Tese de Mestrado

Estudo de Enovelamento de Polímeros em Rede Quadrada Bidimensional

Celina Maria de Souza Costa

CENTRO BRASILEIRO DE PESQUISAS FÍSICAS
Rio de Janeiro, junho de 1998

Tese de Mestrado

Estudo de Enovelamento de Polímeros em Rede Quadrada Bidimensional

Celina Maria de Souza Costa

Tese submetida ao Departamento de Matéria Condensada e Espectroscopia
como requisito para obtenção do grau
de mestre em Física.

Orientador
Fernando de Magalhães Coutinho Vieira
Co-orientador
Léa Jaccoud El-Jaick

A minha mãe

Agradecimentos

A todos aqueles que de alguma maneira colaboraram com este trabalho. Em especial

- ao Fernando pelas discussões, por estar sempre disponível para conversar e esclarecer as dúvidas e principalmente pela confiança;
- à Léa pelas correções,
- ao Paulo pelas sugestões e correções;
- ao Pedro pelo auxilio com o PDB e pelas figuras de proteínas.
- ao Marcelo pelas discussões muito proveitosas;
- ao Jim pela ajuda inestimável com os desenhos e, principalmente, pelo apoio nas horas de crise;
- à direção do CAp-UFRJ pela compreensão, pelo apoio e pelas condições de trabalho.

Resumo

Neste trabalho é realizada uma análise estatística da dependência do tempo de enovelamento de polímeros com diversos parâmetros relacionados tanto com a seqüência primária, quanto com as estruturas de menor energia. O modelo utilizado representa os polímeros como cadeias de monômeros hidrofóbicos (H) e polares (P) descrevendo caminhos sem interseção na rede quadrada bidimensional. Para obter as estruturas enoveladas de menor energia foi desenvolvido um programa para gerar seqüências de tamanho 22 e 24 e explorar exaustivamente o espaço de configurações. Para simular o tempo de enovelamento foi desenvolvido um programa de busca baseado em algoritmo genético. Nove parâmetros são analisados, dos quais sete apresentam algum tipo de influência sobre o tempo de enovelamento, o principal deles sendo a freqüência de degenerescência do estado fundamental. É analisada a dependência de cada parâmetro com essa freqüência e a interdependência entre eles. A análise mostra que fatores relacionados com a ordem dos contatos entre os monômeros que interagem apresentam maior influência sobre o tempo de enovelamento do que aqueles relacionados com a hidrofobicidade das cadeias.

Abstract

In this work we perform a statistical analysis of the dependence of the folding time of polymers with respect to various parameters related to both their primary structures and their minimum energy states. The model used for the polymers is of short self-avoiding chains of hydrophobic (H) and polar (P) monomers configured on a two-dimensional square lattice. Minimum energy states were calculated for a number of random HP sequences of lengths 22 and 24 by using an exhaustive search. A program with a search mechanism based on a genetic algorithm was used to simulate the folding time. Nine parameters are analysed, of which seven present some level of influence on the folding time, the principal factor being the degree of degeneracy of the ground state. The dependence of each parameter on the degeneracy is studied, as well as their mutual dependence. The analysis shows that factors related to the primary sequence distance between interacting monomers has a greater influence over the folding time than those related to the hydrophobicity of the chains.

Índice

Lista de Figuras

Lista de Tabelas

Introdução

O estudo da estrutura de polímeros tem contribuído de modo decisivo para a compreensão dos mecanismos envolvidos nos processos biológicos, onde desempenham papel fundamental, relacionado, em geral, à sua estrutura espacial. Ácidos nucléicos, proteínas e polissacarídeos são exemplos de polímeros com atividade biológica. Em particular as proteínas são um exemplo de interação entre estrutura tridimensional e atividade biológica, repousando sua atividade na capacidade de reconhecer e interagir com moléculas altamente diversas, desempenhando papéis cruciais em, virtualmente, todos os processos biológicos.

Apesar da disponibilidade de um grande número de estruturas tridimensionais de proteínas, cujo conhecimento deriva de técnicas de cristalografia e ressonância magnética nuclear, pouco se tem avançado na elucidação dos mecanismos pelos quais a cadeia polipeptídica da proteína se enovela para adquirir sua configuração espacial biologicamente ativa. Mesmo o papel relativo de cada uma das forças que atuam no enovelamento protéico ainda não é bem esclarecido. Um dos problemas centrais do estudo de proteínas, conhecido como paradoxo de Levinthal, diz respeito ao tempo necessário para a proteína atingir essa conformação, também chamada de conformação nativa. Esse tempo cresceria exponencialmente com o comprimento da cadeia e seria da ordem de 10^{87}s., para uma pequena

proteína de 100 aminoácidos, caso o espaço de configurações fosse explorado aleatoriamente pela molécula.

Desse modo, a investigação dos mecanismos pelos quais as proteínas se enovelam tem merecido a atenção de um grande número de pesquisadores que vêm desenvolvendo diversos métodos de abordagem do problema, desde modelos simplificados, que buscam capturar apenas a essência do mecanismo de enovelamennto, até modelos mais realísticos, que procuram incluir o maior número possível de variáveis que possam afetar o processo. O modelo utilizado neste trabalho se enquadra no primeiro grupo, remetendo a princípios gerais do enovelamento, mais do que a aspectos atômicos.

A proposta deste trabalho é desenvolver um estudo das características de polímeros definidos num modelo simplificado na rede quadrada bidimensional, que reproduz alguns aspectos das proteínas reais, e relacioná-las com o tempo de enovelamento do polímero.

No capítulo 1 apresentamos um resumo da literatura referente aos principais aspectos envolvidos no enovelamento protéico, bem como resultados experimentais relacionados à questão. No capítulo 2 descrevemos três abordagens teóricas do problema de enovelamento e apresentamos o modelo de polímero utilizado neste trabalho. O capítulo 3 apresenta os métodos utilizados para a obtenção de uma amostra de polímeros construídos na rede quadrada bidimensional. Descrevemos o algoritmo, desenvolvido neste trabalho, para explorar o espaço de configurações dos polímeros definidos pelo modelo, e o algoritmo genético utilizado no programa que associa um tempo de enovelamento a cada polímero investigado. É também apresentado um rápido resumo sobre algoritmos genéticos. No capítulo 4 são definidas e analisadas, estatisticamente, características dos

polímeros que possam interferir no tempo de enovelamento da amostra. Finalmente, concluímos com uma análise dos principais resultados obtidos e perspectivas de possíveis desdobramentos deste trabalho.

Capítulo 1

Proteínas

Proteínas são polímeros lineares flexíveis, caracterizados por uma estrutura covalente, determinada por informação genética, e que apresentam a capacidade de adotar conformações tridimensionais relativamente definidas. Existem 20 tipos de aminoácidos que participam da estrutura das proteínas. Esses aminoácidos são identificados pela presença de 20 diferentes tipos de cadeias laterais, ligadas ao carbono α (Figura 1-1). As propriedades físico-químicas, tais como polaridade, acidez, aromaticidade, flexibilidade conformacional, tendência a estabelecer pontes de hidrogênio e reatividade química, variam consideravelmente entre os 20 tipos de aminoácidos, e são responsáveis, em grande parte, pelo amplo espectro de propriedades das proteínas. A seqüência de aminoácidos ligados entre si na cadeia polipeptídica é chamada de estrutura primária da proteína (Figura 1-2) e técnicas para sua determinação têm sido desenvolvidas desde 1953 [1].

A estrutura espacial biologicamente ativa da molécula, chamada de configuração nativa, é determinada pela seqüência de aminoácidos da proteína [2]. A conformação da

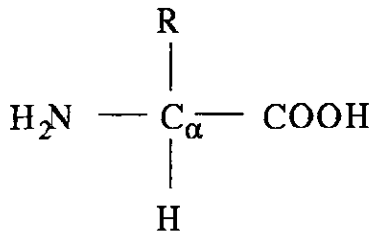


Figura 1-1: Fórmula estrutural geral dos aminoácidos. R é chamada de cadeia lateral e pode ser de 20 tipos diferentes.

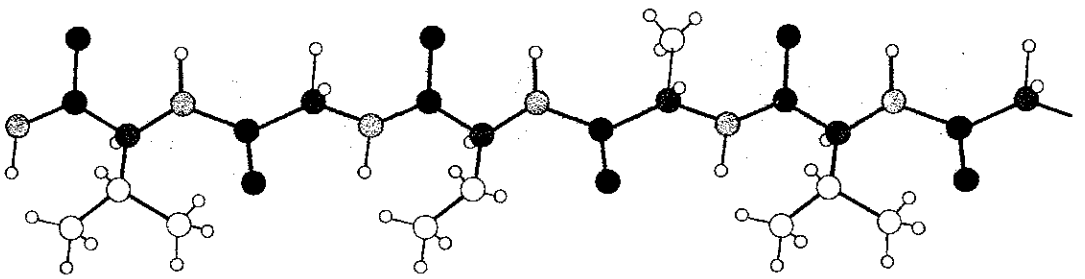


Figura 1-2: Estrutura primária da proteína. Uma cadeia polipeptídica, completamente estendida, mostrando a planaridade de cada grupo peptídico.

espinha dorsal polipeptídica pode ser especificada pelos ângulos torsionais em torno das ligações $C_\alpha - N(\phi)$ e $C_\alpha - C(\psi)$ (Figura 1-3). Entretanto, cada conformação não é especificada por uma única seqüência. Mudanças na seqüência de aminoácidos têm ocorrido durante divergência evolucionária, resultando em variantes com conformações muito similares, mesmo quando as similaridades remanescentes nas seqüências primárias são mínimas. Seqüências locais podem apresentar degenerescência estrutural: pentapeptídeos e hexapeptídeos idênticos podem adotar estruturas secundárias bastante diferentes [3]. Por outro lado, não há proteínas conhecidas, com estruturas primárias idênticas, que apresentam conformações nativas diferentes. A conformação, muito mais do que a seqüência de aminoácidos, tem sido mantida ao longo da evolução, presumivelmente por seleção natural [4].

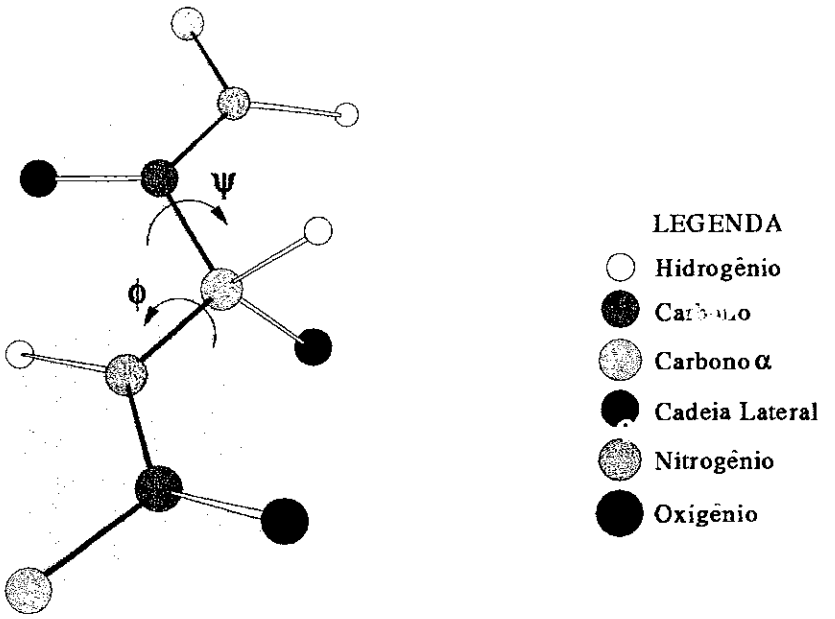


Figura 1-3: Ângulos torsionais ϕ e ψ da cadeia protéica. Os únicos movimentos razoavelmente livres são rotações em torno das ligações $C_\alpha - N(\phi)$ e $C_\alpha - C(\psi)$. Existem muitas restrições estéricas sobre os ângulos de torsão do esqueleto polipeptídico que limitam suas conformações.

Técnicas de difração de Raio-X e espectroscopia de Ressonância Magnética Nuclear, RMN, aliadas ao conhecimento da estrutura primária, têm permitido descrever as conformações detalhadas de um número crescente de proteínas. A determinação precisa das posições dos átomos nas estruturas tridimensionais, permitidas por essas técnicas, tem levado à constatação de que as diferentes estruturas das proteínas globulares, embora únicas e complexas, compartilham algumas propriedades.

A estrutura nativa de proteínas globulares é marcadamente compacta, com uma razão de compactação da região interna da ordem de 0,75. Esse valor é obtido pela razão entre o volume definido pelos raios de van der Waals dos átomos presentes numa dada região e o volume total da região. Em comparação, líquidos orgânicos, como gotas de óleo, têm razões entre 0,60 e 0,70 [1]. O interior da proteína é, portanto, eficientemente compacto,

há poucas cavidades, e quaisquer moléculas de água internas são, aparentemente, parte integral da estrutura, formando pontes de hidrogênio com grupos polares da proteína. No interior da estrutura raramente ocorrem grupos carregados ou polares não pareados em pontes de hidrogênio.

Em geral, o esqueleto polipeptídico forma estruturas secundárias denominadas hélices-α e folhas-β (Figuras 1-4 e 1-5). Numa proteína globular, essas estruturas secundárias, localizadas mais internamente, atravessam a configuração seguindo um curso moderadamente direto e, ao atingir um extremo, mudam abruptamente de direção, formando uma alça e seguindo para o outro lado, novamente num caminho mais ou menos direto (Figura 1-6). As alças (Figura 1-5), portanto, ocorrem na superfície da proteína. Cerca de 60% dos resíduos, na maior parte das proteínas globulares, estão localizados em hélices-α e folhas-β. A distribuição desses elementos, porém, varia amplamente em diferentes proteínas, algumas sendo formadas quase inteiramente por hélices-α e outras por folhas-β. Esse arranjo tridimensional das estruturas secundárias e das cadeias laterais é chamado de estrutura terciária da proteina.

Cadeias polipeptídicas com mais de 200 resíduos são, em geral, enoveladas em duas ou mais subunidades estruturais denominadas domínios, o que dá a essas proteínas a aparência bi ou multilobular. A maior parte dos domínios consiste de 100 a 200 resíduos de aminoácidos.. Os domínios parecem se enovelar individualmente e então agregar-se. A maior parte das proteínas é formada por mais de uma cadeia polipeptídica. As várias subunidades polipeptídicas se associam num arranjo geométrico específico, conhecido como estrutura quaternária. Embora mudanças conformacionais ocorram na agregação

Figura 1-4: Representações de uma hélice-α da proteína barnase, uma ribonuclease. À esquerda temos uma representação do esqueleto polipeptídico da estrutura e à direita uma representação em *cartoon*. Nas duas figuras a hélice-α aparece em rosa.

Figura 1-5: Representações de folhas-β da proteína barnase, uma ribonuclease. À esquerda temos uma representação do esqueleto da estrutura e à direita uma representação em *cartoon*. Nas duas figuras as folhas-β aparecem em amarelo. O trecho em azul representa uma alça.

dos diferentes domínios e na formação da estrutura quaternária, a questão primária do enovelamento de proteínas, de determinar como um domínio individual se enovela, permanece.

Sob certas condições as proteínas podem perder sua atividade biológica. Esse fenômeno é chamado de desnaturação e vem sendo estudado há mais de 60 anos. Em torno de 1925, o processo de desnaturação era visto como hidrólise da ligação peptídica ou desidratação da proteína. A noção de que a desnaturação era um processo de desenovelamento foi pri-

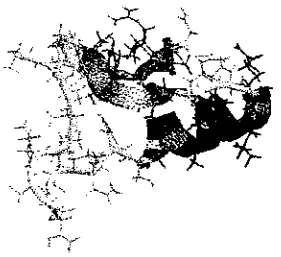


Figura 1-6: Configuração espacial da defensina-A de inseto, uma proteína do sistema imunológico. As folhas-β são mostradas em amarelo e as hélices-α em rosa.

meiro apresentada por Wu, em 1929 [5]. Ele propôs que proteínas nativas apresentariam modelos regulares de enovelamento da cadeia repetidos em uma rede tridimensional, um pouco semelhante a um cristal, mantidos juntos por ligações não covalentes. "Desnaturação é a quebra dessas ligações lábeis. Ao invés de ser compacta, a proteína torna-se uma estrutura difusa. A superfície é alterada e o interior da molécula é exposto" [5].

O estado desnaturado de uma proteína é difícil de se caracterizar. Ao menos sob algumas condições de desnaturação, uma proteína desenovelada apresenta as propriedades hidrodinâmicas esperadas para novelos aleatórios (conjunto de conformações totalmente desordenadas, com rápidas flutuações). Existem evidências de que algumas pequenas proteínas desnaturadas ou peptídeos não mantêm qualquer "estrutura residual" ou conformação estável não aleatória em solução aquosa; interações entre partes de cadeias polipeptídicas são muito fracas [4].

Muitos experimentos têm sido realizados no sentido de verificar a reversibilidade do processo de enovelamento e a estabilidade termodinâmica das estruturas nativas. A reversibilidade do processo de desnaturação e sua independência das condições iniciais foram

testadas pelos experimentos clássicos de Anfinsen [2]. No estudo, por meio da monitoração das ligações dissulfeto, sobre a renaturação de ribonuclease pancreática bovina (RNase A) *in vitro*, foi observada a reaquisição da estrutura nativa e da atividade da enzima, na ausência de outras proteínas, a partir de uma mistura de configurações que apresentavam as ligações dissulfeto distribuídas aleatoriamente.

Experimentos de calorimetria têm indicado que a desnaturação é reversível para muitas proteínas globulares pequenas com um único domínio [6] e, também, para algumas proteínas maiores, com múltiplos domínios [7]. Esses experimentos não implicam, porém, que a reversibilidade seja completamente geral para outras condições ou para outras proteínas.

Sabe-se que o enovelamento de algumas proteínas pode ser catalizado por outras proteínas assistentes, tais como "chaperones" [8], [9], [10], que dirigem o enovelamento e previnem a agregação de formas desenoveladas do polímero. Além disso, são conhecidos exemplos de proteínas que não têm a habilidade de se desnaturar e renaturar reversivelmente. Alguns membros da família das serpinas podem existir em duas diferentes conformações enoveladas, sendo que a forma biologicamente ativa parece ser a forma metaestável, de modo que existe a possibilidade de, em alguns casos, a conformação nativa ser determinada pelo caminho do enovelamento, mais do que pela estrutura mais estável [11].

Entretanto a existência dessas proteínas e de "chaperones" não deve excluir a hipótese da reversibilidade, e nem afetar a natureza do estado nativo ou as forças que dirigem o enovelamento [12].

Investigando a natureza e as magnitudes das forças dominantes no enovelamento de proteínas, Dill [12] analisou a contribuição de interações eletrostáticas, de pontes de hidrogênio e interações de Van der Waals, de interações locais decorrentes da existência de certas preferências conformacionais de di ou tripeptídeos, do efeito hidrofóbico e da entropia conformacional. Seu trabalho mostra a existência de fortes evidências que permitem considerar a interação hidrófobica e a entropia conformacional como contribuições dominantes ao enovelamento de proteínas, embora outras forças mais fracas também possam afetar a estabilidade.

Essa conclusão é apoiada por dados obtidos por Ponnuswamy & Gromiha [13] que utilizaram resultados experimentais de estrutura cristalina e de diferença de energia livre, ΔG, entre os estados enovelados e desnaturados de 14 proteínas conhecidas, para desenvolver um modelo utilizado na previsão da estabilidade conformacional de outras 24 proteínas. Por esse modelo, mostraram que a variação de energia livre devida a interações hidrofóbicas de enovelamento é diretamente proporcional ao número de resíduos. Além disso o aumento na energia livre devido a essa interação hidrofóbica de uma proteína particular é devido a uma composição mais alta de grupos não polares e cisteínas, e/ou devido à mais baixa composição de grupos polares. Os resultados relativos à contribuição proporcional de todas as forças analisadas apontaram para a seguinte descrição da estabilidade de proteínas [13]: quando a estrutura desnaturada começa a se enovelar progressivamente, a contribuição das interações hidrofóbicas aumenta, superando o efeito da entropia conformacional, e dirigindo a cadeia para o estado nativo; então as outras contribuições, especialmente pontes de hidrogênio e interações de van der Waals,

emergem para manter a estrutura enovelada.

Uma interpretação da interação hidrofóbica, baseada nas grandes variações de entropia obtidas a partir dos coeficientes de partição de pequenas moléculas apolares entre solventes apolares e água, supõe que interações entre moléculas apolares (H-H) aumentaria a entropia da água e esse aumento na entropia da água proveria a maior parte da energia para direcionar a associação de solutos não polares em água [12], [14]. Por outro lado a diferença observada de entropia entre moléculas apolares em solventes apolares e as mesmas moléculas em água pode ser interpretada, igualmente bem, como o resultado de uma perda de entropia da molécula apolar em água (H-W) em relação à entropia que ela possuia no solvente apolar (H-H); de uma diminuição na entropia da água com a introdução da molécula apolar em relação à entropia existente anteriormente, ou por uma mistura dos dois efeitos. Desse ponto de vista, um aumento de entropia resultaria, inteiramente, da substituição das ligações mais fortes com os dipolos da água (H-W) por ligações mais fracas entre as moléculas apolares (H-H) [15].

Embora a presença de uma grande organização interna nas proteínas globulares, compreendida pela combinação de estruturas irregulares (novelos e alças) e estruturas regulares tais como hélices-α e folhas-β, possa parecer difícil de conciliar com o quadro de forças dominantes no enovelamento, exaustivas simulações de todas as possíveis conformações de cadeias têm mostrado que uma arquitetura interna semelhante à de proteínas é conseqüência natural de vínculos estéricos em polímeros compactos [16]. Essas simulações indicam que qualquer polímero flexível, quando direcionado para uma compactação por qualquer força, apresentará arquitetura interna regular. Existem poucos modos eficientes

de preencher o espaço para obter uma configuração compacta e a maior parte deles envolve hélices-α e folhas-β. Assim, dados somente a liberdade conformacional e restrições estéricas em cadeias flexíveis, existe uma tendência à formação de hélices e folhas, mesmo na ausência de outras forças. Os resultados apresentados por Chan *et al.* [16] também foram obtidos para cadeias configuradas em duas e três dimensões, sobre diferentes tipos de redes, e por diferentes métodos [17] [18] [19].

Um aspecto da estrutura interna, que não é uma conseqüência de forças de compactação apenas, é a distribuição espacial das alças [20]. Alças ocorrem amplamente na superfície de proteínas e parecem ser uma conseqüência da natureza polar dos resíduos presentes [21]. Uma vez que os resíduos centrais das alças são geometricamente incapazes de formar pontes de hidrogênio no interior da estrutura, então sua tendência a estabelecer pontes de hidrogênio pode ser melhor satisfeita pela interação com a água na superfície da proteína [12].

Dados disponíveis indicam que proteínas *in vitro* podem enovelar-se à sua estrutura nativa num intervalo de tempo que varia de dezenas de segundos a minutos, na ausência de pontes de enxofre [22]. A principal questão, portanto, diz respeito a como uma cadeia polipeptídica consegue atingir rapidamente sua forma nativa, apesar do grande número de conformações acessíveis a ela. Um cálculo realizado primeiramente por Cyrus Levinthal [23], demonstra que o processo de enovelamento não deve ocorrer através da exploração aleatória de todas as conformações disponíveis. Assumindo que os $2n$ ângulos torsionais, ϕ e ψ (Figura 1-3), de uma cadeia com n resíduos, tenham, cada um, apenas três conformações estáveis, isto fornece $3^{2n} \approx 10^n$ possíveis conformações para a proteína.

Esta é uma estimativa grosseira, uma vez que as cadeias laterais são ignoradas. Se uma proteína puder explorar novas conformações à velocidade com que ligações simples podem se reorientar, então ela poderá encontrar $\approx 10^{13}$ conformações por segundo, o que é, sem dúvida, uma superestimativa. Podemos então calcular o tempo, t, em segundos, necessário para uma proteína explorar todas as conformações disponíveis a ela:

$$t = 10^{n-13}s.$$

Para uma pequena proteína de $n = 100$ resíduos, $t = 10^{87}$s, o que é um tempo imensamente maior do que a idade aparente do universo (6×10^{17}s). Mesmo a menor proteína levaria um tempo absurdamente grande para explorar todas as suas possíveis conformações. Além disso, esse tempo cresce exponencialmente com o tamanho da cadeia. Entretanto, como já foi dito, muitas proteínas se enovelam a suas conformações nativas em menos do que alguns segundos.

Muitos trabalhos têm sido desenvolvidos no sentido de determinar como o paradoxo de Levinthal é resolvido. A transição entre o estado nativo e os estados desnaturados tem sido estudada exaustivamente [24] [25] e os resultados experimentais sugerem que a renaturação deve ser iniciada por

(1) colapso de regiões hidrofóbicas para o interior da molécula,

(2) formação de estruturas secundárias estáveis, que fornecem um referencial para o enovelamento subseqüente e

(3) formação de interações covalentes, tais como ligações dissulfeto, que estabilizam o polipeptídeo em conformações particulares.

Têm sido obtidas evidências que suportam cada um desses mecanismos e é provável que todos os três devem operar em conjunto durante os estágios iniciais da renaturação [26]. Enovelamento subseqüente parece ocorrer através de um número limitado de caminhos, envolvendo intermediários distintos (glóbulos fluidos ou intermediários compactos) que têm significativa estrutura secundária e uma forma compacta, mas sem uma bem definida estrutura terciária, e com uma exposição maior de uma superfície hidrofóbica do que as moléculas completamente enoveladas. Esses intermediários parecem estar em rápido equilíbrio com o estado completamente desnaturado e são apenas lentamente convertidos ao estado nativo. Assim, a fase em que o processo de renaturação se torna mais lento, freqüentemente ocorre em um estágio muito tardio, imediatamente antes de a proteína adotar sua conformação nativa final [26]

Por outro lado, modelos teóricos têm sido propostos considerando que o enovelamento se desenvolve em dois estágios, primeiro ocorre um aumento da compactação da cadeia e em seguida uma reconfiguração dos resíduos polares na superfície e não polares no centro. Desse modo, o colapso inicial reduziria drasticamente o espaço conformacional, permitindo às proteínas atingir seu estado de energia mínima [12].

Simulações por Monte Carlo, realizadas por Shakhnovich *et al.* [27], [28], em um modelo simplificado, consistindo de uma cadeia com 27 monômeros ocupando os vértices de uma rede cúbica, forneceram resultados consistentes com essa hipótese. O resultado mais importante encontrado pelos autores foi que seqüências com rápido enovelamento

Capítulo 2

Abordagens Teóricas

A complexidade e individualidade das estruturas tridimensionais de proteínas conhecidas têm fornecido poucos indícios para a elucidação dos mecanismos pelos quais essas configurações são atingidas. Uma grande quantidade de estudos têm sido dedicados ao problema mas o mecanismo de enovelamento de proteínas permanece desconhecido ainda [4], [19], [30]. Para tentar entender os mecanismos microscópicos envolvidos no enovelamento e na desnaturação, e prever a configuração espacial de proteínas, diferentes abordagens têm sido adotadas, tanto experimentais quanto teóricas. Do ponto de vista teórico, a principal técnica escolhida tem sido simulações computacionais de modelos de proteínas. As abordagens computacionais utilizadas podem ser divididas em três categorias principais – dinâmica molecular e estudos de minimização de energia em modelos realísticos, modelagem comparativa ou por homologia e estudos Monte Carlo e suas variantes – simulação por recozimento e algoritmos genéticos – de modelos altamente simplificados.

apresentam um pronunciado mínimo na superfície de energia, isto é, a estrutura mais estável encontra-se num nível de energia muito mais baixo do que as estruturas vizinhas, enquanto seqüências com lento enovelamento têm espectros de energia aproximadamente contínuos, com a estrutura mais estável muito próxima, em energia, das suas vizinhas. Além disso, observaram que o enovelamento começa por um rápido colapso, a partir de um estado formado por novelos aleatórios, para um glóbulo semi-compacto aleatório. Em seguida, prossegue em uma busca lenta, através de estados semi-compactos, até encontrar um estado de transição a partir do qual a cadeia se enovela rapidamente ao estado nativo. Assim, embora o modelo tenha um espaço de configurações que exibe o paradoxo de Levinthal (o número total de conformações para uma cadeia, nesse modelo, é $6^{25} \approx 10^{20}$), certas seqüências podem encontrar o mínimo global explorando um número reduzido de conformações no espaço de glóbulos semi-compactos ($\approx 10^{10}$) e encontrando muitos estados de transição ($\approx 10^{3}$). Dessa forma, eles concluíram que a habilidade de enovelar-se rapidamente parece estar presente nas mesmas seqüências que podem formar estruturas termodinamicamente estáveis. Assim, além de selecionar seqüências em termos da necessidade de um mínimo global único [29], a evolução deve também ter selecionado seqüências em termos de necessidades cinéticas de enovelamento [27].

Como pode ser visto, o problema de como as proteínas se enovelam é extremamente complexo e os estudos até agora dedicados a sua compreensão têm apenas permitido vislumbrar esboços de soluções. No próximo capítulo descrevemos sucintamente alguns métodos teóricos utilizados na abordagem do problema.

2.1 Dinâmica Molecular e minimização de energia

Mecânica e dinâmica molecular são extensivamente usadas na predição da estrutura de proteínas, complementando abordagens experimentais e análises de banco de dados. Esses métodos são capazes de prever conformações ao nível atômico. Este alto grau de resolução, entretanto, tem um grande custo computacional, colocando limites no tamanho do sistema, na extensão do espaço conformacional explorada e no tamanho das simulações. Simulações de dinâmica molecular têm sido aplicadas a problemas como a predição da estrutura tridimensional de polipeptídeos pequenos a partir de sua seqüência de aminoácidos, a análise da estabilidade conformacional de peptídeos [31] e a predição da conformação de regiões locais de proteínas, utilizando abordagens que reduzem o custo computacional [32]. Os recursos computacionais atuais não permitem obter a estrutura enovelada de uma proteína *ab initio*, a partir unicamente do conhecimento de sua seqüência primária, mas o enovelamento pode ser obtido com a utilização de informações adicionais, impondo-se algumas restrições.

Dois métodos formam a base da mecânica molecular: minimização de energia e dinâmica molecular. Ambos utilizam uma função energia potencial para descrever as interações atômicas, que são modeladas usando relativamente simples formas matemáticas de interações ligadas e não-ligadas.

Métodos de simulação de dinâmica molecular são baseados na solução numérica das equações de movimento de Newton para todos os átomos do sistema [32]. A derivada da função energia potencial com respeito à posição atômica é usada para calcular as forças

sobre cada átomo do sistema. A aceleração de cada átomo é então determinada, a partir dessas forças, usando as equações de movimento de Newton. Dadas uma posição e uma velocidade atômicas iniciais, a nova posição é obtida integrando-se a aceleração. Desse modo, as simulações de dinâmica molecular descrevem a evolução temporal de um sistema atômico, isto é produzem uma trajetória no espaço de fase.

Os métodos de minimização de energia ajustam as coordenadas do sistema, a partir de uma estrutura tridimensional inicial, de modo a obter configurações de energia mínima locais ou globais [33]. Dinâmica molecular pode ser usada em conjunto com cálculos de energia livre para fornecer a superfície de energia livre conformacional.

É também possível introduzir o controle de temperatura em dinâmica molecular. Esta possibilidade é responsável por muitas conexões úteis entre as trajetórias moleculares e quantidades termodinâmicas estatísticas de interesse nos cálculos de energia livre.

2.2 Modelagem Comparativa

Uma previsão da estrutura tridimensional de uma proteína, a partir da sua seqüência de aminoácidos, pode ser obtida quando a estrutura de um ou mais homólogos é conhecida. Informações estruturais podem ser extrapoladas para a nova seqüência e um modelo tridimensional pode ser obtido. Esta abordagem é conhecida como modelagem comparativa ou modelagem por homologia.

O primeiro passo na modelagem comparativa é identificar as estruturas tridimensionais conhecidas, que formarão uma base para a estrutura desconhecida. Utilizando-se

diversos métodos de comparação da proteína a ser modelada com as seqüências e estruturas armazenadas nos bancos de dados, pode-se determinar o grau de identidade da seqüência com os homólogos de estrutura conhecida. Nesse sentido foram desenvolvidos diferentes esquemas para pontuar a equivalência de cada um dos 210 possíveis pares de aminoácidos [34].

Uma vez escolhido o esquema de pontuação e identificados os homólogos de estrutura tridimensional conhecida, o próximo passo é gerar um alinhamento da seqüência a ser modelada com aquela de estrutura conhecida. O alinhamento é tanto mais difícil quanto mais baixo for o grau de identidade de seqüência. O alinhamento torna-se extremamente difícil quando a identidade da seqüência é menor do que 25%. Uma das técnicas de modelagem comparativa [34] utiliza a extrapolação de fragmentos equivalentes dos homólogos identificados para a seqüência de estrutura desconhecida, seguida da combinação desses fragmentos para obter a nova configuração.

O procedimento consiste em selecionar segmentos rígidos, a partir dos homólogos, e montá-los numa base tridimensional, obtida a partir da superposição de regiões estruturalmente conservadas em estruturas homólogas conhecidas, em geral hélices e folhas. A contribuição de cada homólogo é ponderada pelo grau de identidade das seqüências. As regiões variáveis, representadas por alças e laços, são as mais difíceis de serem modeladas, uma vez que variam muito em comprimento, seqüência e conformação, mesmo entre proteínas de uma dada família [35]. Inconsistências na geometria do modelo podem estar presentes, particularmente nas regiões de ancoragem, onde um laço é mesclado com a região interna, mais estruturalemnte definida da proteína. Para retificar tais incon-

sistências e corrigir contatos próximos pode-se utilizar procedimentos de minimização de energia que não deverão alterar significativamente a estrutura.

2.3 Simulações em rede de modelos simplificados

Representações de proteínas na rede têm sido utilizadas há bastante tempo. Gō *et al.* [36] têm empregado uma série de modelos simplificados em redes bi e tridimensionais, nos quais interações são permitidas apenas entre resíduos em contato no estado nativo, para investigar a contribuição de interações de curto e longo alcance para as transições conformacionais. O modelo foi bem sucedido em obter a estrutura enovelada do inibidor da tripsina pancreática, a partir do estado desnaturado, mas o mesmo não ocorreu para outra proteína estudada, a lisozima. Mais recentemente, Chan e Dill [16] têm explorado exaustivamente o espaço de seqüências e o espaço conformacional de polímeros compactos na rede bidimensional e encontrado que conformações compactas são dominadas por estruturas secundárias. Além disso, Krigbaum e Lin [37] têm utilizado um modelo hexagonal de rede para investigar a eficiência relativa do enovelamento sob potenciais de interação local *versus* centrossimétricos. Skolnick e Kolinski [38] têm desenvolvido uma série de modelos de proteínas globulares em rede de diamante. Shakhnovich [39] tem utilizado modelos em rede quadrada tridimensional para investigar a acessibilidade cinética da estrutura de menor energia para seqüências termodinamicamente selecionadas de modo a terem um pronunciado mínimo de energia. O'Toole & Panagiotopoulos [40] utilizaram outro modelo tridimensional em rede quadrada para modelar a termodinâmica da transição entre

formas nativas e desnaturadas de proteínas. Analisando seqüências arbitrárias, de comprimento maior do que 48 resíduos, eles determinaram a curva de desnaturação térmica e observaram que esta varia muito com a seqüência, aproximando-se, em alguns casos especiais, da forma observada para proteínas reais. Essas seqüências especiais parecem ser simples arranjos periódicos de resíduos.

Nesses modelos em rede são utilizadas seqüências de duas letras (hidrofóbico-hidrofílico) ou seqüências de 20 letras para representar os aminoácidos, bem como uma variedade de potenciais para descrever as interações entre os resíduos.

Um modelo bastante simples, baseado apenas nas forças dominantes envolvidas, que captura a essência dos componentes importantes do enovelamento de proteínas e ainda permite computar as energias de todas as configurações de uma cadeia, bem como encontrar as configurações com a menor energia possível, foi desenvolvido por Lau & Dill em 1989 [41]. Nele, as proteínas são modeladas como cadeias curtas, que sofrem enovelamento descrevendo um caminho sem intersecção numa rede quadrada bidimensional. Os aminoácidos podem ser de dois tipos, H (hidrofóbicos) ou P (outros). A fração de aminoácidos do tipo H é dada por Φ e do tipo P por $1 - \Phi$. Para cada contato HH, entre dois monômeros hidrofóbicos, não conectados, ocupando pontos vizinhos na rede, não diagonais, é atribuída uma energia de contato igual a ϵ, ($\epsilon < 0$). Todas as outras interações, entre todos os outros possíveis tipos de vizinhos, têm energia igual a 0.

Em vários aspectos, este modelo bidimensional mimetiza as propriedades de proteínas reais [42] [43] [44]. As conformações de energia mínima têm a propriedade de ser maximamente compactas, apresentar um caroço predominantemente composto por resíduos

não-polares e possuir uma considerável quantidade de equivalentes bidimensionais de estruturas secundárias: hélices-α e folhas-β. As distribuições de hélices e folhas paralelas e anti-paralelas nas cadeias curtas, representadas na rede bidimensional, são similares àquelas observadas em proteínas conhecidas [41]. Além disso o modelo H-P para cadeias curtas na rede bidimensional é consistente com experimentos sobre mutabilidade de proteínas nos seguintes aspectos [41] [43]:

(1) sítios na superfície são altamente mutáveis, isto é, a maior parte das mutações simples na superfície não levam a mudanças na estrutura,

(2) sítios H localizados no caroço hidrofóbico são sensíveis à mutação, isto é, quando um resíduo H do caroço é substituído por um resíduo P, ocorre uma perda de estabilidade,

(3) a maior parte das mutações entre seqüências que codificam uma única estrutura de energia mínima são neutras no sentido de que elas não mudam a estrutura, implicando que proteínas devem ser altamente plásticas à mutação, e

(4) existem muitas seqüências convergentes que podem se enovelar a uma dada estrutura nativa.

O modelo permite a exploração exaustiva do espaço conformacional de cadeias pequenas (até 24 monômeros) para identificar as conformações de energia mínima. O fato de o modelo ser bidimensional e não tridimensional oferece certas vantagens [44]. Para proteínas reais, com 100 resíduos, em três dimensões, argumentos geométricos simples

mostram que apenas cerca de 20% a 30% dos resíduos podem estar no caroço; 70% a 80% devem estar na superfície de uma conformação compacta. Esta razão é recuperada na rede bidimensional para polímeros a partir de 16 monômeros de comprimento até 24. Por outro lado, em três dimensões, o menor modelo significativo é um cubo com 27 monômeros. Tal modelo em 3D apresenta duas limitações:

(1) existe apenas um monômero representando o caroço hidrofóbico, e

(2) mesmo cadeias com 27 monômeros são muito longas para simulações exaustivas, de modo que o espaço de configurações não pode ser explorado completamente [44].

Esta é uma abordagem voltada para questões de princípios gerais mais do que para detalhes atômicos, pois são considerados apenas a natureza compacta das conformações enoveladas.

Capítulo 3

Métodos

3.1 O modelo HP

Neste trabalho foi utilizado o modelo H-P, desenvolvido por Lau & Dill [41], descrito no capítulo 2. Para o cálculo da energia de uma configuração foram atribuidos os valores -1, para cada par de monômeros HH não ligados, ocupando pontos vizinhos na rede, e 0, para qualquer outro contato (PP ou HP). Uma vez que, durante as simulações, as interações hidrofóbicas são privilegiadas pela função energia, as configurações de energia mínima das cadeias são representadas por estruturas compactas, com um caroço hidrofóbico, enquanto os monômeros do tipo (P) são "forçados" para a superfície. A Figura 3-1 mostra essas configurações para uma dada seqüência de 24 monômeros.

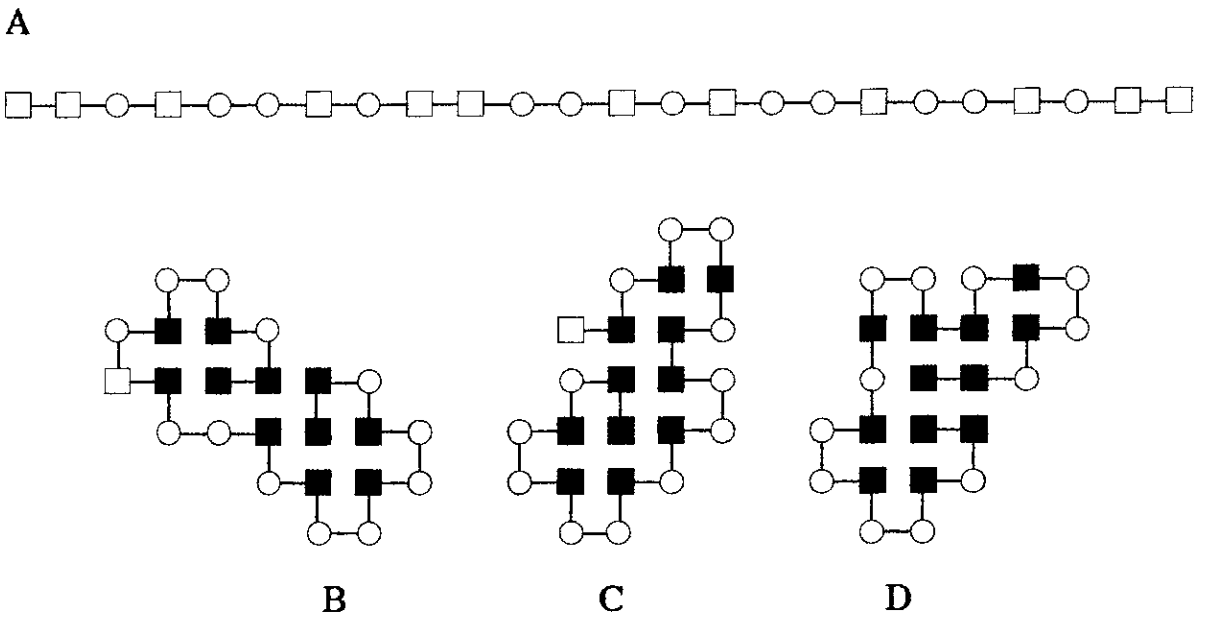


Figura 3-1: Exemplo de uma seqüência de comprimento 24 (A) e suas estruturas de menor energia (B, C e D). Círculos representam resíduos P e quadrados representam resíduos H. Os resíduos H em contato estão representados em preto.

3.2 Obtenção da amostra

Uma vez escolhido o modelo, o passo seguinte foi obter uma amostra significativa de seqüências primárias, com seus respectivos valores de energia e configurações de energia mínima, que pudesse ser analisada estatisticamente. No sentido de manter, na essência, um paralelo entre os elementos dessa população e proteínas reais foram definidas algumas características que nortearam a geração da amostra.

O maior número possível de configurações de uma dada seqüência, que podem ser construídas na rede quadrada, sem considerações de volume excluído, i.e., permitindo a superposição de monômeros, é dado por $\frac{3^{N-2}+1}{2}$, onde N é o número de elementos da cadeia. Para $N = 24$ esse número é da ordem de 2^{32}, que corresponde ao tamanho do maior inteiro que pode ser armazenado e manipulado eficientemente pelos computadores

utilizados neste trabalho. Foram geradas e analisadas, inicialmente, seqüências com 18, 20, 22 e 24 monômeros.

Dentre as seqüências geradas, foram selecionadas aquelas cujo espaço conformacional apresentasse um número reduzido de configurações com energia mínima, uma vez que polipeptídeos reais se enovelam a uma configuração espacial definida. Ao ser analisada a fração do espaço de configurações correspondente às estruturas de menor energia, para os quatro tamanhos de seqüências, optou-se por descartar as de tamanho 18 e 20 pois, para elas, a fração correspondente à menor degenerescência do estado fundamental (10^{-6}) era muito superior às frações relativas aos tamanhos 22 (10^{-9}) e 24 (10^{-10}). Dentre estas últimas foram, portanto, escolhidas aquelas com freqüência de degenerescência do estado de energia mínimo menor ou igual a 10^{-7}, eliminando as seqüências de tamanho 18 e 20. Para as seqüências de tamanho 24 essa freqüência corresponde a aproximadamente 300 estruturas diferentes e para as seqüências de tamanho 22, 100 estruturas.

Para obtenção e caracterização da amostra foi elaborado um programa, em linguagem PASCAL, que gera seqüências aleatórias de zeros (resíduos P) e uns (resíduos H), e para cada seqüência, constrói, uma a uma, todas as configurações permitidas na rede, gerando as coordenadas de cada resíduo e calculando a energia de cada estrutura. O programa armazena o número de configurações para cada nível de energia e as coordenadas das configurações de energia mínima. Se a freqüência de configurações com energia mínima ultrapassa o valor estipulado de 10^{-7}, a seqüência é desprezada.

O procedimento seguido para a obtenção e caracterização da amostra pode ser resumido nos seguintes passos:

(1) Geração de seqüências aleatórias de zeros (P) e uns (H), representando a seqüência primária H-P do polímero.

(2) Enumeração completa do espaço de configurações de cada seqüência para avaliar a energia de todas as configurações e determinar o número de estruturas em cada nível de energia.

(3) Seleção das seqüências com freqüência de degenerescência do estado de menor energia menor ou igual a 10^{-7}.

Com este procedimento foi obtida uma amostra com 71 cadeias de tamanho 22 e 122 de tamanho 24, cada uma delas caracterizada pela seqüência primária de monômeros H e P, pela distribuição de energia no espaço de configurações e pela conformação de suas estruturas de energia mínima. Toda a amostra foi gerada utilizando-se microcomputadores PC DX4-100. Para uma seqüência de comprimento 22 o tempo gasto para explorar integralmente o espaço de configurações foi de aproximadamente 1 hora, enquanto seqüências de tamanho 24 utilizavam cerca de 8 horas no mesmo equipamento. Vários procedimentos foram testados até conseguirmos um tempo computacional razoável. Nesses procedimentos o passo mais dispendioso é o cálculo da energia de cada configuração.

O passo seguinte consistiu em elaborar um procedimento que permitisse associar a cada seqüência obtida um tempo de busca computacional pela estrutura de menor energia conhecida. Com isso esperava-se identificar possíveis características das seqüências, independentes da degenerescência do estado de energia mínima, que pudessem ser identificadas como apresentando influência significativa no tempo de enovelamento. Dentre

as várias possibilidades de métodos de busca disponíveis, optou-se por elaborar um programa em PASCAL utilizando algorítmo genético [45], método de busca cuja eficiência já havia sido testada em trabalhos anteriores [46].

3.3 Algoritmos genéticos

Algoritmos genéticos são procedimentos estocásticos de busca que trabalham com populações de pontos do espaço de configurações e não com pontos isolados. Essencialmente, um algoritmo genético é composto de três operadores:

1. Reprodução
2. Permutação
3. Mutação

aplicados a uma população inicial de indivíduos, gerada aleatoriamente, sobre a qual se processa a busca. Para permitir o tratamento pelos operadores, em geral, cada indivíduo costuma ser representado por uma seqüência de caracteres de comprimento finito definida sobre um alfabeto finito. A representação mais simples e mais geral consiste em utilizar um alfabeto binário para construir as seqüências, embora alfabetos de cardinalidade mais alta possam ser usados [47].

O procedimento tem início com a reprodução, um processo pelo qual as seqüências individuais, avaliadas por uma função custo, são copiadas para a geração seguinte com probabilidade proporcional a essa função, que representa o valor adaptativo do indivíduo

na população. Copiar seqüências com maior valor adaptativo significa que seqüências com valor mais alto terão uma maior probabilidade de contribuir com um ou mais descendentes para a próxima geração. Após a reprodução, pares de indivíduos são selecionados aleatoriamente e cada par sofre permutação. Para isso, uma posição é sorteada ao longo da seqüência e as porções iniciais e terminais são trocadas, criando-se duas novas seqüências. A mutação, num algoritmo genético simples, é a alteração aleatória do valor de uma posição da seqüência. Corresponde a percorrer um caminho aleatório através do espaço das soluções, e é utilizada para prevenir a perda prematura de informações importantes [48]. Após uma rodada de reprodução, permutação e mutação, uma nova população de seqüências substitui a primeira. Ao transformar um conjunto prévio de bons indivíduos em um novo conjunto, os operadores geram uma nova população com indivíduos cujo valor adaptativo médio é, em geral, maior do que o da geração anterior. Após a repetição desse ciclo por muitas gerações, o valor adaptativo médio da população geralmente cresce, e os indivíduos representam soluções melhores do problema definido pela função custo.

O procedimento descrito acima pode ser aplicado de muitas maneiras diferentes para resolver um amplo espectro de problemas. Ao se desenvolver um algoritmo genético para resolver um problema específico existem duas principais decisões envolvidas:

(1) especificar o mapeamento entre as seqüências de caracteres e os candidatos a soluções (comumente definido como o problema de representação) e

(2) definir uma medida concreta do valor adaptativo (função custo).

A representação particular e a função custo selecionadas determinarão o sucesso do algoritmo genético na resolução do problema proposto.

3.4 Algoritmo para busca de estados de energia mínima no modelo HP

Unger & Moult [46] desenvolveram um procedimento de busca por algoritmo genético, apropriado para utilização em simulações de enovelamento de seqüências representando proteínas em uma rede bidimensional. O procedimento foi utilizado para encontrar a conformação de menor energia de cada uma de 8 seqüências, com comprimentos variando entre 20 e 64. Quando comparado com métodos do tipo Monte Carlo o algoritmo genético encontrou uma solução rapidamente em 7 casos contra os 3 casos (seqüências mais curtas) resolvidos pelo método Monte Carlo, isto é, em todos os casos o número de passos computacionais equivalentes, gastos para se chegar à estrutura de menor energia, foi menor para o procedimento baseado em algoritmo genético do que para aqueles baseados em Monte Carlo. O algoritmo desenvolvido neste trabalho é baseado na descrição apresentada em [46].

Para a obtenção de uma grandeza que representasse o tempo de enovelamento de cada uma das seqüências geradas aleatoriamente, foi desenvolvido um programa em linguagem PASCAL, utilizando o método de algoritmos genéticos. Neste programa, a população de soluções é formada por configurações de uma mesma seqüência primária, representadas pelas coordenadas de cada um de seus resíduos, que são tratadas diretamente pelos ope-

radores genéticos. A população é formada por N configurações de uma dada seqüência. A cada geração, cada configuração é submetida a um número L de passos de mutação, onde L foi escolhido igual ao comprimento da cadeia. Cada passo consiste em sortear uma posição na cadeia e submetê-la a um giro de 90°, 0° ou −90° nesse ponto (Figura 3-2). A cada mutação, se o giro produz uma configuração válida, i. e., sem superposição de monômeros, a energia E_2 da nova estrutura é calculada e comparada à energia E_1 da estrutura original. Se $E_2 \leq E_1$, então a nova conformação é aceita, senão a nova configuração é aceita de acordo com o critério Metropolis, da forma:

$$Rnd < \exp\left[\frac{E_1 - E_2}{k_B T(t)}\right],$$

onde Rnd é um número aleatório entre 0 e 1, $k_B = 1$ por simplicidade e $T(t)$ é gradualmente diminuida durante a simulação. Se a mudança não for aceita, a configuração inicial é mantida e uma nova mutação é tentada até se completar o número L de passos. A mutação, neste caso, corresponde ao método de recozimento simulado.

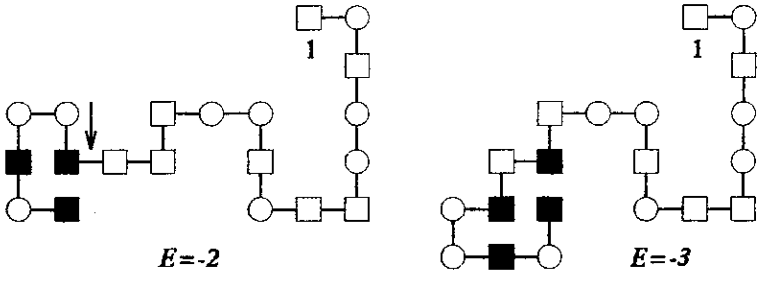


Figura 3-2: **Mutação.** Um monômero é escolhido aleatoriamente e a porção terminal da cadeia é girada em torno desse monômero. No exemplo, o resíduo de número 14 foi selecionado e uma rotação de 90° do segmento 15 a 20 transformou a estrutura A na estrutura B. O movimento é sempre aceito se leva a uma estrutura de energia mais baixa, ou aceito não deterministicamente de acordo com o aumento de energia. Os resíduos H em contato estão representados em preto.

Ao final deste estágio a operação de permutação é realizada. Cada estrutura, S_i, é selecionada para permutação de acordo com uma probabilidade $p(S_i)$, igual à sua contribuição fracional à energia total da população:

$$p(S_i) = \frac{E_i}{\sum_{j=1}^{N} E_j}.$$

Para um par de estruturas selecionadas é escolhido um ponto aleatório ao longo da seqüência e a porção inicial da primeira estrutura é conectada à porção terminal da segunda. Como existem três maneiras de juntar as partes, em ângulos de $-90°$, $0°$ e $90°$, essas possibilidades são testadas em ordem aleatória até encontrar uma que seja válida, i.e na qual não haja superposição de monômeros (Figura 3-3). Se nenhuma das posições fornecer uma estrutura válida, o par é abandonado e outro é sorteado em seu lugar. Uma vez encontrada uma configuração consistente, S_k, sua energia E_k é calculada e comparada à média das energias das estruturas-mãe, $E_{ij} = \frac{1}{2}(E_i + E_j)$. A estrutura é aceita se $E_k \leq E_{ij}$ ou se o aumento de energia satisfizer a relação:

$$Rnd < \exp\left[\frac{E_{ij} - E_k}{k_B T(t)}\right].$$

Novamente utilizamos aqui o método de recozimento simulado. A operação de permutação é repetida até que $N - 1$ estruturas híbridas tenham sido criadas para formar a nova geração. A configuração de menor energia de cada geração é copiada para a geração seguinte, completanto os N indivíduos.

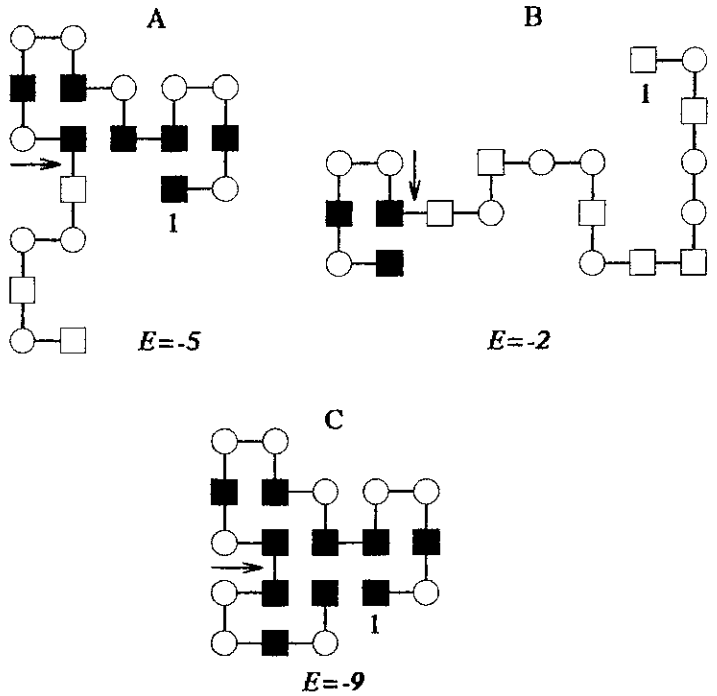


Figura 3-3: **Permutação.** Pares de estruturas são selecionados com base em sua energia. Neste exemplo a posição de corte foi aleatoriamente definida após o resíduo 14. Os 14 primeiros resíduos da estrutura A foram ligados aos 6 últimos da estrutura B, após uma rotação de 90° no ponto de junção, dando origem a estrutura em C. A energia da estrutura híbrida, neste caso, é menor que as energias das estruturas-mãe. O híbrido é sempre aceito se sua energia for menor do que a média das energias dos pais, ou aceito não deterministicamente, de acordo com o aumento de energia. Em cada estrutura, estão representados em preto os resíduos H vizinhos na rede

Diferentemente do algoritmo genético clássico, o algoritmo aqui desenvolvido não utiliza uma probabilidade constante de mutação e permutação. Toda mutação e permutação que resultam na diminuição de energia são aceitos, e, para modificações que elevam a energia utilizamos o critério Metropolis [49] de aceitação. Neste caso, a probabilidade de mutação e permutação é inversamente proporcional ao aumento de energia, segundo uma distribuição de Boltzman. A probabilidade é tanto maior quanto menor for o aumento de energia gerado pelo operador, em uma dada temperatura. Além disso, como já foi dito, a temperatura é diminuída ao longo da simulação, seguindo um esquema de recozimento.

Para a permutação, a temperatura varia no intervalo [0,3 ... 0,1], segundo a expressão

$$T(t) = T(1)\,0{,}99^{t-1}. \tag{3.1}$$

Nesse intervalo as variações de energia para cima maiores do que 2 ($\Delta E > 2$) não são aceitas. Uma variação igual a 1 tem probabilidade 0,1 de ser aceita no início da simulação. Isto significa que durante a maior parte da simulação apenas serão aceitas, na fase de permutação, variações que reduzam a energia da configuração.

Para a mutação a temperatura varia no intervalo [2 ... 0,6] segundo a expressão

$$T(t) = T(1)\,0{,}97^{t-1}. \tag{3.2}$$

Nesse intervalo, a probabilidade de aceitação de aumentos de energia maiores do que 8 é menor do que 0,03. Aumentos de energia iguais a 3 têm probabilidade aproximada de 0,2 de serem aceitos no início da simulação. Essa probabilidade cai para cerca de 0,06 na metade da simulação, quando a temperatura chega a 1,1. Nesse ponto (100 gerações) variações maiores do que 5 têm probabilidade menor do que 0,02 de serem aceitas e variações iguais a 2 são aceitas com probabilidade aproximada de 0,15. Ou seja, a partir da metade da simulação, praticamente apenas serão aceitas variações que diminuam a energia ou que a aumentem de uma ou duas unidades.

Neste trabalho, a população inicial era sempre formada por 100 estruturas estendidas de uma mesma seqüência, e o procedimento rodado por, no máximo, 200 gerações, ou até que uma configuração de energia mínima fosse encontrada. O estágio de permutação

começava com $T(t) = 0{,}3$, reduzida a cada 5 gerações segundo a expressão (3.1). Para o estágio de mutação o esquema de resfriamento começava com $T(t) = 2$ e a cada 5 gerações a temperatura era modificada segundo a expressão (3.2).

3.5 Simulação do tempo de enovelamento

Com o procedimento descrito, passamos, então, a verificar o comportamento de nossa amostra quanto ao número de passos do algoritmo necessários para atingir o estado de energia mínima.

O programa foi testado inicialmente, com as seqüências estudadas por Unger & Moult [46], com resultados semelhantes, e em seguida aplicado às seqüências da amostra. Para cada seqüência a simulação foi repetida 6 vezes, com números aleatórios diferentes, e registrado o número de gerações necessárias para se chegar a uma configuração de energia mínima, em cada simulação. Quando nenhuma estrutura desse tipo era encontrada em até 200 gerações registrava-se o valor 300 para a corrida. Esse valor foi tomado para representar as seqüências com tempo muito alto de enovelamento e foi verificado que os resultados aqui apresentados não eram sensíveis a mudanças a partir desse valor. O número médio de gerações encontrado foi então utilizado como estimativa do tempo de enovelamento, G, para a seqüência.

As Figuras 3-4 e 3-5 mostram a evolução da energia mínima e média da população durante uma simulação do algoritmo genético para uma seqüência que se enovela e para outra que não se enovela dentro do número limite de passos determinado, respectivamente.

As tabelas no Apêndice mostram as seqüências que compõem a amostra analisada neste trabalho, os valores de energia mínima e a degenerescência absoluta obtidos para cada seqüência.

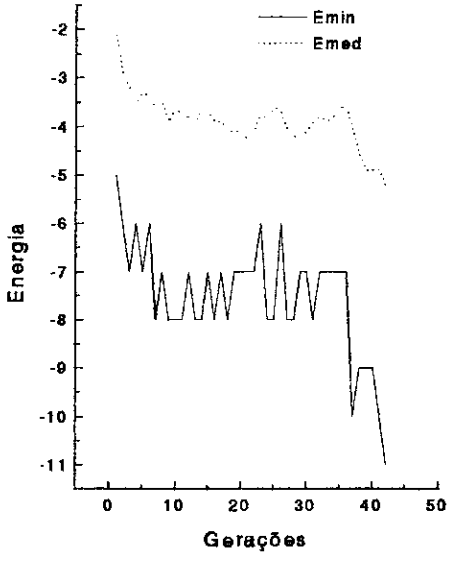


Figura 3-4: A figura mostra a evolução da energia mínima e média da população durante a busca pelo estado de menor energia da seqüência HPHHHHHPHHPHPPHPHPPPHPPH que apresenta uma única configuração com $E_{min} = -11$. Esse estado foi encontrado na 42ª geração.

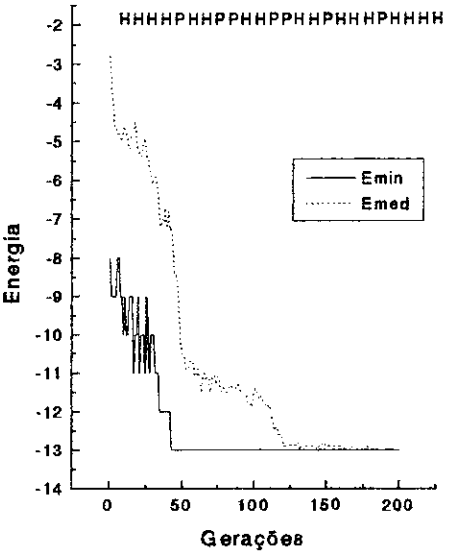


Figura 3-5: A figura mostra a evolução da energia mínima e média da população durante a busca pelo estado de menor energia da seqüência HHHHPHHPPHHPPHHPHHHPHHHH, que apresenta uma única configuração com $E_{min} = -14$. A configuração não foi encontrada em nenhuma das 6 simulações, em 200 gerações.

Capítulo 4

Resultados

4.1 Fatores que afetam o tempo de enovelamento

Ao se tentar determinar os fatores que têm maior influência sobre o enovelamento de polímeros encontram-se diversas possibilidades a serem testadas. Essas possibilidades podem ser divididas em dois grandes grupos, um deles composto por características do polímero que são conhecidas *a priori*, antes de o enovelamento ocorrer, e o outro composto por características que são conhecidas apenas *a posteriori*, em geral, após uma exaustiva busca no espaço de configurações do polímero. O conhecimento de características *a posteriori* e sua relação com o enovelamento podem fornecer informações úteis para o desenvolvimento de modelos que possam representar melhor o comportamento de proteínas reais. Também, o conhecimento da influência de uma característica ligada à estrutura primária, que apresente alguma influência sobre o tempo de enovelamento, abre a possibilidade de se construir algoritmos de busca que utilizem essa informação inicial e, portanto, sejam

mais eficientes.

Entre os fatores incluídos no grupo das características *a priori* podem ser mencionados

- número total de zeros e uns na seqüência primária,
- comprimento da maior subcadeia (de zeros ou de uns) na seqüência primária,
- contribuição de cada freqüência (no sentido de análise de Fourier) para a cadeia,

enquanto no segundo grupo tem-se

- energia mínima,
- tamanho do espaço de configurações,
- degenerescência do estado de menor energia,
- exposição da cauda do polímero,
- energia interna média do espaço de configurações.

Trabalhos anteriores têm mostrado a relação entre alguns desses fatores: Guttman *et al.* [50] obtiveram o número total de conformações válidas possíveis, $\Omega_0(L)$, (considerando apenas volume excluído) para uma seqüência de comprimento L, numa rede quadrada bidimensional:

$$\Omega_0(L) \approx A\,\mu^L\,L^\gamma$$

onde $\mu \approx 2{,}63$ e $\gamma \approx 0{,}333$. Chan & Dill [43], trabalhando com o mesmo modelo HP utilizado aqui, encontraram uma correlação entre a degenerescência do estado de menor energia e a composição hidrofóbica para cadeias de um dado comprimento.

Neste capítulo são definidas várias características de uma seqüência de resíduos HP, relacionadas à seqüência primária e às estruturas de energia mínima. É feita uma análise estatística para verificar se a característica tem uma influência determinante sobre o tempo de enovelamento, e se cada característica pode ser considerada independente dos outros fatores estudados. Todas as análises estatísticas, neste trabalho, foram realizadas considerando um grau de confiança (aceitabilidade) maior ou igual a, 95%, o que corresponde a aceitar coeficientes de regressão maiores, em valor absoluto, do que duas vezes o desvio padrão, em valor absoluto.

Como era de se esperar, foi encontrado que a freqüência de degenerescência do estado de energia mínima, definida como a razão entre o número de configurações com energia mínima e o número total de configurações para uma dada seqüência com N resíduos, desempenha papel importante na facilidade com que o polímero se enovela. De fato, quanto maior o número de estados de energia mínima disponível, mais rapidamente (na média) deve ser possível se chegar a um desses estados. Na próxima seção é analisada esta dependência para o conjunto dos dados.

Nas seções seguintes são definidos alguns parâmetros, associados a características *a priori* e *a posteriori*, e, em seguida, busca-se relacioná-los, estatisticamente, com o tempo de enovelamento, G, da amostra. É possível que uma dependência com respeito a um fator particular seja reflexo de uma forte ligação entre aquele fator e a degenerescência. É necessário, portanto, muito cuidado na determinação de qualquer dependência aparente encontrada com o tempo de enovelamento, para garantir que ela seja independente da degenerescência (ou de qualquer outro fator considerado importante para a análise).

4.2 Degenerescência do estado de energia mínima

Para cada uma das 193 seqüências escolhidas, o tempo de enovelamento, G, definido como o número médio de gerações necessárias para se alcançar uma configuração com energia mínima, obtido após 6 corridas do programa de busca, como descrito no capítulo 3, foi comparado com a freqüência de degenerescência do estado de menor energia, d. A relação foi ajustada a uma lei de potência da forma

$$G = d_0\, d^{\mu} \tag{4.1}$$

pela aplicação de uma regressão linear em $\log(G)$ e $\log(d)$. No presente modelo, a maior degenerescência possível ocorre quando se tem $\Omega(N)$ configurações com energia $E = 0$, o que corresponde ao espaço conformacional inteiro de uma dada seqüência. Teoricamente pode-se argumentar que se a freqüência de degenerescência for 1, i.e., se todas as configurações tiverem a mesma energia, então a energia mínima será encontrada na primeira geração. Substituindo $d_0 = 1$ em (4.1) e tomando os logarítmos, a regressão deverá ser da forma

$$\log(G) = \log(d_0) + \mu\, \log(d),$$

com $\log(d_0) = 0$, i.e., a regressão deve passar pela origem. De fato, uma regressão linear com o pacote Origin, forneceu os resultados da Tabela 4.1.

Pode-se observar que o desvio padrão do coeficiente linear da regressão equivale ao valor absoluto do coeficiente, o que é consistente com com $d_0 = 1$. O coeficiente de

Parâmetro	Valor	σ
$\log(d_0)$	$-0{,}26$	0,27
μ	$-0{,}27$	0,03

Tabela 4.1: Estatisticas para a regressão linear livre de $\log(G)$ com $\log(d)$.

correlação encontrado foi $R = -0{,}50$, de modo que a freqüência de degenerescência responde por aproximadamente 25% da variação em G.

Quando a regressão é forçada a passar pela origem, encontra-se

$$\mu = -0{,}2446 \pm 0{,}0002.$$

O gráfico de G em função de d em escala $\log \times \log$ é mostrado na Figura 4-1.

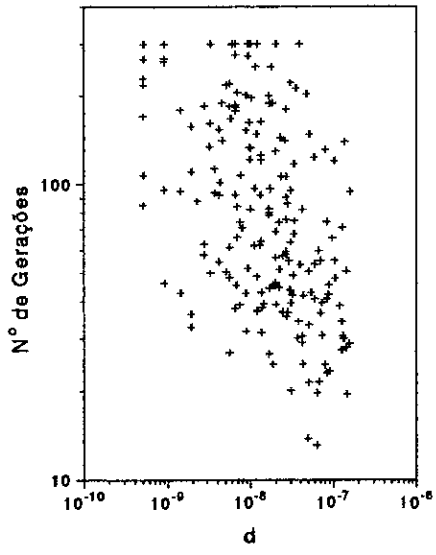


Figura 4-1: Gráfico de G em função de d, em escala $\log \times \log$.

A partir do gráfico da Figura 4-1 pode-se observar que o tempo de enovelamento não depende exclusivamente da degenerescência, uma vez que há seqüências com baixa

degenerescência e rápido enovelamento, bem como seqüências com degenerescência mais alta e enovelamento lento. Isto leva a considerar a possibilidade de identificar outras características das estruturas estudadas, além da degenerescência, que contribuam para facilitar o enovelamento.

A seguir são definidas e analisadas algumas dessas possíveis características, começando por aquelas possíveis de serem identificadas *a priori*.

4.3 Composição hidrofóbica da cadeia

Considerando que, no presente modelo, o enovelamento está baseado na maior probabilidade de reprodução e aceite de estruturas que apresentam monômeros vizinhos na rede do tipo H (com valor 1), conforme descrito no capítulo 3, foi investigada a possibilidade de haver uma relação entre a freqüência, Φ, desses monômeros na seqüência primária e o tempo de busca pelo estado de energia mínima, G.

Trabalhos anteriores indicam que o número de estruturas com energia mínima para uma dada seqüência depende da sua composição hidrofóbica. A Figura 4-2 mostra a correlação entre a degenerescência do estado fundamental e o número de resíduos H, obtida por Chan & Dill [43], para o conjunto de todas as seqüências de comprimento 14. Pode-se observar que não existem seqüências de degenerescência 1 para $\Phi > \frac{10}{14}$ ou $\Phi < \frac{4}{14}$. Além disso, analisando o subespaço com degenerescência 1, das seqüências H-P de tamanho 14, os autores encontraram que a freqüência média de resíduos H nessas cadeias era igual a 0,52.

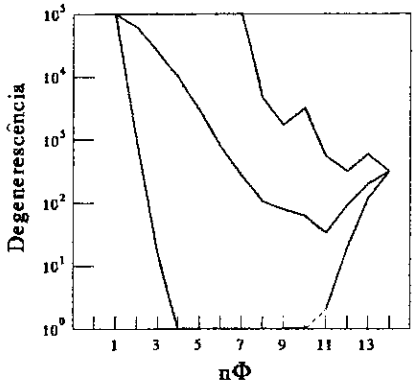


Figura 4-2: Correlação da degenerescência com a composição hidrofóbica para todas as cadeias H/P de comprimento 14, obtida por Chan & Dill [43]. A curva do meio representa a média sobre todas as seqüências com o mesmo Φ. A curva inferior representa a variação com os valores mínimos de Φ e a curva superior com os máximos.

Portanto, no presente trabalho, ao serem selecionadas apenas seqüências com baixa degenerescência, espera-se estar restringindo a amostra a um intervalo de Φ pouco maior do que aquele onde é possível encontrar cadeias com degenerescência igual a 1, ou seja, não se deve encontrar nessas cadeias uma composição hidrofóbica muito maior do que $\frac{10}{14}$ ou muito menor do que $\frac{4}{14}$. De fato, a a distribuição na Figura 4-3 mostra que, para o conjunto de seqüências analisadas neste trabalhos, Φ varia no intervalo $0{,}3 < \Phi < 0{,}75$, que coincide com os extremos mostrados na Figura 4-2. O valor médio de Φ, para o conjunto analisado é igual a 0,52.

É razoável, então, supor que as seqüências estudadas estão em uma faixa onde existe pouca ou nenhuma dependência da freqüência de degenerescência do estado de energia mínima com Φ. A Tabela 4.2 mostra os valores dos coeficientes obtidos para as regressões lineares em $d = d_0\,\gamma^{\Phi}$, da forma

$$\log(d) = \log(d_0) + \Phi\,\log(\gamma),$$

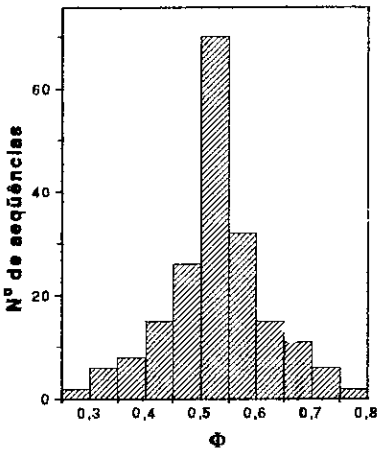


Figura 4-3: Distribuição de Φ entre as 193 seqüências analisadas neste trabalho, com baixa degenerescência do estado fundamental ($< 10^3$). Φ varia entre 0,28 e 0,77. O valor médio de Φ é $0{,}53 \pm 0{,}09$.

e $d = d_0\,\Phi^{\alpha}$, da forma

$$\log(d) = \log(d_0) + \alpha \log(\Phi).$$

Parâmetro	Valor	σ
$\log(d_0)$	$-7{,}8$	0,3
$\log(\gamma)$	$-0{,}03$	0,47
α	$-0{,}08$	0,54

Tabela 4.2: Estatística para as regressões lineares de $\log(d)$ com Φ e $\log(d)$ com $\log(\Phi)$. Os coeficientes angulares das duas regressões são menores, em valor absoluto, do que os respectivos desvios padrões.

Os coeficientes angulares obtidos para os dois tipos de dependência analisados são menores, em valor absoluto, do que os respectivos desvios padrões. Esses resultados reforçam a hipótese de as seqüências analisadas estarem numa faixa em que as duas variáveis são independentes entre si.

Procurou-se, então, estabelecer a existência, ou não, de dependência entre o tempo de enovelamento, G, e a composição hidrofóbica da cadeia, representada por Φ. Regressões lineares de $\log(G)$ com $\log(\Phi)$ e $\log(G)$ com Φ forneceram coeficientes maiores, em valores absolutos, do que dois desvios padrões, com coeficientes de correlação muito próximos.

Nenhuma das duas leis de dependência, $G = G_0\,\Phi^{\alpha}$ ou $G = G_0\,\gamma^{\Phi}$, permite uma extrapolação, fisicamente consistente, para o comportamento quando Φ tende a zero e quando Φ tende a um, provavelmente pelo fato de o conjunto de seqüências apresentar valores de degenerescência baixos e muito próximos. No primeiro caso, para $\Phi = 0$, as seqüências seriam formadas unicamente por monômeros do tipo P, o que daria uma freqüência de degenerescência igual a um. O tempo de enovelamento, neste caso, seria de uma geração pois todas as configurações apresentariam a mesma energia zero. No segundo caso, $\Phi = 1$, todos os monômeros seriam do tipo H e haveria uma grande degenerescência, correspondendo a todos os caminhos sem interseção ligando 22 ou 24 pontos numa rede quadrada 5x5, o que deveria também fornecer um tempo de enovelamento pequeno.

Partindo da observação do gráfico de G em função de Φ, mostrado na Figura 4-4, procurou-se, então, obter alguma indicação qualitativa da possível dependência entre as duas variáveis, no intervalo considerado. Para isso foram traçados dois outros gráficos, tomando-se os tempos médio e mínimo de enovelamento, para seqüências com o mesmo Φ, em função de Φ (Figuras 4-5 e 4-6, respectivamente).

Mesmo considerando que os valores médios e mínimos obtidos não têm, todos, o mesmo peso estatístico, uma vez que para o cálculo de cada um deles foi considerado um número diferente de seqüências, a observação da Figura 4-5 sugere que pode haver

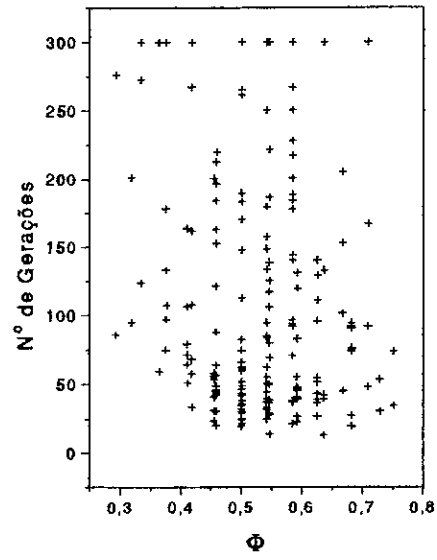


Figura 4-4: Gráfico de G em função de Φ para todas as seqüências estudadas.

um decaimento inicial, enquanto o comportamento para Φ grande não é muito definido. Por outro lado a Figura 4-6 indica um comportamento simétrico em torno de um valor $\Phi_0 \neq 0$, no centro do intervalo.

De posse dessas indicações passou-se a investigar algumas funções que pudessem fornecer uma indicação qualitativa do comportamento da amostra ajuste. Em particular foram utilizadas 2 funções, que compartilham entre si a propriedade de apresentar 3 parâmetros, um a mais do que uma representação linear, o que pode ser considerado um bom equilíbrio entre a restrição imposta por um ajuste linear, e a excessiva liberdade introduzida por uma função com grande número de parâmetros, que podem ser mal determinados e pouco significativos.

A função

$$G = G_0 + G_1\,\Phi + G_2\,\Phi^2, \tag{4.2}$$

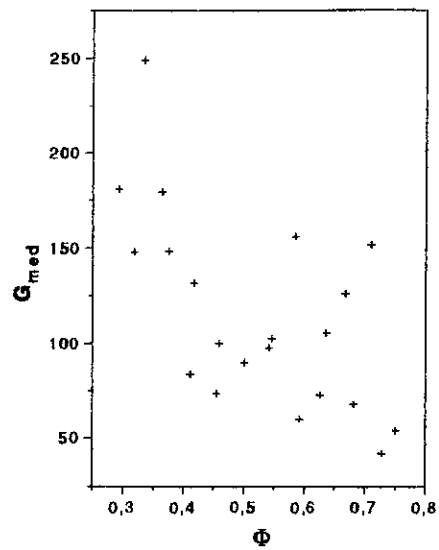


Figura 4-5: Gráfico do tempo médio de enovelamento para as seqüências com o mesmo Φ, em função de Φ.

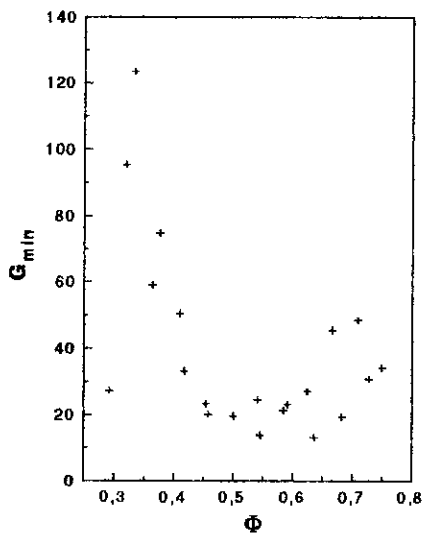


Figura 4-6: Gráfico do tempo mínimo de enovelamento para seqüências com o mesmo Φ, em função de Φ.

corresponde a uma generalização natural de uma relação linear, pelo acréscimo de um termo de dependência quadrática. Além disso, uma parábola é um candidato natural ao ajuste dos pontos representados na Figura 4-6. O ajuste da função ao conjunto dos pontos da Figura 4-4, que representa a amostra completa, forneceu os parâmetros

$$\begin{aligned} G_0 &= 450 \pm 130, \\ G_1 &= -1200 \pm 500, \\ G_2 &= 1000 \pm 480, \end{aligned}$$

com $\chi^2 = 6406$. O mínimo da função se encontra em $\Phi = 0{,}6$.

Outra função,

$$G = G_0 \exp \frac{(\Phi - \Phi_0)^2}{\sigma}, \tag{4.3}$$

que apresenta concavidade semelhante à da função (4.2) forneceu os resultados

$$\begin{aligned} G_0 &= 92 \pm 7 \\ \Phi_0 &= 0{,}60 \pm 0{,}05 \\ \sigma &= 0{,}12 \pm 0{,}05 \end{aligned}$$

com $\chi^2 = 6366$.

Não se pretende com este procedimento encontrar uma expressão analítica para a dependência entre G e Φ, mas sim identificar uma tendência geral do comportamento do tempo de enovelamento em função da composição hidrofóbica, num intervalo onde a

degenerescência tenha pouca influência. Nesse sentido, os bons resultados obtidos com ajuste tanto da função (4.2) quanto da função (4.3) parecem indicar que existe uma freqüência de resíduos hidrofóbicos na cadeia, da ordem de 0,6, para o qual esse tempo é mínimo. Quando Φ se afasta desse valor, em qualquer direção, o tempo de enovelamento aumenta. Os dados, porém, não permitem determinar a forma explícita da dependência. De fato um ajuste de uma função do tipo $\cosh(x)$ também forneceu bons resultados. Por outro lado, os resultados não permitem descartar um decaimento proporcional a $\Phi^{-0,8}$ como aquele obtido pela regressão linear de $\log(G)$ com $\log(\Phi)$.

4.4 Transformada de Fourier

No sentido de identificar possíveis padrões de repetição de monômeros cada seqüência primária foi submetida a uma Transformada de Fourier (FFT). Para isso foi elaborado um programa em PASCAL, baseado no algoritmo apresentado em [51]. A função na qual a transformada foi aplicada corresponde a uma representação da seqüência H-P por uma seqüência de *zeros* e *uns* considerada como uma função contínua, composta por intervalos lineares crescentes (de 0 a 1), decrescentes (de 1 a 0) e constantes (iguais a 0 ou a 1) (Figura 4-7). Esta abordagem procura estabelecer um paralelo com a construção do gráfico de hidrofobicidade, bastante utilizado na análise experimental de seqüências primárias de proteínas reais.

Para verificar se o método seria capaz de indicar um padrão regular de repetições, o programa foi testado com seqüências muito regulares de zeros e uns, de tamanho 32,

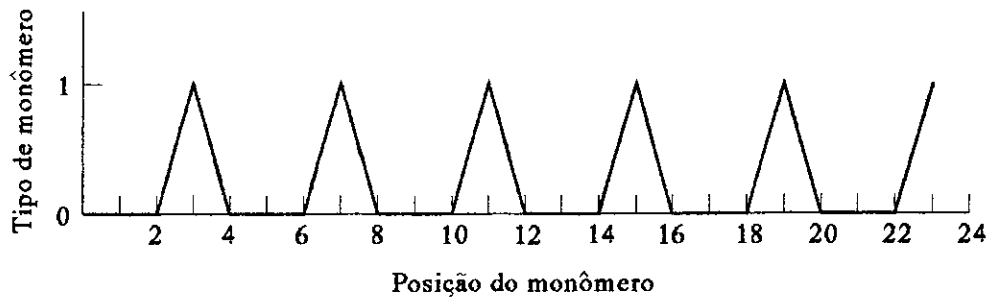


Figura 4-7: Representação da seqüência PPPHPPPHPPPHPPPHPPPHPPPH como uma função contínua, com valores 0 (P) e 1 (H), para a obtenção da Transformada de Fourier.

22 e 24. Como o algoritmo opera com um número de pontos da forma 2^n, e o conjunto de pontos das funções representando as seqüências H-P não obedece a essa relação, pois $2^4 < L < 2^5$, foram acrescentados, ao final da seqüência, tantos zeros quanto necessários para se chegar a 32 pontos, a mesma técnica utilizada no pacote "Origin". O gráfico da Figura 4-8 mostra o resultado obtido para uma seqüência regular, com 32 pontos. A mesma seqüência, porém com 24 pontos, forneceu o resultado mostrado na Figura 4-9. A diferença entre as duas é o acréscimo de zeros para completar 32 pontos na seqüência da Figura 4-9. Podemos observar que nos dois casos obtem-se a mesma freqüência esperada, 0,25. O acréscimo de zeros para completar o número de pontos não interfere com a identificação das freqüências dominantes, embora apareçam interferências de outras freqüências.

Após o teste todas as seqüências foram submetidas ao processo de transformada e identificada a freqüência de maior amplitude de cada seqüência,definida como freqüência dominante f. Um exemplo do resultado para uma dada seqüência aleatória é mostrado na Figura 4-10.

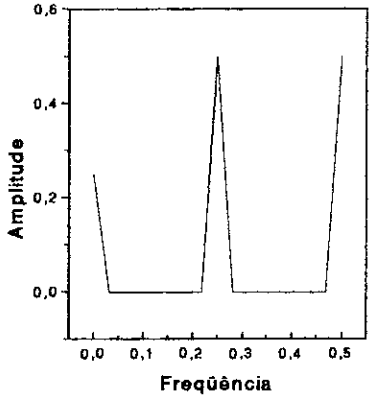


Figura 4-8: Transformada de Fourier para a seqüência regular de 32 monômeros PP-PHPPPHPPPHPPPHPPPHPPPHPPPHPPPH.

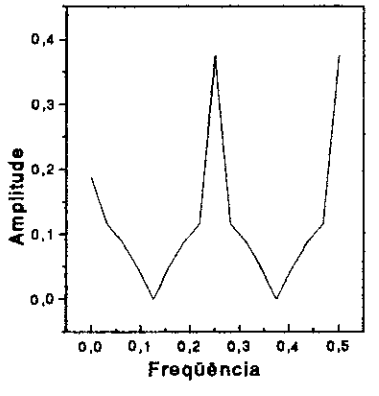


Figura 4-9: Transformada de Fourier para a seqüência regular de 24 monômeros PP-PHPPPHPPPHPPPHPPPHPPPH. A função foi completada com zeros até atingir 32 pontos.

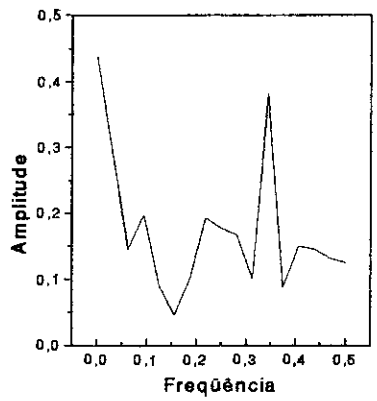


Figura 4-10: Transformada de Fourier para a seqüência aleatória HPHPPHHPHHHH-PHHPPHPPHPHH. Os 32 pontos foram completados com zeros.

A distribuição das freqüências dominantes para o conjunto de dados e para as seqüências que satisfazem a relação

$$\log(G) \leq \mu \log(d).$$

é mostrada na Figura 4-11.

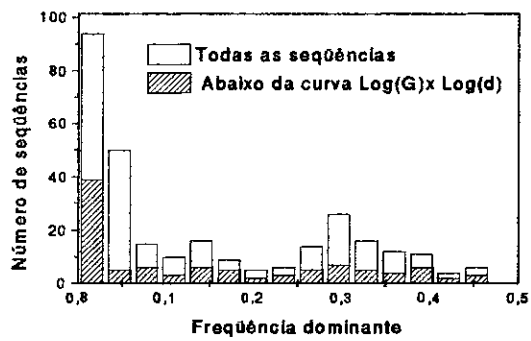


Figura 4-11: Distribuição das freqüências de Fourier dominantes para o conjunto de seqüências estudadas e para aquelas que satisfazem a relação $\log(G) \leq \mu \log(d)$.

A relação entre as freqüências dominantes, f, e o tempo de enovelamento, G, foi

analisada estatisticamente e o resultado da regressão linear é dado por

$$\log(G) = (-0{,}24 \pm 0{,}17)\, f + (1{,}93 \pm 0{,}04).$$

O coeficiente angular da regressão é menor, em valor absoluto, do que duas vezes o desvios padrão, o que indica não haver dependência da forma proposta, entre G e f.

4.5 Comprimento da maior subseqüência de resíduos P e de resíduos H

Ainda dentro da perspectiva de buscar características da seqüencia primária que pudessem estar associadas ao tempo de busca, medimos o comprimento das subseqüências de resíduos H, presentes em cada cadeia, e caracterizamos cada uma delas pelo maior comprimento encontrado, l_{H}. Repetimos o mesmo para as subseqüências P, com o parâmetro l_{P}. A dependência entre G e l_{H} e G e l_{P} foi analisada por uma regressão linear da forma

$$\log(G) = \log(l_0) + l\, \log(\gamma)$$

substituindo-se l por l_{H} ou l_{P}. A análise estatística para o parâmetro l_{H} é mostrada na primeira linha da Tabela 4.3. Dentro do grau de confiança de 95% não foi possível encontrar dependência entre $\log(G)$ e l_{H}.

O mesmo tratamento aplicado a l_{P} forneceu os resultados mostrados na segunda linha

Estatística	$\log(\gamma)$	$\log(l_0)$	R^2
l_H	$0{,}003 \pm 0{,}016$	$1{,}88 \pm 0{,}07$	0,00
l_P	$0{,}09 \pm 0{,}02$	$1{,}59 \pm 0{,}06$	0,12

Tabela 4.3: Estatísticas para as regressões lineares de $\log(G)$ versus l_P e l_H.

da Tabela 4.3. Neste caso, maior subseqüência de P, os dois coeficientes da regressão são maiores do que duas vezes o desvio padrão, o que permite considerar a dependência de G com o parâmetro analisado. O valor de R^2, obtido a partir do coeficiente de correlação (0,34) indica ser l_P responsável por cerca de 12% da variação em G.

Uma análise estatística da relação entre l_P e o logaritmo da freqüência de degenerescência $\log(d)$ indica que são características linearmente independentes, conforme a equação da regressão

$$\log(d) = (0{,}017 \pm 0{,}034)\, l_P - (7{,}91 \pm 0{,}12).$$

Por outro lado deve-se tomar cuidado ao procurar estabelecer a lei de dependência entre G e l_P. Observando o gráfico da Figura 4-12 pode-se pensar que o tempo de busca cresce à medida que o tamanho da subseqüência aumenta. Porém, é necessário considerar que o conjunto de dados analisado foi escolhido de modo a representar apenas aquelas cadeias que têm degenerescência relativamente pequena, e não o conjunto de todas as seqüências de um dado comprimento. Além disso, quando $l_P = L$, i.e., quando todos os resíduos são do tipo P, a freqüência de degenerescência torna-se 1 e, portanto, G também deverá ser igual a 1, o oposto do que ocorre no gráfico da Figura 4-12.

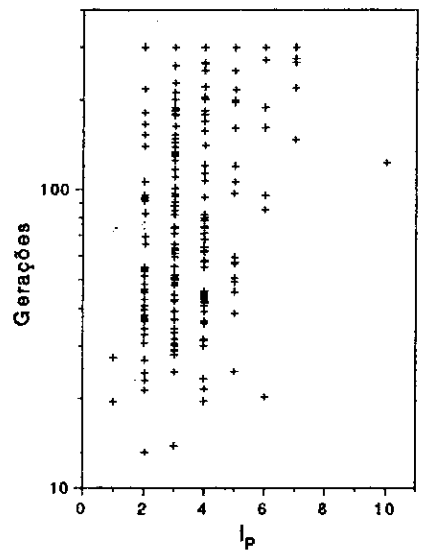


Figura 4-12: Gráfico de G versus l_P com G em escala logarítmica.

Nas próximas seções serão consideradas as características conhecidas após a investigação do espaço de configurações, relacionadas às estruturas de menor energia.

4.6 Grau de proteção das extremidades

Uma situação que pode dificultar a obtenção da estrutura de energia mínima diz respeito à profundidade em que uma ou as duas extremidades do polímero estão situadas, dentro da configuração enovelada. Para verificar a influência da distância das extremidades à borda da estrutura sobre o tempo de enovelamento, foi desenvolvida uma medida do "grau de proteção" da extremidade de um polímero, descrita a seguir. Um monômero é definido como *interno* se todas as oito posições em torno dele, na rede quadrada, são ocupadas por outros monômeros ou, se vazias, inteiramente cercadas pela cadeia. A Figura 4-13 mostra uma estrutura típica, com monômeros internos representados em

preto e monômeros externos em branco. Pode-se, então, definir o grau de proteção da extremidade do polímero como o número de ligações que devem ser percorridas, a partir da extremidade, ao longo da cadeia, até encontrar um monômero externo. No exemplo da Figura 4-13 esse grau é 4 para a extremidade mais interna.

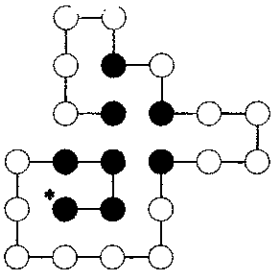


Figura 4-13: Cálculo do grau de proteção das extremidades de uma cadeia. Círculos vazios correspondem a monômeros externos e círculo cheios a monômeros internos. O monômero na extremidade assinalada com * está a uma distância de 4 ligações do monômero externo mais próximo, que é o de número 5. Portanto, para essa extremidade o grau de proteção é 4. Para a outra extremidade o grau é 1.

Na elaboração de um algoritmo para obter esses valores devem ser levadas em conta situações como aquela mostrada na Figura 4-14, nas quais a extremidade do polímero é adjacente a "vazios" internos à estrutura enovelada. Na prática, raramente apareçem buracos internos com mais de um ou dois pontos vazios, uma vez que as estruturas com energia mínima tendem a ser muito compactas, porém a ocorrência desses casos não pode ser descartada.

É razoável supor que o enovelamento do polímero será dificultado pelo grau de proteção das extremidades, C, de acordo com a relação

$$G = C_0 \rho^C,$$

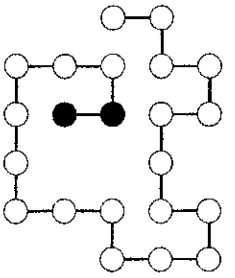
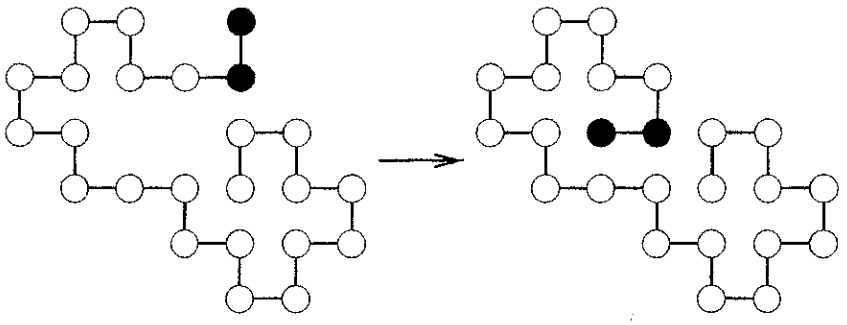


Figura 4-14: Um exemplo de uma configuração com buraco interno. Os monômeros destacados em preto são vizinhos de espaços vazios no interior da configuração.

onde ρ e C_0 podem ser determinados estatisticamente. A justificativa para o modelo pode ser obtida a partir da análise da Figura 4-15, supondo-se que, para atingir a estrutura de energia mínima dada pela configuração B seja necessário partir da estrutura mais aberta, representada em A. Cada um dos dois últimos monômeros da cadeia deverão ter suas coordenadas modificadas de forma única para que se encaixem nas novas posições. Se a probabilidade de ocorrer uma modificação, que leve um monômero para uma coordenada particular, for ρ, então a probabilidade de combinar duas modificações que coloquem dois monômeros nas posições corretas será ρ^2. Portanto, em geral, para obter uma estrutura alvo compacta única, a partir de uma estrutura aberta, é introduzido um fator de dificuldade da forma ρ^C, onde C é a distância da extremidade mais interna da cadeia até a borda da estrutura, definida como grau de proteção da extremidade do polímero.

Para o conjunto de configurações de energia mínima de uma dada seqüência, foram formuladas e testadas três possíveis definições da dificuldade introduzida por extremidades protegidas. Obviamente é esperado que todas elas sejam altamente interdependentes, mas o ponto em questão é selecionar qual das três definições responde melhor ao ajuste estatístico. As definições utilizadas foram



A B

Figura 4-15: Formação de uma estrutura fechada a partir de outra aberta. Para passar da estrutura A para a estrutura B cada um dos dois monômeros da extremidade superior em A deverão ter suas coordenadas modificadas de forma única para que se posicionem como em B.

DM1 Média das distâncias das duas extremidades à borda, para o conjunto de configurações de energia mínima.

DM2 Média das distâncias da extremidade mais interna à borda, para o conjunto de configurações de energia mínima.

DA1 Distância da extremidade mais interna à borda, para a configuração de energia mínima mais aberta. Para obter esta grandeza, primeiro era identificado o valor de C para a extremidade mais interna de cada configuração e, em seguida, escolhido o menor desses valores para representar o conjunto.

A justificativa para a definição do parâmetro DA1 é que, dado um conjunto de configurações de energia mínima, o processo de enovelamento do polímero irá conduzir mais naturalmente àquela estrutura mais fácil de atingir. Portanto, para medir a dificuldade de atingir uma configuração de energia mínima, a estrutura mais fácil de obter caracteriza melhor o grupo do que a média.

As estatísticas da regressão linear em $\log(G)$ e C para as três definições do parâmetro C forneceu os resultados da Tabela 4.4.

Estatística	$\log(\rho)$	$\log(C_0)$	R^2
DM1	0,08 ± 0,03	1,84 ± 0,04	0,02
DM2	0,04 ± 0,02	1,84 ± 0,04	0,02
DA1	0,11 ± 0,02	1,83 ± 0,03	0,12

Tabela 4.4: Estatísticas da regressão linear de $\log(G)$ contra C.

Os dados da tabela indicam que todas as estatísticas são aceitáveis num grau de confiança de duas vezes o desvio padrão, entretanto existe uma maior dependência com a estatística DA1, evidenciada no valor de R^2, confirmando a justificativa apresentada na definição dos parâmetros.

Na seção 4.2 foi encontrado que a freqüência de degenerescência responde por cerca de 25% da variação em G. Se os dois fatores analisados forem linearmente independentes, poderiamos pensar que a degenerescência e o grau de internalização deverão ser responsáveis, em conjunto, por cerca de 37% da variação em G. Porém uma análise estatística da relação entre $\log(d)$ e C forneceu a relação

$$\log(d) = (-0{,}43 \pm 0{,}11)\, C + (-2{,}0 \pm 0{,}7).$$

Isso indica dependência entre C e $\log(d)$, pois o coeficiente angular é maior, em valor absoluto do que duas vezes o desvio padrão. Como a freqüência de degenerescência explica cerca de 25% da variação de G contra os 12% explicados pelo grau de internalização, é possível que esta última relação se deva à dependência de C com d, sendo d o fator

primário na variação de G. A diminuição da freqüência de degenerescência com o aumento do grau de proteção das extremidades possivelmente está relacionada com a diminuição da liberdade para o arranjo dos monômeros à medida que as extremidades se localizam mais internamente na estrutura compacta.

4.7 Ordem do contato entre dois monômeros

Chan & Dill [16], trabalhando com um modelo de enovelamento de polímero em rede quadrada bidimensional, e considerando apenas efeito de volume excluido, observaram que a presença de um primeiro contato presumido entre quaisquer dois monômeros da cadeia influencia fortemente a formação de um segundo contato e que a liberdade conformacional do polímero dependerá da distância entre os monômeros do contato presumido. Eles definiram a *ordem do contato*, k, como a distância entre dois monômeros em contato, medida ao longo do comprimento da cadeia (Figura 4-16).

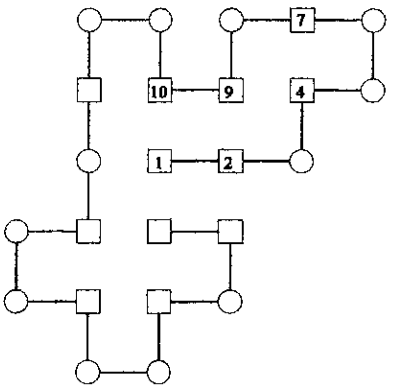


Figura 4-16: A ordem, k, do contato entre dois monômeros é definida como a distância ao longo da cadeia entre os dois. Assim, os monômeros 1 e 10 formam um contato de ordem 9, os monômeros 2 e 9 formam um contato de ordem 7, os monômeros 4 e 7, um contato de ordem 3 e os monômeros 4 e 9 formam um contato de ordem 5.

Com base nessa idéia, buscou-se uma relação entre a ordem dos contatos presentes nas configurações de energia mínima de cada seqüência e o tempo de enovelamento G. Uma vez que, no presente modelo, os contatos relevantes para o cálculo de energia, e, portanto, para a liberdade conformacional, são aqueles entre monômeros H, foi identificada a ordem de cada um dos contatos desse tipo em cada conformação de energia mínima. Cada seqüência foi, então, caracterizada por uma ordem média $\overline{k}$, obtida a partir da média dos valores de todos os k para todas as suas configurações de energia mínima. Como as seqüências analisadas têm comprimentos diferentes, a ordem média dos contatos foi normalizada pelo tamanho, N, da seqüência. A Figura 4-17 mostra um exemplo do cálculo de $\overline{k}$ para uma seqüência de 24 monômeros que apresenta 3 configurações com energia mínima.

A relação entre G e $\overline{k}$ foi ajustada por uma expressão da forma

$$G = G_k \lambda^{\overline{k}}$$

por meio de uma regressão linear sobre $\log(G)$ e $\overline{k}$, cujo resultado é a relação

$$\log(G) = (1{,}39 \pm 0{,}12) + (1{,}5 \pm 0{,}4)\,\overline{k}.$$

O parâmetro $\log(\lambda)$ obtido é maior do que 2 desvios padrões. O coeficiente de regressão é 0,29 de modo que a ordem média dos contatos é responsável por cerca de 8% da variação em G. O gráfico de G em função de $\overline{k}$, com G em escala logarítmica, é mostrado na

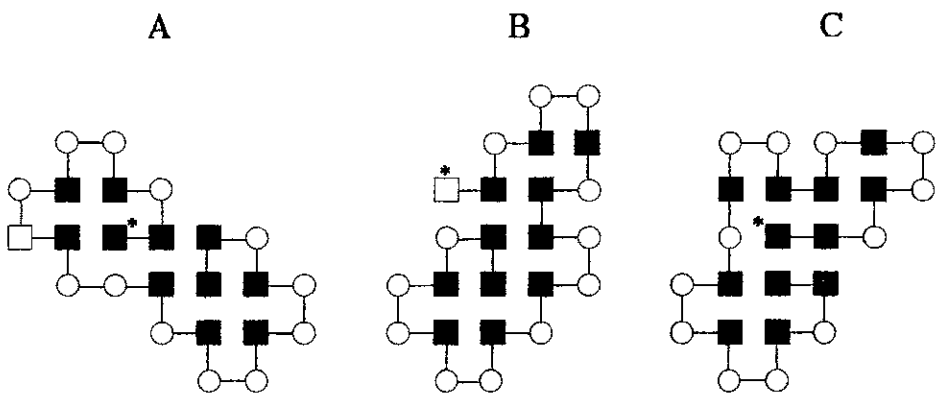


Figura 4-17: Cálculo do valor de $\overline{k}$ para uma seqüência com degenerescência do estado fundamental igual a 3. Para cada uma das configurações da figura são calculados a quantidade e a ordem (k_n) de cada contato. Esses valores são mostrados na Tabela 4.5. A ordem média dos contatos da seqüência, $\overline{k}$, é dada pela média das ordens de todas as configurações, normalizada pelo tamanho da cadeia. Para o conjunto das estruturas mostradas $\overline{k} = 0{,}31$. Os asteriscos indicam o início da cadeia.

Estrutura/k_n	3	5	7	9	11	13	21	23
A	6	–	–	2	2	–	1	–
B	5	1	1	1	1	1	1	–
C	5	1	1	2	–	–	1	1

Tabela 4.5: Quantidade e tipo dos contatos presentes nas estruturas mostradas na Figura 4-17. A média dos valores encontrados normalizada pelo tamanho da seqüência fornece $\overline{k} = 0{,}31$.

Figura 4-18. Estes resultados permitem afirmar que o fator $\overline{k}$ tem alguma influência no tempo de busca, indicando que, à medida que a ordem do contato cresce, o tempo de busca também aumenta, isto é, contatos de ordem mais baixa devem favorecer o enovelamento.

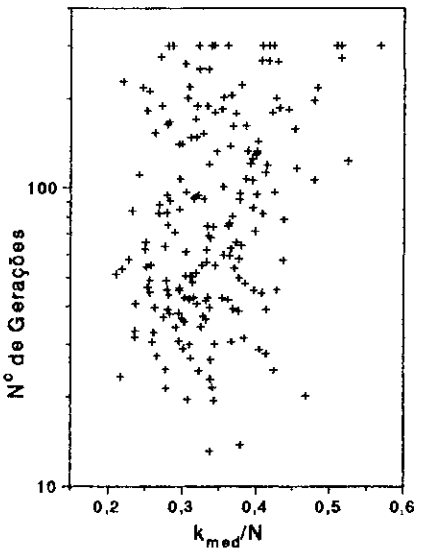


Figura 4-18: Gráfico de G em função de $\overline{k}$, com G em escala logarítmica.

Além disso, uma análise estatística da relação entre a ordem média dos contatos, $\overline{k}$, e a freqüência de degenerescência, d, indica que são características independentes, conforme a expressão abaixo:

$$\log(d) = (0{,}03 \pm 0{,}7)\,\overline{k} - (7{,}8 \pm 0{,}2).$$

Uma regressão múltipla nas variáveis $\overline{k}$ e $\log(d)$ forneceu um coeficiente de regressão $R = 0{,}588$, significando que os dois fatores juntos são responsáveis por cerca de 35% da variação em G, consistentemente com a análise individual das dependências.

4.8 Número de contatos de ordem 3

Simulações de enovelamento, considerando apenas efeito de volume excluído, realizadas por Chan & Dill [16] na rede quadrada bidimensional indicaram que hélices são as conformações mais favorecidas dentre todas as que apresentam dois contatos de ordem 3 quando se trata de preencher compactamente o plano. Por outro lado, os resultados da seção anterior indicam que contatos de baixa ordem facilitam o enovelamento. Para verificar a influência de contatos de baixa ordem sobre o tempo de enovelamento da cadeia foi definido o parâmetro $\overline{k}_3$ como o número médio de contatos de ordem 3 presentes nas configurações de energia mínima, normalizado pelo tamanho da seqüência. O fator foi então relacionado com o tempo de enovelamento, G.

O resultado de uma regressão linear da forma

$$\log(G) = \log(\alpha)\,\overline{k}_3 + \log(G_0)$$

forneceu os coeficientes

$$\log(G) = (-1{,}31 \pm 0{,}47)\,\overline{k}_3 + (2{,}08 \pm 0{,}07),$$

com um valor de $R^2 = 0{,}04$. A dependência encontrada é consistente com aquela encontrada na seção anterior, i.e., um maior número de contatos de ordem 3 parece favorecer o enovelamento. Um gráfico da dependência entre os dois fatores é mostrado na Figura 4-19. Um análise da relação entre $\overline{k}_3$ e $\log(d)$ indica que os dois fatores podem ser

considerados linearmente independentes.

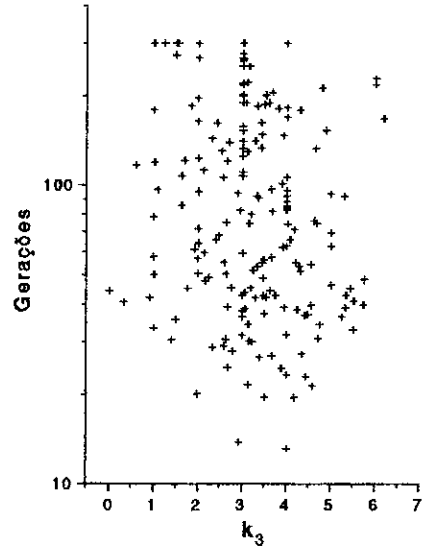


Figura 4-19: Gráfico de G em função de $\bar{k}_3$, com G em escala logarítmica.

4.9 Segundo momento das distâncias dos monômeros H nas configurações de energia mínima

A compactação das estruturas de menor energia é uma característica de polímeros com propriedades similares às de proteínas. Durante o processo de enovelamento, estruturas completamente esticadas tendem rapidamente para configurações mais enoveladas somente em função da entropia. Por outro lado, o grau de compactação da estrutura não é suficiente para caracterizar o estado de energia mínima de uma dada seqüência. Para o modelo aqui utilizado, esse estado é caracterizado pelo maior número possível de contatos entre monômeros hidrofóbicos. Portanto as estruturas de menor energia tenderão a

formar um caroço hidrofóbico compacto.

Nesta seção, para medir quão compactamente estão arranjados os monômeros H nas estruturas de menor energia, foi desenvolvida uma analogia entre a distribuição desses monômeros numa configuração fechada e uma distribuição superficial de massas discretas.

O momento de inércia de um conjunto de massas discretas, m_i, situadas nas coordenadas (x_i, y_i) de um plano, em torno de um eixo perpendicular ao plano, passando pelo centro de massa $(\overline{x}, \overline{y})$, é definido como

$$I = \sum_i m_i r_i^2 = \sum_i m_i[(x_i - \overline{x})^2 + (y_i - \overline{y})^2]. \tag{4.4}$$

Nessa analogia foi tomado $m_i = 0$ ou 1, correspondendo ao valor do monômero naquele ponto (H=1 e P=0). O centro de massa foi substituído pelo centro geométrico dos resíduos H na configuração, com a massa total, M, da estrutura igual ao número total de monômeros hidrofóbicos na cadeia, $N\Phi$:

$$\overline{x} = \frac{1}{N\Phi} \sum_i m_i x_i, \quad \overline{y} = \frac{1}{N\Phi} \sum_i m_i y_i.$$

Com o momento de inércia dos monômeros hidrofóbicos definido como na equação (4.4), cada cadeia foi caracterizada por um valor I_{H} correspondendo à média sobre as estruturas de energia mínima. Um gráfico do tempo de enovelamento em função de I_{H} é mostrado na Figura 4-20.

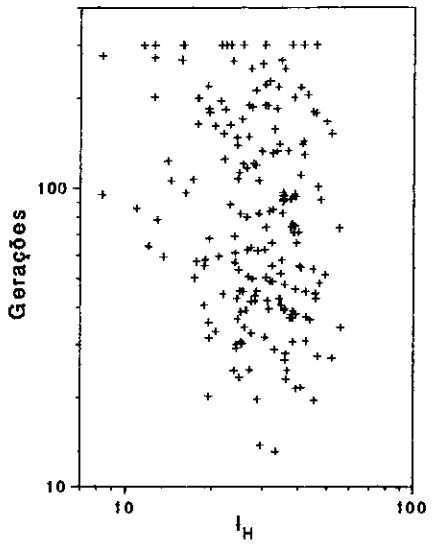


Figura 4-20: Gráfico do número de gerações em função do segundo momento das distâncias dos monômeros H em relação ao centro do caroço hidrofóbico em escala log × log.

Uma regressão da forma

$$\log(G) = \log(I_0) + q \log(I_H)$$

foi testada e produziu os resultados mostrados na Tabela 4.6. A dependência é consistente com o comportamento esperado de G quando I_H tende a zero. Dentro do grau de confiança de 95% utilizado neste trabalho, pode-se observar que o número médio de gerações necessárias para encontrar uma estrutura com energia mínima é proporcional a $I_H^{-0.5}$.

Parâmetro	Valor	σ
$\log(I_0)$	2,62	0,22
q	−0,50	0,15
R^2	0,05	

Tabela 4.6: Estatística para a regressão linear $\log(G) = \log(I_0) + q\ \log(I)$.

4.10 Dependência entre os fatores analisados

Na Tabela 4.7 é apresentado um quadro que resume os resultados qualitativos obtidos para os fatores analisados.

	d	Φ	f	l_{H}	l_{P}	C	$\overline{k}$	$\overline{k}_3$	I_{H}
G	sim	sim	não	não	sim	sim	sim	sim	sim
d	—	não	—	—	não	sim	não	não	não

Tabela 4.7: Existência ou não de dependência entre tempo de enovelamento (G) e degenerescência do estado de menor energia (d) e todos os outros fatores.

Dos 9 fatores analisados foram encontrados 7 que apresentam algum tipo de influência no tempo de enovelamento: a freqüência de degenerescência (d), o coeficiente de hidrofobicidade (Φ), o comprimento da maior subseqüência de monômeros do tipo P (l_{P}), o grau de proteção das extremidades (C), a ordem me'dia dos contatos ($\overline{k}$), o número médio de contatos de ordem 3 ($\overline{k}_3$) e o segundo momento das distâncias dos monômeros H nas configurações de energia mínima (I_{H}). Além disso, 5 deles são independentes da freqüência de degenerescência (d) e apenas o grau de proteção das extremidades, (C), apresentou essa dependência.

Uma vez identificados os parâmetros que têm alguma influência no tempo de enovela-

mento e que são independentes da freqüência de degenerescência, passou-se a analisar a dependência mútua entre eles. Uma análise estatística desses 5 parâmetros revelou uma interdependência entre todos eles. A Tabela 4.8 apresenta os coeficientes de correlação obtidos a partir de regressões lineares aplicadas a cada par de fatores.

Variáveis	$\overline{k}_3$	$\overline{k}$	l_{P}	$\log(\Phi)$
$\log(I_{\mathrm{H}})$	0,54	−0,33	0,69	0,83
$\overline{k}_3$	—	−0,49	−0,53	0,59
$\overline{k}$	—	—	0,45	−0,25
l_{P}	—	—	—	0,64

Tabela 4.8: Coeficientes de correlação para as regressões lineares entre as características, independentes de $\log(d)$. As variáveis na primeira linha são aquelas consideradas independentes na regressão.

Os coeficientes da Tabela 4.8 indicam uma forte correlação entre o segundo momento dos resíduos hidrofóbicos nas configurações compactas, I_{H}, e o grau de hidrofobicidade da seqüência primária, Φ, bem como entre I_{H} e o comprimento da maior subseqüência de resíduos P, l_{P}.

A influência relativa de cada fator no tempo de enovelamento pode ser obtida da observação da Tabela 4.9 que apresenta os coeficientes de correlação para as regressões lineares entre $\log(G)$ e cada fator analisado.

A maior dependência encontrada foi com a freqüência de degenerescência, d, que responde por cerca de 25% da dependência. Em seguida vêm os fatores tamanho da maior subseqüência de resíduos P, l_{P}, e presença de contatos de ordem 3, $\overline{k}_3$, respondendo, cada um por cerca de 11%. A menor dependência foi encontrada com a composição hidrofóbica, Φ, responsável por 4% da dependência.

Variável Independente	R
$\log(d)$	−0,50
l_{P}	0,34
$\overline{k}_3$	0,33
$\overline{k}$	0,29
$\log(I_{\mathrm{H}})$	−0,23
$\log(\Phi)$	−0,19

Tabela 4.9: Valores dos coeficientes de correlação R para as regressões de $\log(G)$ em função de $\log(d)$, $\log(\Phi)$, l_{P}, $\overline{k}$, $\overline{k}_3$ e $\log(I_{\mathrm{H}})$.

Como, com excessão da freqüência de degenerescência, todos os fatores estão relacionados entre si, podemos destacar l_{P} entre as características *a priori* e $\overline{k}_3$ entre as características *a posteriori*, como os mais significativos no tempo de enovelamento. A relação entre os dois pode ser descrita da seguinte maneira: se l_{P} é grande, a ordem do contato entre os dois monômeros H, localizados nas extremidades das subseqüências também será grande. Portanto, espera-se que uma subseqüência grande de resíduos P aumente o tempo de enovelamento, uma vez que reduz o número de contatos de ordem 3 entre monômeros H.

Conclusão

Durante o desenvolvimento deste trabalho foram definidos e analisados parâmetros, relacionados às seqüências dos monômeros e às estruturas de menor energia de polímeros, que pudessem afetar o tempo de enovelamento numa rede quadrada bidimensional. Os polímeros foram definidos como seqüências de dois tipos de monômeros, H e P, e as interações consideradas apenas entre pares de vizinhos não conectados HH.

Um conjunto de seqüências foi gerado e seqüências com baixa degenerescência foram selecionadas após exploração exaustiva do espaço conformacional, por meio de um programa desenvolvido para esse fim. Essa amostra foi utilizada na análise estatística da dependência do tempo de enovelamento com cada um dos parâmetros definidos. O tempo de enovelamento, G, foi definido a partir de parâmetros obtidos por meio de um procedimento de busca baseado em algoritmo genético. Os resultados das simulações mostraram haver seqüências com rápido enovelamento e outras com enovelamento mais lento, bem como algumas que não se enovelavam até o limite de passos computacionais definidos pelo procedimento de busca.

Uma análise estatística mostrou que a freqüência de degenerescência, d, do estado de menor energia das estruturas enoveladas respondia por 25% da dependência com o tempo

de enovelamento na amostra. Portanto, a partir dessa primeira constatação, tornou-se clara a necessidade de buscar outros parâmetros, tanto das seqüências primárias H-P, quanto das estruturas de menor energia, que pudessem afetar, em maior ou menor grau, o tempo de enovelamento.

Dos 8 fatores analisados, com influência sobre G, foram encontrados 5 independentes da freqüência de degenerescência, d. Apenas um dos fatores com influência sobre o tempo de enovelamento, o grau de proteção, C, das extremidades do polímero nas configurações de menor energia, apresentou dependência também com a freqüência de degenerescência. Não foi possível, portanto, avaliar o grau de dependência exclusiva de G com C, ou mesmo se a dependência encontrada se deve apenas à dependência com d. Para isso, uma possível abordagem seria analisar o fator para um conjunto de seqüências com a menor degenerescência possível, isto é, com apenas uma estrutura de energia mínima.

Todos os parâmetros analisados, com excessão da freqüência de degenerescência, são interdependentes em maior ou menor grau. O maior coeficiente de correlação foi encontrado entre o grau de hidrofobicidade, Φ, e o segundo momento dos monômeros H, I_{H}, numa regressão log × log. A menor correlação, apareceu entre a ordem média dos contatos nas estruturas de energia mínima, $\overline{k}$, e $\log(\Phi)$. A dependência mútua encontrada para todos os fatores analisados merece uma investigação mais detalhada, utilizando recursos computacionais mais poderosos, que permitam gerar populações de seqüências maiores e mais homogêneas, para uma análise estatística mais refinada.

Fatores relacionados com o alcance das interações, como a freqüência de contatos de ordem 3, $\overline{k}_3$, o comprimento da maior subseqüência de monômeros P, e a ordem média dos

contatos apresentaram maior influência sobre o tempo de enovelamento do que aqueles relacionados com a presença de monômeros hidrofóbicos nas cadeias, como o grau de hidrofobicidade, Φ, e o segundo momento dos monômeros hidrofóbicos, I_H.

O fator $\overline{k}_3$ está relacionado a interações locais nas estruturas nativas e o resultado aqui obtido sugere que a presença de interações desse tipo favorecem o enovelamento, resultado que também aparece na análise de $\overline{k}$, ordem média de todos os contatos. Na literatura, temos resultados de modelos teóricos em rede que prevêm tanto um menor tempo de enovelamento com um maior número de contatos locais [52], [53] quanto com um maior número de contatos não-locais [54]. Por outro lado, o resultado encontrado neste trabalho é consistente com aqueles descritos em publicação recente [55], que indicam uma correlação significativa entre o tempo de enovelamento e a ordem média dos contatos para um conjunto de 12 proteínas reais, não homólogas, com um único domínio. Esse tempo foi menor para proteínas com ordem média dos contatos mais baixa. Esse resultado é bastante interessante, considerando a simplicidade do modelo, em contraste com modelos mais realísticos (3D) que fornecem resultados opostos [54]. Isto permite pensar na utilização sistemática do modelo para investigar o papel desses parâmetros no enovelamento.

A análise dos dados também sugeriu uma possível dependência entre o tempo de enovelamento, G e a composição hidrofóbica, Φ, das cadeias. A pequena correlação obtida para a regressão linear de $\log(G)$ com $\log(\Phi)$ parece reforçar a hipótese de que a dependência entre esses dois fatores, no intervalo considerado, não é bem descrita por uma lei de potência. O comportamento simétrico, em torno de um valor diferente de zero,

das funções que se ajustaram bem à relação $G \times \Phi$, parece apontar para a existência de um valor mínimo de Φ, ou uma faixa de valores, no qual o tempo de enovelamento seja mínimo. Os dados disponíveis não permitiram, porém, testar essa hipótese. Para verificar a validade desta observação um caminho possível é analisar o tempo de enovelamento para seqüências de um dado comprimento N, num intervalo bastante restrito de Φ. Inicialmente devemos encontrar os valores extremos de Φ, para os quais existem seqüências de comprimento N com degenerescência do estado fundamental única e em seguida gerar essas seqüências para posterior análise. Além disso, nas extremidades do intervalo analisado, onde o tempo de enovelamento é máximo, deve ocorrer uma descontinuidade da função, uma vez que em $\Phi = 0$, onde a degenerescencia é máxima e em $\Phi = 1$ onde a degenerescência também é grande, o tempo de enovelamento deve ser mínimo.

Outros parâmetros como a freqüência de Fourier dominante, obtida a partir da aplicação de uma Transformada de Fourier às seqüências, tomadas como funções com valores zero ou um, e o comprimento da maior subseqüência de monômeros H na cadeia parecem não interferir diretamente no tempo de enovelamento. Apesar de parecer uma característica importante, que pudesse refletir algum grau de regularidade na distribuição de resíduos H e P nas seqüências que enovelam mais rapidamente, a freqüência dominante de cada seqüência não forneceu nenhum tipo de informação que pudesse ser relacionada com o tempo de enovelamento. Por outro lado, é possível que uma investigação mais elaborada com o uso da análise de Fourier possa fornecer outras informações relevantes que permitam estabelecer uma relação com os parâmetros analisados.

Por fim, é importante ressaltar que neste trabalho optou-se por uma abordagem que

permitisse, em primeiro lugar, identificar alguns parâmetros relevantes para o tempo de enovelamento. Nesse sentido, foi utilizada uma amostra sobre a qual uma análise estatística preliminar pôde selecionar aqueles que poderão servir para definir uma amostra mais homogênea que possa sevir de base para uma investigação mais refinada do papel de cada fator no processo de enovelamento.

Bibliografia

[1] D. Voet and J. Voet. *Biochemistry.* John Wiley & Sons, Inc., New York, 1995.

[2] C. B. Anfinsen. *Science*, 181:223–230, 1973.

[3] M. J. E. Stern. In *Protein Structure Prediction*, pages 1–30. IRL Press, Oxford, 1996. Ed: Michael J. E. Sternberg.

[4] T. E. Creighton. *J. Phys. Chem.*, 89:2452–2459, 1985.

[5] H. Wu. *apud* K. A. Dill. *Biochemistry*, 29:7133–7155, 1990.

[6] M. M. Santoro and D. W. Bolen. *Biochemistry*, 27:8063, 1988.

[7] P. L. Privalov. *Annu. Rev. Biophys. Biophys. Chem.*, 18:47, 1989.

[8] M. B. Yaffe et al. *Nature*, 358:245–248, 1992.

[9] V. A. Lewis, G. M. Hynes, D. Zheng, H. Saibil, and K. Willison. *Nature*, 358:249–252, 1992.

[10] C. D. Sfatos, A. M. Gutin, V. I. Abkevich, and E. I. Shakhnovich. *Biochemistry*, 35:334–339, 1996.

[11] R. L. Baldwin. *Nature*, 369:183–184, 1994.

[12] K. A. Dill. *Biochemistry*, 29:7133–7155, 1990.

[13] P.K. Ponnuswamy and M. M. Gromiha. *J. Theor. Biol.*, 166:63–74, 1992.

[14] D. O. V. Alonso and K. A. Dill. *Biochemistry*, 30:5975–5985, 1991.

[15] G. Weber. *J. Phys. Chem.*, 97:7108–7115, 1993.

[16] H. S. Chan and K. A. Dill. *J. Chem. Phys*, 90(1):492–509, 1988.

[17] H. S. Chan and K. A. Dill. *J. Chem. Phys.*, 92:3118–3135, 1990.

[18] A. Sikorski and J. Skolnick. *Proc. Natl. Acad. Sci. U.S.A*, 86:2668, 1989.

[19] J. Skolnick, A. Kolinski, and R. Yaris. *Proc. Natl. Acad. Sci. U.S.A*, 86:1229, 1989.

[20] H. S. Chan and K. A. Dill. *Macromolecules*, 22:4559, 1989.

[21] G. D. Rose, L. M. Gierasch, and J. A. Smith. *Adv. Protein Chem.*, 37:1–109, 1985.

[22] M. Karplus and D. L. Weaver. *Nature*, 260:404–406, 1976.

[23] C. Levinthal. *J. Chem. Phys.*, 65:44–45, 1965.

[24] T. E.Creighton. *Biochem. J.*, 270:1–16, 1990.

[25] O. B.Ptitsyn. *J.Prot.Chem.*, 6:272–293, 1991.

[26] M. J. Gething and J. Sambrook. *Nature*, 355:33–45, 1992.

[27] E. Shakhnovich, G. Farztdinov, A. M. Gutin, and M. Karplus. *The American Physical Society*, 67:1665–1668, 1991.

[28] A. Sali, E. Shakhnovich, and M. Karplus. *Nature*, 369:248–251, 1994.

[29] E. I. Shakhnovich and A. M. Gutin. *Nature*, 346:773–775, 1990.

[30] C. Chothia. *J. Mol. Biol.*, 105:1, 1976.

[31] P. G. Pascutti, K. C. Mundim, A. S. Ito, and P. M. Bisch. submitted to *Biophys. J.*, 1998.

[32] M. E. Karpen and C. L. Brooks III. In *Protein Structure Prediction*, pages 229–262. IRL Press, Oxford, 1996. Ed: Michael J. E. Sternberg.

[33] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus. *J. Comp. Chem.*, 4:187–217, 1983.

[34] G. J. Barton. In *Protein Structure Prediction*, pages 31–64. IRL Press, Oxford, 1996. Ed: Michael J. E. Sternberg.

[35] N. Srinivasan, K. Guruprasad, and T. L. Blundell. In *Protein Structure Prediction*, pages 111–140. IRL Press, Oxford, 1996. Ed: Michael J. E. Sternberg.

[36] N. Gō and H. Abe. *Biopolymers*, 20:1013–1031, 1981.

[37] W. R. Krigbaum and S. F. Lin. *Macromolecules*, 15:1143–1145, 1982.

[38] J. Skolnick and A. Kolinski. *J. Mol. Biol.*, 212:787–817, 1989.

[39] E. I. Shakhnovich. *Phys. Rev. Letters*, 72:3907–3910, 1994.

[40] E. M. O'Toole and A. Z. Panagiotopoulos. *J. Chem. Phys.*, 97:8644–8652, 1992.

[41] K. F. Lau and K. A. Dill. *Proc. Nat. Acad. Sci., U.S.A*, 97:638–642, 1990.

[42] K. A. Dill. *Biochemistry*, 24:1501, 1985.

[43] H S Chan and K A Dill. *J. Chem. Phys.*, 95:3775–3787, 1991.

[44] R. Miller, C. A. Danko, M. J. Fasolka, and A. C. Balazs. *J. Chem. Phys.*, 96:768–780, 1992.

[45] J. H. Holland. *Adaptation in Natural and Artificial Systems.* Univ. of Michigan Press, Ann Arbor, MI, 1975.

[46] R. Unger and J. Moult. *J. Mol. Biol.*, 231:75–81, 1993.

[47] S. Forrest. *Science*, 261:872–878, 1993.

[48] D. E. Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning.* Addison-Wesley, Reading, MA, 1989.

[49] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. J. Teller. *J. Chem. Phys.*, 21:1087–1092, 1953.

[50] A J Guttman, B W Ninham, and C J Thompson. *Phys. Rev.*, 172:554–558, 1968.

[51] W H Press, S A Teukolsky, B P Flannery, and W T Vetterling. *Numerical Recipes: The Art of Scientific Computing.* Cambridge University Press, Cambridge, Inglaterra, 1987.

[52] N. Gō and H. Taketomi. *Proc. Natl. Acad. Sci. USA*, 75:559–563, 1978.

[53] R. Unger and J. Moult. *J. Mol. Biol.*, 259:988–994, 1996.

[54] V. I. Abkevich, A. M. Gutin, and E. I. Shakhnovich. *J. Mol. Biol.*, 252:460–471, 1995.

[55] H. S. Chan. *Nature*, 392:761–763, 1998.

Apêndice

Nas próximas páginas encontra-se uma relação das seqüências analisadas neste trabalho. As primeiras duas tabelas contêm as seqüências de tamanho 22 e as demais contêm as seqüências de tamanho 24.

Degenerescência	Energia Mínima	Seqüência
2	−10	11010111110011000010 10
2	−10	10011110100001010011 01
2	−10	10100011111101000101 01
4	−9	00100111001010110101 01
4	−12	10011101100111111111 001
4	−10	11001100111010001011 01
4	−9	00100110010110011101 10
6	−10	11000111011111100100 10
6	−11	11001100010110111111 01
6	−10	00110111111011110000 11
8	−10	10100100011000111111 01
8	−7	01010001011010011000 10
8	−8	01100001100111001101 01
8	−8	01001100110000110000 11
8	−9	11110000111000100100 11
10	−11	10101100111110100101 01
10	−8	11100000100101101001 10
10	−10	11101111010100101000 01
10	−8	00101000010011010001 11
10	−10	10100111011010101111 00
12	−10	10110101110010011011 10
12	−10	11111100010111101001 00
12	−6	10001000110000000110 11
12	−10	00101001010011011111 01
14	−8	10011000100100000101 11
16	−7	01000100100110000010 11
16	−10	10100100101001110111 01
16	−7	11000011110001010100 01
16	−11	01100111101111111000 11
16	−12	11110011110101110010 11
16	−9	11000011110011011000 01
18	−8	10100100001011101111 10
18	−6	11001000000100001011 00
18	−11	10010010111110010101 11
20	−9	01110100011010100011 11
24	−8	00110100110011100101 00

Degenerescência	Energia Mínima	Seqüência
28	−6	00010110010010010000010
30	−9	11001110100111000010110
30	−6	01010010100010101110000
36	−7	01101010000111100011100
36	−12	11011001101101111111101
38	−11	11100111001110011110011
40	−8	10010010001100011111110
42	−10	10110111011001001011101
42	−9	11000011110111100101000
44	−10	01011110010011011111101
44	−12	11001101111111111010011
48	−9	01100011011000111101111
48	−9	11001001001001111001100
50	−10	11101011101100100100011
50	−10	00101101011110110011011
50	−11	11110011110001111001111
52	−10	00011111111001110110011
52	−10	11110010000101100111111
54	−8	01011001011011000010011
58	−8	11010101100011010110000
60	−9	11111101110000011011000
62	−9	10101100011110001110111
62	−7	00010110001011110100011
72	−9	10111111001010110000011
76	−8	11001110100110001001100
76	−6	11000100101000101011000
76	−11	01010111101011101101111
78	−7	01101010010001110001011
80	−9	10011001111001100001011
82	−9	00011010011110001101111
86	−9	10100010110011110101011
86	−7	01001101101000100000011
88	−11	11110101101011011110100
94	−10	11111110111111000011000
94	−9	11000110110010100111011

Degenerescência	Energia Mínima	Seqüência
2	−11	101111101101001010001001
2	−14	111101100110011011101111
2	−11	110100111110010000101001
2	−8	100100010010000110010110
2	−12	101001101111011001001011
2	−11	010001001111110111101001
2	−10	111101000100001001100111
2	−9	100101000000011001111001
4	−10	010011100101001110011001
4	−12	011111110101001101001011
4	−9	101000101011000110101111
4	−10	010100111011000110010011
4	−11	101101000101011101111010
4	−11	111101110100001010111010
6	−10	101010001111101101010010
6	−11	110100101100101001001011
6	−8	101100001010011000010100
8	−11	110010111000010100110111
8	−11	100100111110010011001110
8	−10	101111100111101110001010
8	−13	110101001101111010100111
10	−9	001010011101100100010110
12	−12	100110011010110001011111
12	−9	010100101010110100001001
12	−11	010110110001111001001011
14	−8	110110100110000100000110
14	−10	110011111000100010110101
16	−11	011110111101011010010100
16	−10	111000010000111101010011
18	−12	110110011101101111010101
18	−12	100110010111011111001110
18	−12	101100010111110111010111
18	−11	011011111100101010011100
20	−12	001011101011111100110101
20	−9	010001001001010111101110
22	−10	011101100100011111001010
22	−11	111011110000011110010101
24	−9	111111010000100001100001
24	−13	110110101111011111100110
24	−12	110101001110111001010111
24	−7	010000000101110111111000

Degenerescência	Energia Mínima	Seqüência
24	−9	011110001000111110010001
24	−13	011111111001011110110011
26	−8	011000001011000011100101
28	−7	011001110100000100001010
28	−13	100111111111011011001001
28	−11	110010100010111011011101
28	−6	000010011011000000010010
30	−12	111111111000011010111001
30	−10	011010011110100111010100
30	−10	010010101101101011001001
32	−10	110100001001001101110110
32	−12	111011011000011110100111
32	−8	110110001010000100100011
34	−11	101000010110111101011110
38	−10	010011101011110110001100
38	−10	111001010001010110101001
38	−10	010011010111110100011011
38	−8	011101000100110001110010
40	−7	011010000001001000001101
40	−7	011010000001001000001101
40	−9	011100110100110100100011
42	−7	001000101010110010010001
42	−9	101000011010011110001001
42	−9	111101000000100110101100
44	−9	111100100000111001010100
48	−10	100101001100001011110110
48	−7	110010001011000001000011
50	−10	001000111101001000111111
50	−10	011110010100001111110101
52	−6	001100001001001011000010
52	−9	011001000111010011011010
52	−10	110100101011001100101110
56	−10	000010101101011111001001
56	−13	111111110010011110011101
58	−8	101100001000101101100011
60	−11	100110111110110011010100
62	−9	110111010100110001000111
72	−8	111000111000110110000100
72	−10	011111100011111011001000
74	−10	011110100011110001011101
76	−9	111111011101100000110000

Degenerescência	Energia Mínima	Seqüência
78	−9	110000001101101110110010
80	−8	100110000011110010101100
80	−10	001100100011110101101011
84	−8	001100001101100001010111
86	−8	111001001110100000101100
86	−11	111110100011011101001110
94	−13	111111010111111011101010
94	−9	011101001110001011010110
96	−11	010100111011110110110110
98	−10	011110010101000110101111
104	−8	111000001100110001100100
104	−10	110010010001011101111010
112	−10	011110010000111110111100
114	−10	110010110010001101101101
120	−6	100000011010010001000001
120	−11	101001101111110100010011
124	−10	010110000110010011111011
128	−8	100001111001101010100100
130	−8	001011110001101001101000
132	−9	111001001100000011000111
138	−8	010101110000111000011010
140	−8	101100011100000101101010
142	−8	110100110100001000111000
146	−7	100101000010010001010011
152	−8	010001001001011011111000
158	−9	011001100101100011111000
164	−13	011111100111111010111011
168	−9	000001100110110101001111
182	−9	110100011001111000111000
184	−8	101000110100110100011010
184	−9	111111100100001000010011
188	−9	111100000101100110001101
190	−9	011011110010000101101100
220	−7	010001100011100011100100
220	−9	011000000011110100111101
220	−11	001011001011111001101101
234	−10	011010001010111110111011
250	−7	100110110000000000100101
278	−9	011000011111010010010011
286	−9	101001000010101101101101

"ESTUDO DE ENOVELAMENTO DE POLÍMEROS EM REDE QUADRADA BIDIMENSIONAL"

Celina Maria de Souza Costa

Tese de Mestrado apresentada no Centro Brasileiro de Pesquisas Físicas, do Conselho Nacional de Desenvolvimento Científico e Tecnológico, fazendo parte da Banca Examinadora os seguintes profesores:

Fernando de Magalhães Coutinho Vieira - Presidente

Sergio Teixeira Ferreira

Myriam Malvina Segre de Giambiagi

Léa Jaccoud El-Jaick

Rio de Janeiro, 13 de julho de 1998