hélices feita anteriormente foi analisada quanto a suas cadeias laterais (Figura 25). Na sobreposição das hélices I e II da estrutura experimental (azul e vermelha respectivamente) e da estrutura simulada (laranja), apenas o esqueleto de carbonos α de cada hélice é mostrado, como suas cadeias laterais. Para a maioria dos resíduos de aminoácidos de PA_Z, as cadeias laterais "saem" das hélices apontando na mesma direção e diferem na sua orientação a partir dos seus ângulos χ_2.

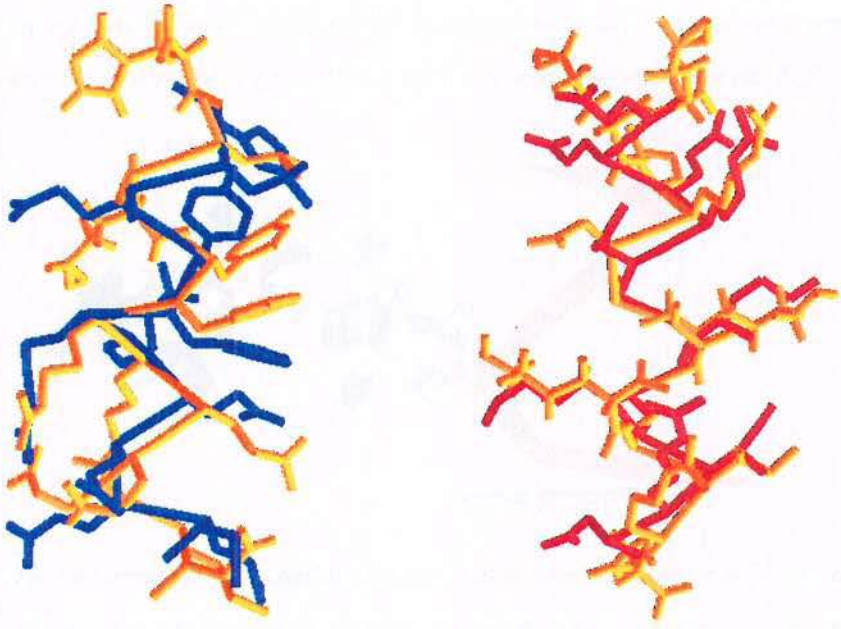


Figura 25. Sobreposição das cadeias laterais dos resíduos de aminoácidos de cada hélice, separadamente como na Figura 23 e 24, porém aqui apenas as cadeias laterais e os carbonos α estão ilustrados para maior clareza da figura. À esquerda, sobreposição da hélice I de 1zdb e de PA_Z, coloridas como na Figura 23, e à direita, sobreposição da hélice II, com as hélices coloridas como na Figura 24. As variações de posição destas cadeias laterais ocorrem nos ângulos de rotação entre os átomos que as compõem, a partir de χ_2, e não nos ângulos que são definidos pela cadeia principal (χ_1, as cadeias laterais "saem" da cadeia principal em direções congruentes).

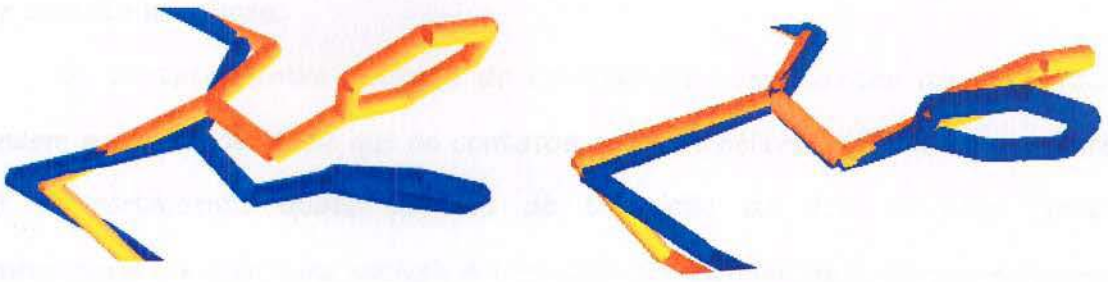


Figura 26. Visão mais detalhada da orientação de uma das cadeias laterais da hélice I, resíduo de fenilalanina, Phe9. À esquerda sobreposição do átomo de carbono dos resíduos da cadeia principal (RMSD = 1.35 Å) e à direita, o mesmo resíduo, com sobreposição dos átomos constituintes da cadeia lateral (RMSD = 0.07 Å), mostrando a diferença na sua orientação a partir apenas do ângulo diedro χ_2; o mesmo ocorre nos demais aminoácidos de PA_Z.

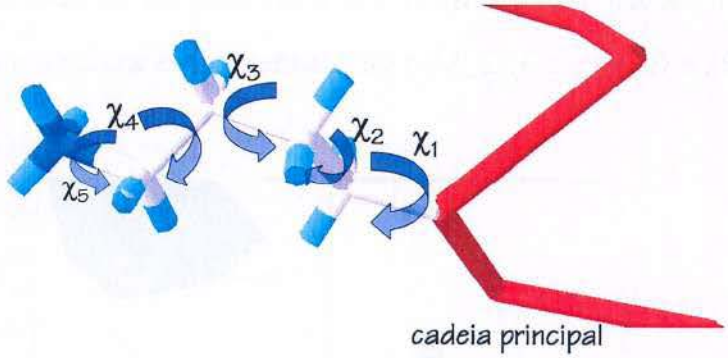


Figura 27. As conformações das cadeias laterais, assim como as conformações da cadeia principal, são descritas por ângulos de rotação interna (ângulos de torção), denominados χ (chi) $_1$, χ_2, χ_3 ... Cadeias laterais diferentes possuem quantidades de graus de liberdade diferentes, aqui é mostrada a cadeia lateral do resíduos de lisina (Lys26), que encontra-se na hélice II, e possui cinco graus de liberdade de rotação. As conformações das cadeias laterais tendem a serem conservadas; resíduos homólogos, em proteínas relacionadas, têm conformação similar das cadeias laterais em decorrência das interações que ocorrem com os resíduos vizinhos [42].

Posto que a orientação das cadeias laterais, e consequentemente as interações entre aminoácidos vizinhos nas hélices, não possui diferenças significativas entre a estrutura experimental — 1zdb, e a que foi gerada pela

simulação pela DM, passamos à análise dos resíduos constituintes da volta (*turn*) que conecta as hélices.

Os contatos entre resíduos de aminoácidos constituintes da volta (*turn*) tendem a ser menos lábeis que os contatos entre as hélices, além de apresentarem um comportamento quase perfeito de transição de dois estados (estado desenovelado → estrutura nativa). A formação correta destes contatos é essencial para que motivos estruturais do tipo *alpha-helical-hairpins* e *three-helix-bundles* atinjam suas estruturas supersecundária e terciárias corretamente, pois as ligações entre estes aminoácidos determinam a orientação espacial final das hélices [17].

A distribuição dos ângulos diedros phi e psi, no gráfico de Ramachandran, dos resíduos de aminoácidos constituintes da volta mostra claramente a diferença existente entre a estrutura experimental 1zdb e PA_Z (Figuras 28 e 29).

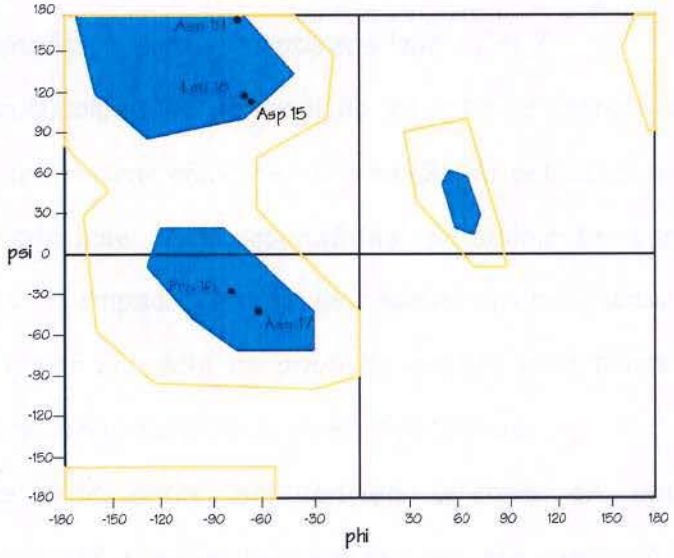


Figura 28. Gráfico de Ramachandran com a distribuição dos resíduos de aminoácidos constituintes da volta (*turn*) — Asp15, Pro16, Asn17, Leu18 e Asn19, da estrutura experimental 1zdb.

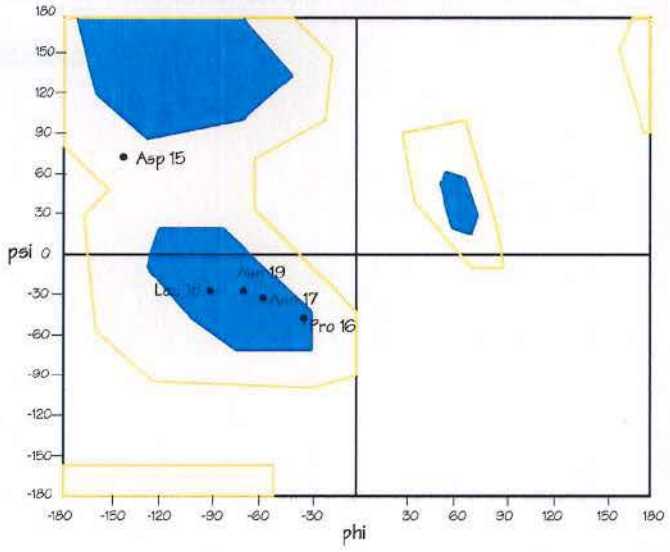


Figura 29. *Gráfico de Ramachandran com a distribuição dos resíduos de aminoácidos constituintes da volta (turn) — Asp15, Pro16, Asn17, Leu18 e Asn19, da estrutura gerada computacionalmente por simulação pela DM, PA_Z.*

A acentuada diferença entre os ângulos φ e ψ destes resíduos pode ser responsável pelos altos valores de RMSD observados quando comparados todos resíduos de aminoácidos dos polipeptídeos 1zdb e PA_Z.

Nosso protocolo inicial de predição *ab initio* de estruturas por simulação pela DM apesar de se mostrar eficiente na predição de estruturas secundárias corretas (hélices) e estruturas supersecundárias parcialmente corretas (*alpha-helical-hairpin*), incluindo o empacotamento de cadeias laterais; ainda requer refinamentos que permitam maior acurácia na predição destas estruturas antes que se possa aplicá-lo a polipeptídeos maiores ou mesmo proteínas.

A similaridade entre as cadeias laterais da estrutura experimental determinada por NMR e as estruturas obtidas por simulação da DM são um bom indicativo de que estamos de fato no caminho correto para a obtenção de modelos com menores valores de RMSD e melhor empacotamento dos seus microdomínios (estruturas secundárias).

Avaliações mais detalhadas das simulações de PA_Z, considerando-se as variações de alguns parâmetros e suas conseqüências na trajetória dinâmica do

polipeptídeo estão em andamento em nosso laboratório e constituem parte de meu projeto de mestrado.

5. Conclusão

O volume de dados depositados em bancos de dados tais como *GenBank* e *Protein Data Bank* (Figuras 1 e 3) [43 e 44] vem dobrando de tamanho em média a cada quinze meses; e como resultado deste fenômeno, computadores se tornaram indispensáveis nas pesquisas biológicas. A bioinformática é freqüentemente definida como a aplicação de técnicas computacionais para entender e organizar informações relacionadas a macromoléculas biológicas [2], sendo ferramenta indispensável nos estudos de genômica estrutural.

Nesta era pós-genômica, o entendimento da relação entre seqüência e estrutura de proteínas tem um papel crucial, com grande impacto nas pesquisas genéticas, bioquímicas e farmacêuticas [45, 46 e 47]; além da sua importância no maior entendimento das doenças relacionadas a mutações pontuais [48] e no desenho racional de fármacos assistido por computador.

Enquanto os métodos de modelagem comparativa são limitados a famílias de proteínas que possuem ao menos uma proteína conhecida, a predição *ab initio* não possui tal limitação e é uma ferramenta particularmente útil para caracterizar novas famílias e ou motivos estruturais. Estimativas do New York Structural Genomics Research Consortium [49] apontam que, para cada nova seqüência caracterizada, em média 100 outras seqüências protéicas poderão ser também modeladas. Estes valores ilustram e justificam as premissas da genômica estrutural e do estudo destas seqüências. Estudos de uma única proteína modelo, ou de seus domínios, em associação com a predição de estruturas irão contribuir substancialmente para a caracterização eficiente de grandes conjuntos de macromoléculas, além de prover importantes informações funcionais a partir das estruturas determinadas [50].

Os protocolos de predição *ab initio* de estruturas tridimensionais de proteínas que estamos testando e padronizando em nosso laboratório pretendem um melhor entendimento dos processos que regem a formação das estruturas supersecundárias e terciárias de proteínas para que possamos predizê-las

corretamente. Os resultados aqui apresentados e discutidos indicam que estamos de fato no caminho correto para atingir este objetivo; os problemas que encontramos referem-se principalmente às cargas dos aminoácidos dos modelos estudados, e estão sendo parcialmente contornados com a adoção de resíduos neutralizados. Resultados desta mesma metodologia de predição de estruturas encontrados na literatura referem-se apenas a modelos cujos resíduos são predominantemente hidrofóbicos (não carregados) e para modelos protéicos ainda menores [15]; e esses resultados foram reproduzidos sem maiores problemas em nosso laboratório.

Análises das diferentes simulações feitas do modelo PA_Z ainda estão em andamento no LABIO e fazem parte do projeto de mestrado, assim como estudos de outros modelos protéicos constituídos por outros motivos estruturais além de *alpha* hélices apenas. Simulações preliminares de um modelo *beta-hairpin* vêem apresentando resultados positivos e serão retomados no decorrer de 2004. Motivos estruturais mais complexos, constituídos de associações de domínios *alpha* e *beta*, o objetivo maior desta linha de pesquisa, serão abordados uma vez que tenhamos um protocolo de simulação bem padronizado; algumas simulações-teste feitas com um dos domínios da proteína G de *Streptococcus* (código PDB 1igd) e com a proteína inibidora de quimiotripsina 2 de cevada (código PDB 3ci2), ambas compostas de *alpha* hélices e folhas *beta* apresentaram resultados iniciais muito promissores e encorajadores para a continuidade deste trabalho.

6. Bibliografia

[1] Norin, M.; Sundström, M. **Structural proteomics: lessons learnt from the early case studies.** *Il Farmaco*, 57: 947-951, 2002.

[2] Luscombe, N.M.; Greenbaum, D. and Gerstein, M. **What is Bioinformatics? A proposed definition and overview of the field.** *Method Inform Med*, 40:346-358, 2001.

[3] Blundell, T.L.; Mizuguchi, K. **Structural genomics: an overview.** *Prog Biophys. Mol. Biol.*, 73:289-295, 2000.

[4] Adams, M.D. *et al.* **The genome sequence of *Drosophila melanogaster*.** *Science*, 287:2185-2195, 2000.

[5] Rubin, G.M. *et al.* **Comparative genomics of the eukaryotes.** *Science*, 287:2204-2215, 2000.

[6] Vitkup, D. *et al.* **Completeness in structural genomics.** *Nat. Struct. Biol.*, 8: 559-566, 2001.

[7] Smith, C.M. **Molecular modeling in the genomics era.** *The Scientist*, 15: 24-17, 2001.

[8] Moult, J. **Predicting protein three-dimensional structure.** *Curr. Opin. Biotechnol.*, 10:583-588, 1999.

[9] Anfinsen, C.B. **Studies on the principles that govern the folding of protein chains.** Nobel Lecture, December 11, 1972.

[10] Osguthorpe, D.J. ***Ab Initio* protein folding.** *Current Opinion in Structural Biology*, 10:146-152, 2000.

[11] Bonneau, R.; Baker, D.A. **Ab initio protein structure prediction: progress and prospects** . *Annu. Ver. Biophys. Biomol. Struct.*, 30:173-189, 2001.

[12] Levinthal, C. **Mossbauer Spectroscopy in Biological System.** Proceedings of a meeting held at Allerton House, Monticello, IL. Eds. Debrunner P., Tsibris J. C. M. and Munck E. University of Illinois Press, pp. 22-24, 1969.

[13] Geney R, Simmerling C, Ojima I. **Computational analysis of the paclitaxel binding site in alpha-tubulin.** *The American Chemical Society*, 222: 65, 2001.

[14] Xia, Y.; Huang, E. S.; Levitt, M. and Samudrala, R. **Ab Initio Construction of Protein Tertiary Structures Using a Hierarchical Approach.** *Jounal of Molecular Biologyl.* 300: 171-185, 2000.

[15] Chowdhury, S.; Lee, M. C.; Xiong G. and Duan Y. ***Ab Initio* Folding Simulation of the Trp-cage Mini-protein Approaches NMR Resolution,** *Journal of Molecular Biology.* 327:711-717, 2003.

[16] Onufriev, A.; Case D. A. and Bashford D. **Structural Details, Pathways, and Energetics of Unfolding Apomyoglobin.** *Jounal of Molecular Biology.* 325: 555—567, 2003.

[17] Shakhnovich, E. I. and Berriz, G. F. **Characterization of the folding kinetics of a three-helix bundle protein via a minimalist Langevin model.** *Journal of Molecular Biology.* 310: 673-685, 2001.

[18] Pauling, L. & Corey, R. B. **Atomic Coordinates and Structure Factors for Two Helical Configurations of Polypeptide Chains.** *PNAS.* 37:235-240, 1951.

[19] Pauling, L. & Corey, R. B. **The Structure of Synthetic Polypeptides.** *PNAS.* 37:241-250,1951.

[20] Stryer, L. **Bioquímica** — Stanford University. *Editora Guanabara Koogan*, 4ª ed., 1996.

[21] Ramachandran, G.N.; Saisekharan V. **Conformation of polypeptides and proteins.** *Adv. Protein Chem.* 26:283-437, 1968.

[22] Branden, C.; Tooze, J. **Introduction to Protein Structure.** *Garland Publishing*, 1998.

[23] Karplus, M. **Aspects of protein reaction dynamics: deviations from simple behavior.** *J. Phys. Chem. B*, 104: 11-27, 2000.

[24] Tukerman, M.E.; Martyna, G.J. **Understanding modern molecular dynamics: techniques and applications.** *J. Phys. Chem. B*, 104: 159-178, 2000.

[25] Van Gunsteren, W.F.; Berendsen, H.J.C. **Computer simulation of molecular dynamics: methodology, applications, and perspectives in chemistry.** *Angew. Chem. Int. Ed. Engl.*, 29: 992-1023, 1990.

[26] Bashford, D.; Case, D.A. **Generalized Born models of macromolecular solvation effects.** *Annu. Ver. Phys. Chem.*, 51:129-152, 2000.

[27] 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.** *J. Phys. Chem. B*, 102:9571-9756, 1998.

[28] Norberto de Souza, O.; Ornstein, R.L. **Molecular dynamics simulations of a protein-protein dimer: particle-mesh Ewald eletrostatic model yields far superior results to standard cutoff model.** *Journal of Biomolecular Structure & Dynamics*, 16: 1205-1217, 1999.

[29] Case, D.A; Pearlman, D.A.; Caldwell, J.W.; Cheathem, T.E., III; Ross, W.R.; Simmerling, C.L.; Darden; T.A.; Merz, K..M.; Stanton, R.V.; Cheng, A.L.; Vincent, J. J.; Crowley, M.; Tsui,V.; Radmer, R.J.; Duan,Y.; Pitera, J.; Massova, I.; Seibel, G. L.; Singh, U.C.; Weiner, P. K. and Kollman, P. A., **AMBER 6.0**. University of California, San Francisco, 1999.

[30] Case, D.A.; Pearlman, D.A.; Caldwell, J.W.; Cheatham III, T.E.; Wang, J.; Ross, W.S.; Simmerling, C.L.; Darden, T.A.; Merz, K.M.; Stanton, R.V.; Cheng, A.L.; Vincent, J.J.; Crowley, M.; Tsui, V.; Gohlke, H.; Radmer, R.J.; Duan, Y.; Pitera, J.; Massova, I.; Seibel, G.L.; Singh, U.C.; Weiner, P.K. and Kollman, P.A. **AMBER 7**. University of California, San Francisco, 2002.

[31] Cornell, W.D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz , Jr., K. M.; Ferguson, D. M.; Spellmeyer, D.C.; Fox, T.; Caldwell, J.W.; Kollman, P. A A. **A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules.** *J. Am. Chem. Soc.* 117:5179-5197, 1995.

[32] Pokala, N.; Handel, T.M. **Review: Protein design — where we were, where we are, where we're going.** *Journal of Structural Biology.* 134:269-281, 2001.

[33] Johansson, J.S.; Gibney, B.R.; Skalicky, J.J.; Wand, A.J.; Dutton, P.L. **A Native-like *Three-α-helix-bundle* protein structure-based redesign: A novel maquette scaffold.** *J. Am. Chem. Soc*, 120: 3881-3886, 1998.

[34] Guex, N.; Peitsch, M.C. **SWISS-MODEL and The Swiss-PdbViewer: An environment for comparative protein modeling.** *Eletrophoresis*, 18:2714-2723, 1997. <http://www.expasy.ch/spdbv>

[35] Zhou, Y. & Karplus, M. **Interpreting the folding kinetics of helical proteins.** *Nature*, 401: 400-403, 1999.

[36] Chapeaurouge, A.; Johansson, J.S.; Fereira, S.T. **Folding intermediates of a model *Three-helix-bundle* protein.** *The Journal of Biological Chemistry*, 276:14861-14866, 2001.

[37] Bonneau, R., Tsai, J., Ruczinski, I., Baker, D. **Functional inferences from blind *ab initio* protein structure prediction.** *Journal of Structural Biology*, 134: 186-190, 2001.

[38] Sternberg, M.J.E., Bates, P.A., Kelley, L.A., MacCallum, R.M. **Progress in protein structure prediction: assessment of CASP3.** *Current Opinion in Structural Biology*, 9: 368-373, 1999.

[39] **CASP — Protein Structure Prediction Center** website: <http://predictioncenter.llnl.gov >

[40] Braisted, A. and Wells, JA. **Minimization of a Binding Domain of Protein A.** *PNAS*, 93, 5688-5692, 1996.

[41] Starovasnik, M. A.; Braisted, A. C.; Wells, J. A. **Structural mimicry of a native protein by a minimized binding domain.** *PNAS,* 94: 10080, 1997.

[42] Falcão, P.K.; Baudet, C.; Higa, R.H. & Neshich, G. **Incorporação das propriedades rotâmeros e ocupância em métodos de análise estrutural de proteínas.** Comunicado Técnico — Ministério da Agricultura Pecuária e Abastecimento, 34, 2002.

[43] Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne P.E. **The Protein Data Bank.** *Nucleic Acids Research*, 235-242, 2000.

[44] Benson, D.A.; Karsch-Mizrachi, I.; Lipman, D.J.; Ostell, J. & Wheeler, D,L. **GenBank.** *Nucleic Acids Research*, 31(1):23-7, 2003.

[45] Dobson, C.M.; Sali, A. & Karplus, M. **Protein folding: a perspective form theory and experiment.** *Angew. Chem.*, 37: 868-893, 1998.

[46] Dill, A. & Chan, H.S. **From Levinthal to pathways to funnels.** *Nature Structural Biology*, 4: 10-19, 1997.

[47] Brooks, C.L. III; Gruebele, M.; Onuchic, J.N. & Wolynes P.G. **Chemical physics of protein folding.** *PNAS,* 95: 11037-11038, 1998.

[48] Prusiner, S.B. **Prions.** *PNAS,* 95: 13363-13383, 1998.

[49] New York Structural Genomics Researsh Consortium: <http://nysgc.org/.>, acessado em 05 de Janeiro de 2004.

[50] Baker, D. & Sali, A. **Protein Structure Prediction and Structural Genomics.** *Science*, 294: 93-96, 2001.