UNIVERSIDADE DE SÃO PAULO
FFCLRP - DEPARTAMENTO DE FÍSICA E
MATEMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM FÍSICA APLICADA
À MEDICINA E BIOLOGIA

**“Potencial estéreo-hidrofóbico e
propriedades topológicas no
enovelamento de proteínas”.**

Maria Eulália Pinto Tarragó

Tese apresentada à Faculdade de Filosofia, Ciências e
Letras de Ribeirão Preto da USP, como parte das exigên-
cias para a obtenção do Título de Doutor em Ciências,
Área: Física Aplicada à Medicina e Biologia.

RIBEIRÃO PRETO - SP

2003

UNIVERSIDADE DE SÃO PAULO
FFCLRP - DEPARTAMENTO DE FÍSICA E
MATEMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM FÍSICA APLICADA
À MEDICINA E BIOLOGIA

**“Potencial estéreo-hidrofóbico e
propriedades topológicas no
enovelamento de proteínas”.**

Maria Eulália Pinto Tarragó

Orientador: Antonio Caliri

Tese apresentada à Faculdade de Filosofia, Ciências e
Letras de Ribeirão Preto da USP, como parte das exigên-
cias para a obtenção do Título de Doutor em Ciências,
Área: Física Aplicada à Medicina e Biologia.

RIBEIRÃO PRETO - SP

2003

À minha amada Mamita

Agradecimentos

À Pontifícia Universidade Católica do Rio Grande do Sul, PUCRS, nas pessoas dos diretores da Faculdade de Física, Prof. Antonio Dias Nunes e Prof^a Maria Emília B. Bernasiuk, pelo incentivo e apoio total para a realização deste doutorado.

Ao orientador, Prof. Antonio Caliri, pela oportunidade de compartilhar suas idéias, pelo seu brilhante exemplo, por sua dedicação, carinho, atenção, compreensão e amizade.

Ao querido amigo Vanderlei Haas, pela amizade, momentos de lazer, carinho, coleguismo, dedicação e atenta leitura deste manuscrito.

Ao querido colega e amigo Luiz Faustino Rocha, pela atenção e discussões, e, também, pela linda (e trabalhosa) figura das “Gotas de óleo em água”.

Ao caro colega Roosevelt A. da Silva, pelas discussões e pelas diversas dicas de programação.

Às queridas amigas e colegas Flávia Providello e Inês Regina da Silva, pelo carinho, atenção e dedicação.

A todos os Professores do DFM, da FFCLRP-USP, e aos Professores do DFQ, da FCFRP-USP, pela oportunidade de compartilhar.

Ao Eudereis de Paula e Ivan Pizza pela ajuda inicial na parte de programação.

Ao querido amigo Evamberto Góes, pelo seu constante incentivo.

A todos os colegas, pela alegria de partilhar esta caminhada.

Aos meus amigos ...

À minha família ...

Ao meu amado Sathya Sai Baba ...

Resumo

O entendimento dos princípios básicos do enovelamento proteico pode conduzir a muitas aplicações importantes. Embora não se conheçam todos os aspectos significativos envolvidos neste problema, experimentos e aproximações teóricas têm produzido avanços relevantes na sua compreensão, como, por exemplo, o papel dominante das forças hidrofóbicas. Com o propósito de contribuir para a identificação de alguns parâmetros determinantes no processo de **folding**, o objetivo deste trabalho consiste em estudar o potencial estéreo-hidrofóbico (isto é, o potencial hidrofóbico e um conjunto de especificidades estéricas adequadas) relacionado ao processo de enovelamento de proteínas globulares, evidenciando o papel das restrições estéricas e as características topológicas do estado nativo. Para tanto, empregou-se um modelo simplificado em rede cúbica, estudando-se, através de simulações Monte Carlo, o comportamento de mais de 40 cadeias tipo-proteínas, que apresentam estados nativos caracterizados pelos seguintes parâmetros topológicos: ordem de contato, χ , ordem de longo alcance, χ_{LR} e número ϕ de estruturas tipo grande manivela. Claramente, os resultados das simulações para as cadeias com configurações nativas, caracterizadas por baixo valor de χ e elevado valor de ϕ , são bastante diferenciados daqueles obtidos para as cadeias que apresentam configurações nativas com χ elevado e ϕ pequeno, evidenciando que o potencial estéreo-hidrofóbico, adotado neste trabalho, permite relacionar os comportamentos termodinâmico e cinético da cadeia tipo-proteína com os atributos topológicos da configuração nativa correspondente. Adicionalmente, mostrou-se que as restrições estéricas consideradas introduzem fortes mudanças na atividade configuracional, aumentando a estabilidade do estado nativo, bem como alterando, drasticamente, a curva da capacidade térmica, em função da temperatura, em comparação com os resultados das simulações obtidos com o potencial hidrofóbico (sem restrições estéricas).

Abstract

The understanding of the basic principles of the protein folding process may lead to very important applications. Although all significant aspects of this problem are not yet known, experimental and theoretical results have given important contribution on the subject, as, for instance, about the dominant role of the hydrophobic forces. In order to contribute to the identification of significant ingredients for the folding process, the main goal of this work consists in studying the stereo-hydrophobic potential (that is, the hydrophobic potential and a set of steric specificities) related to the folding process of globular proteins, revealing the importance of the steric constraints and the role of the native structure. A minimalist lattice model was employed for this purpose, and more than forty distinct protein-like chains were studied. These chains were designed based on native structures characterized by topological parameters as contact order, χ , long range order, χ_{LR} and number ϕ of crankshaft-like structures. The Monte Carlo simulation results show clearly that the folding process depends strongly on the topological attributes of the native structure: thermodynamical and kinetic behavior for chains designed from native structures, presenting smaller χ and higher ϕ , are very distinguishable from those with higher χ and lower ϕ . Additionally it was shown that the steric constraints significantly modify the configurational activity, increasing the general conditions for the globule stability, as well changing drastically the shape of the thermal capacity behavior, as a function of the temperature, in comparison with the corresponding results obtained using the hydrophobic potential only, i.e., without the steric specificities.

Conteúdo

1	INTRODUÇÃO	1
2	Proteínas Globulares	5
2.1	Proteínas Globulares, Fibrosas e de Membrana	5
2.2	O Problema do Folding de Proteínas Glo-bulares	8
2.3	Níveis de Estrutura de uma Proteína	10
2.3.1	Estrutura primária	10
2.3.2	Estrutura secundária	12
2.3.3	Estrutura terciária	18
2.4	Seqüência, Estrutura e Função	20
2.5	Interações Presentes nas Proteínas	21
2.6	Aspectos Termodinâmicos do Folding	26
2.6.1	Folding de dois estados e a constante de equilíbrio	27
2.6.2	Energia livre de Gibbs	27
2.6.3	Efeito hidrofóbico e área acessível ao solvente	30
2.6.4	Capacidade térmica	33
2.6.5	Entalpia calorimétrica e Entalpia de van't Hoff	35
2.7	Aspectos Cinéticos do Folding de Proteínas	36
2.8	Parâmetros Topológicos do Estado Nativo	40
2.8.1	Ordem de contato χ	42
2.8.2	Ordem de longo alcance χ_{LR}	42

2.8.3	Distância de contato total χ_T	43
2.8.4	Correlação entre $\ln k_f$ e Parâmetros Topológicos	44
2.9	O Potencial Estatístico Inter-Resíduos	47
2.10	Modelos Computacionais para o Folding	50
2.11	Referências Bibliográficas	59
3	O Modelo Estéreo-Hidrofóbico	67
3.1	Introdução	67
3.2	O Design da Estrutura Nativa	68
3.2.1	Topologia do estado nativo	68
3.2.2	Design da estrutura nativa	70
3.3	Energia Hidrofóbica	75
3.3.1	Modelo hidrofóbico	75
3.3.2	Equivalência entre energia hidrofóbica e energia de contato entre os monômeros	80
3.4	Especificidades Estéricas	82
3.5	Formalismo Entrópico para o Folding	84
3.5.1	Entropia e energia livre	84
3.5.2	Probabilidade de transição entre duas configurações	87
3.6	Movimentos Configuracionais Permitidos	89
3.7	Simulações Monte Carlo	90
3.7.1	Atividade configuracional	91
3.7.2	Cálculo da capacidade térmica	92
3.7.3	Distância ponta-a-ponta e raio de giração	93
3.8	Referências Bibliográficas	95
4	Resultados e Discussão	97
4.1	Distribuição das CSA Segundo Parâmetros Topológicos	97
4.2	Estudo Termodinâmico do Modelo Estéreo-hidrofóbico	101

4.2.1	Resultados para a cadeia 870	101
4.2.2	Resultados para a cadeia 99865	106
4.2.3	Influência das condições iniciais	109
4.2.4	Resultados para as cadeias de Goldstein e de Li	112
4.2.5	Correlação entre quantidades termodinâmicas e pa-râmetros topológicos	115
4.2.6	Discussão dos aspectos termodinâmicos dos resultados . .	119
4.3	Cinética do modelo estéreo-hidrofóbico	121
4.3.1	Resultados para a cadeia 870	122
4.3.2	Resultados para a cadeia 99865	128
4.3.3	Comparando o comportamento cinético de duas ca-deias com mesma ordem de contato	133
4.3.4	Estudo do tempo de enovelamento para 600 cadeias tipo- proteína	135
4.3.5	Estudo do tempo de enovelamento de diversas ca-deias tipo-proteínas	137
4.3.6	Discussão geral dos aspectos cinéticos dos resultados . . .	138
4.4	Importância das Restrições Estéricas	141
4.4.1	Estudo termodinâmico da cadeia 870	141
4.4.2	Estudo cinético da cadeia 5365	144
4.4.3	Discussão do papel das restrições estéricas	148
4.5	Referências Bibliográficas	152

Lista de Figuras

2.1	Representação de uma proteína fibrosa	6
2.2	Representação de uma proteína globular	7
2.3	Representação de alguns tipos de proteínas de membrana	8
2.4	Representação de aminoácidos	11
2.5	Representação de uma hélice- α	13
2.6	Representação de uma folha- β mista	15
2.7	Cadeia polipeptídica e os ângulos diedros (ϕ, ψ)	16
2.8	Mapa de Ramachandran	17
2.9	Potencial de Lennard-Jones	23
2.10	Gotas de óleo em água	31
2.11	Área acessível ao solvente	32
2.12	Variação da área exposta ao solvente, em função da variação da energia livre	33
2.13	Capacidade térmica, à pressão constante, em função da temperatura	35
2.14	Intensidade de fluorescência, em função do tempo	39
2.15	Logaritmo da taxa de folding, em função da concentração final de desnaturante	40
2.16	Logaritmo natural da taxa de folding, em função dos parâmetros topológicos	46
2.17	Representação da configuração nativa, nos trabalhos do grupo de Go	51

2.18	Distribuição das 103446 CSA, em função do parâmetro de otimização $\mathcal{R}$	57
2.19	Estrutura CSA mais “desenhável ” do grupo de Li	58
3.1	Configuração CSA 870	71
3.2	Configuração CSA 99865	72
3.3	Representação de uma CSA, indicando-se as posições das quatro classes de monômeros	73
3.4	Representação esquemática dos contatos que os monômeros p , q e t estabelecem com e sem o resíduo k na cadeia	78
3.5	Representação esquemática dos contatos que os monômeros u , v e z estabelecem com o resíduo k na nova configuração	79
3.6	Representação esquemática do conjunto das restrições estéricas	83
3.7	Movimentos configuracionais permitidos	89
4.1	Distribuição das 103346 configurações CSA, em função da ordem de contato	98
4.2	Distribuição das 103346 configurações CSA, em função da ordem de longo alcance	99
4.3	Distribuição das 103346 configurações CSA, em função do número de estruturas $STTS$	99
4.4	Valor máximo da ordem de contato que ocorre para cada número ϕ de estruturas $STTS$	100
4.5	Ordem de longo alcance, em função da ordem de contato, para as 103346 configurações CSA	100
4.6	Capacidade térmica, em função da temperatura, para a cadeia 870	101
4.7	Número de contatos topológicos, em função da temperatura, para a cadeia 870	103
4.8	Frequência na configuração nativa, em função da temperatura, para a cadeia 870	103

4.9	Energia configuracional para a cadeia 870, em função do tempo, em $T = 0,83$	104
4.10	Energia configuracional para a cadeia 870, em função do tempo, em $T = 0,93$	105
4.11	Energia configuracional para a cadeia 870, em função do tempo, em $T = 1,53$	105
4.12	Capacidade térmica, em função da temperatura, para a cadeia 99865	106
4.13	Números de contatos topológicos, em função da temperatura, para a cadeia 99865	107
4.14	Frequência na configuração nativa, em função da temperatura, para a cadeia 99865	107
4.15	Distância ponta-a-ponta e raio de giração para as cadeias 870 e 99865	108
4.16	Capacidade térmica, em função da temperatura, usando-se doze condições iniciais diferentes	110
4.17	Número de contatos topológicos, em função da temperatura, usando- se doze condições iniciais diferentes	110
4.18	Frequência na configuração nativa, em função da temperatura, usando-se doze condições iniciais diferentes	111
4.19	Capacidade térmica, em função da temperatura, para duas con- figurações iniciais diferentes	111
4.20	CSA de Goldstein e CSA de Li	112
4.21	Capacidade térmica, partindo-se de configurações aleatórias, para as cadeias de Goldstein e de Li	114
4.22	Atividade configuracional, partindo-se de configurações aleatórias, para as cadeias de Goldstein e de Li	114

4.23	Tempo de folding, partindo-se de configurações aleatórias, para as cadeias de Goldstein e de Li	115
4.24	Energia da configuração nativa, em função da ordem de contato e da ordem de longo alcance, para mais de 40 CSA estudadas . . .	116
4.25	Valor máximo da capacidade térmica e “largura à meia altura”, em função da ordem de contato, para mais de 40 CSA estudadas	117
4.26	Ordem de longo alcance, em função da ordem de contato, para o conjunto das CSA estudadas	117
4.27	Temperatura onde o número relativo médio do total de contatos é 0,5, em função da temperatura T_{max}	118
4.28	Temperatura T_{max} , em função da ordem de contato, para mais de 40 CSA estudadas	118
4.29	Tempo de folding, em função da temperatura, T , para a cadeia 870123	
4.30	Frequência relativa na configuração nativa, para a cadeia 870, no processo de enovelamento	124
4.31	Distribuição da energia, em $T = 1, 0$, para a cadeia 870	124
4.32	Distribuição da energia, em $T = 2, 4$, para a cadeia 870	125
4.33	Evolução da energia para a cadeia 870, em $T = 0, 43$, no processo de enovelamento	126
4.34	Evolução da energia para a cadeia 870, em $T = 0, 63$, no processo de enovelamento	126
4.35	Evolução da energia para a cadeia 870, em $T = 1, 0$, no processo de enovelamento	127
4.36	Evolução da energia para a cadeia 870, em $T = 1, 5$, no processo de enovelamento	127
4.37	Evolução da energia para a cadeia 870, em $T = 3, 0$, no processo de enovelamento	128

4.38 Tempo de folding considerando sete condições iniciais diferentes, para a cadeia 99865	129
4.39 Frequência na configuração nativa, para sete condições iniciais diferentes, para a cadeia 99865	130
4.40 Frequência com que a cadeia 99865 visita configurações CSA, em $T = 1, 0$	130
4.41 Distribuição da energia para a cadeia 99865, em $T = 1, 2$	131
4.42 Distribuição da energia para a cadeia 99865, em $T = 1, 8$	131
4.43 Evolução da energia, para a cadeia 99865, em $T = 1, 2$, no processo de enovelamento	132
4.44 Evolução da energia, para a cadeia 99865, em $T = 1, 8$, no processo de enovelamento	132
4.45 Tempo de folding para duas cadeias com CSA nativas caracteri- zadas pela mesma ordem de contato e número de estruturas <i>STTS</i> diferentes	134
4.46 Frequência na configuração nativa para duas cadeias com CSA nativas caracterizadas pela mesma ordem de contato e número de estruturas <i>STTS</i> diferentes	134
4.47 Frequência com que as cadeia visitam configurações CSA, para duas cadeias com CSA nativas caracterizadas pela mesma ordem de contato e número de estruturas <i>STTS</i> diferentes	135
4.48 Decaimento do número de cadeias desnaturadas, em função do tempo, no processo de enovelamento	136
4.49 Tempo de folding para diversas cadeias tipo-proteínas, em função da ordem de contato	138
4.50 Tempo de folding para diversas cadeias tipo-proteínas, em função da ordem de longo alcance	139

4.51	Tempo de folding para diversas cadeias tipo-proteínas, em função do número de estruturas <i>STTS</i>	139
4.52	Capacidade térmica, em função da temperatura, com os potenciais estéreo-hidrofóbico e hidrofóbico	142
4.53	Atividade configuracional, em função da temperatura, com os potenciais estéreo-hidrofóbico e hidrofóbico	142
4.54	Distância ponta-a-ponta e raio de giração com os potenciais estéreo-hidrofóbico e hidrofóbico	143
4.55	Números de contatos total e nativos, em $T = 0,63$, usando-se o potencial hidrofóbico	143
4.56	Tempo de folding, para 30 condições iniciais diferentes, usando-se o potencial estéreo-hidrofóbico	145
4.57	Tempo de folding, para 15 condições iniciais diferentes, usando-se o potencial hidrofóbico	145
4.58	Frequência relativa com que a cadeia visita a configuração nativa, para os potenciais estéreo-hidrofóbico e o hidrofóbico	146
4.59	Número de vezes com que a cadeia visita uma configuração CSA, para os potenciais estéreo-hidrofóbico e hidrofóbico	146
4.60	Energia configuracional, em função do tempo, em $T = 1,03$, com o potencial estéreo-hidrofóbico	147
4.61	Energia configuracional, em função do tempo, em $T = 1,03$, com o potencial hidrofóbico	147

Lista de Tabelas

Lista de Símbolos e Abreviaturas

C/k : capacidade térmica (adimensional)

C_{max}/k : valor máximo da capacidade térmica

CSA : configuração maximamente compacta (Compact Self-Avoiding)

E : energia da cadeia em uma dada configuração

$E^{(n)}$: energia da cadeia na configuração nativa

E_{ij} : energia de contato do par ij

f : frequência relativa com que a cadeia visita a configuração nativa

f_S : frequência relativa com que a cadeia visita a configuração nativa, considerando-se o potencial hidrofóbico (sem restrições estéricas)

f_R : frequência relativa com que a cadeia visita a configuração nativa, considerando-se o potencial estereo-hidrofóbico (com restrições estéricas)

h_i : energia hidrofóbica do monômero i

k : constante de Boltzmann

kT : energia térmica

k_f : taxa de folding

T : temperatura reduzida

T' : temperatura onde os resultados das simulações de desnaturação são dependentes das condições iniciais

T_{max} : temperatura onde ocorre o valor máximo da capacidade térmica

χ : ordem de contato

χ_{LR} : ordem de longo alcance

ϕ : número de estruturas secundária tipo $STTS$

γ_i : nível de hidrofobicidade do monômero i

Ψ : número relativo médio do total de contatos

Ψ_S : número relativo médio do total de contatos, considerando-se o potencial hidrofóbico (sem restrições estéricas)

Ψ_R : número relativo médio do total de contatos, considerando-se o potencial estereo-hidrofóbico (com restrições estéricas)

$\Psi^{(n)}$: número relativo médio dos contatos nativos

$\Psi_S^{(n)}$: número relativo médio dos contatos nativos, considerando-se o potencial hidrofóbico (sem restrições estéricas)

$\Psi_R^{(n)}$: número relativo médio dos contatos nativos, considerando-se o potencial estereo-hidrofóbico (com restrições estéricas)

Capítulo 1

INTRODUÇÃO

O enovelamento ou **folding** de proteínas globulares é considerado um dos problemas biológicos mais complexos que podem ser descrito em termos puramente atômicos e constitui, no momento, um campo muito ativo de pesquisa, envolvendo diversas áreas, como a física, biologia, química, bioquímica e ciências da computação. A questão central do **folding** consiste em prever qual será a estrutura tridimensional bioativa da proteína globular, conhecida como estrutura nativa, a partir de sua seqüência de aminoácidos. O entendimento de seus princípios fundamentais pode conduzir a muitas aplicações práticas, como nas pesquisas que seguirão o projeto genoma, na compreensão de diferentes patologias e no desenho de novas proteínas com funções especiais.

Acredita-se que a complexidade estrutural do estado nativo das proteínas globulares possa ser capturada, de alguma forma, através de uma equação matemática. Neste sentido, recentemente, alguns parâmetros topológicos foram identificados e propostos na literatura, possibilitando a idealização de diversos estudos experimentais. Embora não se conheçam todos os aspectos importantes envolvidos no processo de enovelamento proteico, experimentos e aproximações teóricas têm produzido avanços significativos no seu entendimento.

Os modelos minimalistas, apesar de suas simplicidades, são capazes de cap-

turar alguns aspectos fundamentais do processo de *folding*, sendo a sua principal motivação o papel peculiar das forças hidrofóbicas. A constatação de que a energia hidrofóbica gera a força dominante no processo de enovelamento, permite considerar somente as interações cadeia-solvente e desprezar todos os outros tipos de interações (cadeia-cadeia, solvente-solvente). Esta simplificação é vantajosa, particularmente nas simulações Monte Carlo, pois, assim, pode-se usar diretamente os valores das medidas das energias livres de transferência na probabilidade de transição entre duas configurações. No entanto, as simulações Monte Carlo revelam que um potencial puramente hidrofóbico, apesar de conduzir a cadeia eventualmente à configuração nativa, não garante estabilidade configuracional. Adota-se, então, neste trabalho, o potencial estéreo-hidrofóbico, que considera um conjunto de restrições estéricas juntamente com a energia hidrofóbica, introduzido anteriormente na literatura por da Silva e colaboradores (2001)¹.

Neste estudo, usa-se um modelo computacional simplificado para o enovelamento proteico, no qual a cadeia tipo-proteína é representada por “um colar de 27 contas” em um espaço configuracional do tipo rede cúbica. Em geral, o estado nativo de uma proteína globular, nos modelos computacionais, é representado por uma estrutura maximamente compacta. Existem 103346 maneiras diferentes de dobrar-se, na forma maximamente compacta, uma cadeia de 27 contas, em uma rede cúbica, resultando em estruturas do tipo Compact Self-Avoiding (CSA), que desenham um cubo $3 \times 3 \times 3$.

Com o propósito de contribuir para a identificação de parâmetros determinantes no processo de *folding*, este trabalho estudou o potencial estéreo-hidrofóbico relacionado ao processo de enovelamento de proteínas globulares, evidenciando o papel das restrições estéricas e as características topológicas do estado nativo. Especificamente, procurou-se: (i) levantar as distribuições das 103346

¹Veja-se o artigo “Deterministic Folding: The Role of Entropic Forces and Steric Specificities”, de da Silva e colaboradores, em J. Chem. Phys.(2001) **114**: 4235-4242.

configurações maximamente compactas em relação aos parâmetros topológicos de interesse; (ii) estudar a influência das condições iniciais nas simulações de enovelamento e de desnaturação; (iii) verificar as correlações dos parâmetros topológicos do estado nativo com as propriedades cinéticas do enovelamento; (iv) verificar as correlações dos parâmetros topológicos do estado nativo com as propriedades termodinâmicas das cadeias tipo-proteínas; (v) estudar o papel das restrições estéricas na cinética e na termodinâmica do enovelamento.

No Capítulo 2, esboça-se um panorama das diversas propriedades das proteínas, e, em particular, das proteínas globulares, tais como as propriedades estruturais, topológicas, termodinâmicas e cinéticas; comenta-se também sobre as principais características dos modelos computacionais simplificados usados no estudo do processo de enovelamento de cadeias tipo-proteínas. Introduz-se, no Capítulo 3, os critérios de classificação dos monômeros que constituem a cadeia tipo-proteína, cuja sintaxe, desenvolvida anteriormente por da Silva e colaboradores (2001) (vide nota de rodapé 1), está baseada em considerações geométricas e na regra “mais (menos) hidrofóbico dentro (fora)”. Além disso, apresenta-se o formalismo que constitui o embasamento teórico para o potencial estéreo-hidrofóbico, demonstrando-se a equivalência, para o caso particular em questão, entre o potencial hidrofóbico (relativo à interação solvente-água) e a energia de contato intermonômero.

O Capítulo 4 apresenta e discute os resultados das simulações Monte Carlo de folding e de desnaturação, envolvendo mais de 40 cadeias tipo-proteínas com estruturas nativas com os mais diversos parâmetros topológicos, a saber: ordem de contato, ordem de longo alcance e número de estruturas tipo manivela grande, chamadas estruturas *STTS*. As simulações revelam, por exemplo, que tanto as restrições estéricas consideradas como os atributos topológicos da configuração nativa introduzem mudanças na atividade configuracional, influenciando a estabilidade do estado nativo e alterando a curva da capacidade térmica, em função

da temperatura².

Conclui-se, no Capítulo 5, que as restrições estéricas em conjunto com a energia hidrofóbica, constituindo o potencial estéreo-hidrofóbico, são dois ingredientes fundamentais no processo de enovelamento proteico, capazes de conduzir a cadeia à conformação nativa e prover estabilidade suficiente neste estado, dentro de um certo intervalo de temperatura (vide nota de rodapé 2). Claramente, os resultados das simulações, para as cadeias com configurações nativas caracterizadas por baixo valor de ordem de contato, χ , e elevado valor de número ϕ de estruturas *STTS* são bastante diferenciados daqueles obtidos para as cadeias com configurações nativas com χ elevada e ϕ pequeno, evidenciando que o potencial estéreo-hidrofóbico, adotado neste trabalho, permite relacionar os comportamentos termodinâmico e cinético da cadeia tipo-proteína com os atributos topológicos da sua configuração nativa.

²Veja-se o artigo “Steric Constraints as Folding Coadjuvant”, de Tarragó e colaboradores, em Phys. Rev. E (2003), no prelo.

Capítulo 2

Proteínas Globulares: uma visão geral

2.1 Proteínas Globulares, Fibrosas e de Membrana

As **proteínas** são consideradas macromoléculas essenciais à vida, executando uma diversidade de funções, desde que estejam na sua estrutura tridimensional específica e única, conhecida como conformação nativa. Do ponto de vista químico, as proteínas são seqüências lineares de aminoácidos ligados covalentemente, constituindo cadeias polipeptídicas. Todas as proteínas que ocorrem naturalmente, isto é, que são codificadas geneticamente pelos organismos vivos, podem conter até vinte tipos diferentes de aminoácidos nas suas cadeias, os chamados aminoácidos naturais; por esta razão, elas também são classificadas como heteropolímeros. Existem também proteínas que são funcionalmente ativas quando sua cadeia polipeptídica está ligada a um grupo prostético —um átomo ou grupoamento atômico que não é do tipo aminoácido. Alguns aminoácidos são excelentes ligantes para metais, permitindo às proteínas, conhecidas como metaloproteínas,

Figura 2.1: Representação de um segmento de colágeno. Esta proteína fibrosa tem a forma de longos cordões, podendo atingir 3000 Å de comprimento por 15 Å de diâmetro (Schulz & Schirmer, 1990, p.72, Figura 5.6-c).

incorporarem metais em sua estrutura, tais como ferro, zinco, magnésio e cálcio (Branden & Tooze, 1999); as proteínas que incorporam o grupo prostético hemo são chamadas hemoproteínas; as que incorporam lipídios, são as lipoproteínas e as que incorporam carboidratos, as glicoproteínas, entre outras (Goldberg, 1998; Garrett & Grisham, 1995).

As proteínas podem ser classificadas como fibrosas, globulares e de membrana, de acordo com o tipo de função que executam. As **proteínas fibrosas** geralmente se apresentam na forma de longos cordões e têm funções estruturais; constituem os tendões, cartilagens, ossos, pele, cabelos, unhas, chifres, penas, lã, etc. O colágeno, representado na **Figura 2.1**, é uma das proteínas fibrosas mais abundantes e pode atingir 3000Å de comprimento, por 15Å de diâmetro (Richard et al., 1997).

As **proteínas globulares** são solúveis em água e, como o próprio nome indica, apresentam-se na forma de glóbulos. Tal característica estrutural resulta

Figura 2.2: Representação da molécula de hemoglobina, uma das proteínas globulares mais estudadas. Ela apresenta a forma de um esferóide com dimensões $50 \times 55 \times 65 \text{ \AA}^3$, aproximadamente (Branden & Tooze, 1999, p. 44, Figura 3.13).

da interação dos monômeros com a água, que conduz as partes hidrofóbicas dos aminoácidos (aquelas que não têm afinidade pela água) para o interior da proteína, minimizando o contato com o solvente aquoso (Kauzmann, 1959; Baldwin, 2002). As proteínas globulares são as biomoléculas mais funcionalmente diversas que se conhecem: apresentam-se como enzimas, hormônios, agentes transportadores de moléculas, etc. Elas atuam desde os processos metabólicos, na biossíntese de nucleotídeos, replicação e transcrição do DNA, até processos celulares e regulatórios (Andrade et al., 1999). A **Figura 2.2** mostra uma das proteínas globulares mais estudadas, a molécula de hemoglobina.

As proteínas de membrana apresentam suas cadeias polipeptídicas atravessando a membrana celular (uma ou diversas vezes) ou parcialmente aderidas nesta, como ilustra a **Figura 2.3**. As proteínas de membrana são, de certa forma, parecidas com as proteínas globulares no sentido de como interagem com o meio; a membrana celular é uma camada bilipídica, sendo sua parte interna hidrofóbica e suas duas interfaces, em contato com a água, polares. Estas duas regiões, com afinidades pela água bem diferenciadas, “induzem” a proteína de

Figura 2.3: Representação de quatro tipos de proteínas de membrana. Note-se que a cadeia polipeptídica pode atravessar uma ou muitas vezes a membrana celular, ou pode estar apenas aderida parcialmente nesta (Branden & Tooze, 1999, p.223, Figura 12.1).

membrana a apresentar, nos casos mais simples, no mínimo, duas regiões de hidrofobicidades bem distintas: a parte da proteína que está inserida na membrana é, na sua maioria, hidrofóbica e os aminoácidos que constituem os loops e terminações são, preferencialmente, polares (Branden & Tooze, 1999). Elas podem atuar como canais que permitem a troca de metabólitos e íons entre os meios intra e extracelular ou entre o interior de uma organela e o meio intracelular.

2.2 O Problema do Folding de Proteínas Globulares

Sob a perspectiva deste trabalho, uma proteína é uma biomolécula que apresenta estrutura tridimensional específica, conhecida como **estrutura nativa**, que é determinada unicamente pela sequência de aminoácidos da sua cadeia e pelas condições físico-químicas do meio solvente (Anfinsen, 1973; Micheletti et al., 1999; Privalov, 1997). O problema central do **enovelamento** ou folding consiste em prever qual será a estrutura tridimensional de uma proteína, a partir da sua sequência de aminoácidos (Taketomi et al., 1975). O folding é consid-

erado a segunda etapa do código genético (Dill et al., 1995; Groß, 1996), pois a estrutura de uma proteína já está determinada no gene que a codifica, uma vez que cada gene dita a sequência de aminoácidos de uma proteína (Lattman & Rose, 1993).

As mais sofisticadas técnicas de predição de estrutura de proteínas a partir da sequência —como modelagem comparativa, reconhecimento de enovelamento, ou o método **de novo** de predição de estrutura— conseguem, na maioria das vezes, prever corretamente a topologia da estrutura nativa, mas não conseguem prever detalhes estruturais importantes, como os **loops** (estruturas secundárias irregulares) e certos empacotamentos (Schonbrun et al., 2002); desta forma, o **folding** ainda é um problema em aberto (Berg et al., 2000).

Anfinsen propôs que a estrutura tridimensional nativa de uma proteína corresponde ao mínimo global de energia (Anfinsen, 1973); em outras palavras, ele propôs que a estrutura nativa está determinada termodinamicamente, pois se sabe que a cadeia polipeptídica abandona (ou retorna) facilmente à conformação nativa se as condições do meio solvente se afastam (ou se aproximam) das fisiológicas. Hoje se considera que a forma bioativa pode não corresponder ao mínimo global de energia, podendo se acomodar também em um mínimo local (Berg et al., 2000). Recentemente, estudos computacionais e experimentais têm indicado que os caminhos e as taxas de enovelamento são fortemente determinados pela topologia do estado nativo (Baker, 2000; Ferrara & Caffisch, 2001). Isto ilustra que o problema do **folding** de proteínas abrange não somente a predição da estrutura nativa, a partir da sequência de aminoácidos, mas também a identificação de parâmetros que determinam o processo do enovelamento (Bonneau et al., 2002). Este último é um dos enfoques deste trabalho, em que se identificam parâmetros topológicos do estado nativo de cadeias tipo-proteínas e se estuda a relação destes parâmetros com propriedades cinéticas e termodinâmicas, obtidas de simulações computacionais.

2.3 Níveis de Estrutura de uma Proteína

Costuma-se atribuir diversos níveis de organização estrutural para as proteínas globulares. Nesta seção, serão abordados, de maneira explícita, apenas três níveis: o primário, o secundário e o terciário. O primeiro está relacionado com a sequência de aminoácidos; o segundo, com o arranjo local do “esqueleto” da cadeia polipeptídica; e, o terceiro, com a estrutura tridimensional da molécula. Existem proteínas que exibem estrutura quaternária, que são associações (através de ligações não-covalentes) de estruturas terciárias, formando dímeros —como diversos hormônios de mamíferos que são dímeros de glicoproteínas (Berg et al., 2000)—, trímeros (Schulz & Schirmer, 1990), etc.

2.3.1 Estrutura primária

Uma proteína pode ser visualizada como uma sequência linear de aminoácidos conectados através de ligações covalentes; esta é a sua descrição mais elementar e é conhecida como **estrutura primária**. Cada componente da sequência é um dos vinte aminoácidos que ocorrem naturalmente nas proteínas; os aminoácidos em proteínas também são chamados **resíduos** ou **monômeros**. Todo aminoácido apresenta um átomo de carbono “central”, carbono- α , C_α , ligado a um grupo carboxila, $COOH$, a um grupo amino, NH_2 , a um átomo de hidrogênio, H , e a um radical, R , que representa a cadeia lateral do aminoácido (**Figura 2.4 (a)**). Quando a sequência de uma proteína está sendo montada no meio intracelular, ao ocorrer a união covalente entre dois aminoácidos, também chamada **ligação peptídica** —através de seus terminais amino e carboxila—, ocorre a perda de uma molécula de água; na **Figura 2.4**, observa-se em **(a)**, que o aminoácido isolado apresenta uma hidroxila, OH , e um átomo de hidrogênio, H , a mais do que os aminoácidos da proteína, representados em **(b)**. Devido às ligações peptídicas entre os aminoácidos da sequência, a estrutura primária de uma proteína também é chamada de **cadeia polipeptídica**, a qual apresenta sempre duas

Figura 2.4: Representação esquemática de aminoácidos. Em (a), mostra-se um aminoácido isolado, onde se observa o átomo de carbono- α , ligado ao grupo carboxila, ao grupo amina, a um átomo de hidrogênio e a um radical, R , que representa a cadeia lateral do aminoácido. Em (b), representa-se a cadeia polipeptídica, indicando-se a ligação peptídica entre os aminoácidos n e $n+1$ (adaptado de Branden & Tooze, 1999, p.4, Figura 1.2).

terminações: o terminal C , contendo o grupo carboxila ($COOH$), e o terminal N , contendo o grupo amina (NH_2).

As seqüências de resíduos de aminoácidos das proteínas podem ser dos mais diversos comprimentos; as proteínas que apresentam seqüências muito pequenas (com menos de aproximadamente 50 resíduos) costumam ser chamadas **peptídeos**. Uma das menores seqüências de aminoácidos com função biológica é a de um peptídeo neurocerebral que apresenta somente cinco resíduos: Tyr-Gly-Gly-Phe-Met (Okamoto, 1998) —onde Tyr, Gly, Phe e Met são os símbolos de três letras para os aminoácidos tirosina, glicina, fenilalanina e metionina, respectivamente. Também se encontram seqüências com dezenas de milhares de resíduos,

sendo o comprimento médio das proteínas conhecidas estimado em torno de 367 resíduos (Thornton et al., 1999).

É inimaginavelmente grande (exatamente 20^N) o número de seqüências de comprimento N que se poderia, a princípio, escrever com os 20 aminoácidos naturais. Mesmo num modelo, como o utilizado neste trabalho, onde se tem um “alfabeto” reduzido —somente de 10 letras— e uma cadeia tipo-proteína pequena, com apenas 27 “aminoácidos”, já é impossível registrar o vasto espaço das seqüências possíveis (10^{27}). Diante de números tão grandes permanece a questão: qual o critério que tem levado a natureza a selecionar as seqüências das proteínas?

2.3.2 Estrutura secundária

A estrutura tridimensional final da maioria das proteínas globulares pode ser, grosso modo, descrita como o arranjo de **estruturas secundárias regulares**, tipo hélices- α e/ou folhas- β , conectadas entre si por **estruturas secundárias irregulares** chamadas loops ou turns.

No contexto do folding de proteínas diz-se que existe **contato** ou **interação de curto alcance** entre dois aminoácidos quando estes, além de estarem próximos espacialmente (e, conseqüentemente, interagindo) estão separados ao longo da seqüência por poucas unidades de monômeros (Taketomi et al., 1975). Desta forma, as **hélices- α** , abundantes em proteínas globulares, são as estruturas secundárias que apresentam as interações de curto alcance da forma mais regular que se conhece: cada aminoácido i interage com o aminoácido $i + 4$ através de uma **ligação de hidrogênio** (ver seção 2.5), havendo apenas três aminoácidos entre eles ao longo da cadeia.

O efeito cooperativo das ligações de hidrogênio tem papel importante na estabilização da hélice- α . Na **Figura 2.5** observam-se ligações de hidrogênio entre o oxigênio do carbono do grupo carboxila (CO) do i -ésimo resíduo e o

Figura 2.5: Representação de uma hélice- α . As ligações de hidrogênio, nas hélice- α , ocorrem sempre entre o oxigênio ligado ao carbono do grupo carboxila (C'), do i -ésimo resíduo, e o hidrogênio ligado ao nitrogênio pertencente ao grupo amino, do resíduo $(i + 4)$ (Branden & Tooze, 1999, figura 2.2-b).

hidrogênio ligado ao nitrogênio do grupo amino (NH) do resíduo $i + 4$. Estas interações só não ocorrem para os quatro primeiros grupos aminos e os quatro últimos grupos carboxilas da cadeia, fazendo com que as extremidades das hélices- α sejam geometricamente distorcidas e polares, e localizando-as, com frequência, na superfície da molécula (Baldwin & Rose, 1999; Branden & Tooze, 1999). Encontram-se hélices- α de diversos tamanhos nas proteínas globulares, desde cinco resíduos até mais de 40; o tamanho médio é de 10 resíduos, correspondendo a aproximadamente três voltas e a um comprimento, ao longo do eixo, de 15 Å (Schulz & Schirmer, 1990; Branden & Tooze, 1999).

As **folhas- β** são associações lado a lado de diversas partes da cadeia polipeptídica extendidas em forma de fitas e unidas através de pontes de hidrogênio entre os grupos aminos e carboxilas, formando uma estrutura tipo folha. Se as fitas apresentam as cadeias com a mesma orientação diz-se que as folhas- β são paralelas; caso contrário, são antiparalelas. A **Figura 2.6** mostra uma folha- β mista; da primeira à terceira fita pode-se dizer que a folha- β é antiparalela, e da terceira à quinta, a folha- β é paralela. Cada fita apresenta em média de cinco a dez resíduos de comprimento e uma fita está conectada à outra através de um loop (Branden & Tooze, 1999).

Nos estudos concernentes às estruturas proteicas define-se **contato de longo alcance** entre dois aminoácidos interagentes quando estes estão separados, ao longo da sequência, por várias unidades de monômeros (doze ou mais, conforme Gromiha & Selvaraj, 2001). Observa-se, então, que as estruturas secundárias tipo folhas- β , principalmente as paralelas, apresentam mais contatos de longo alcance do que de curto alcance.

As estruturas tipo hélices- α e folhas- β podem ser descritas pelo conjunto dos **ângulos diedros** (ϕ, ψ) definidos em relação às rotações em torno das ligações covalentes entre pares de átomos específicos do “**esqueleto**” da cadeia polipeptídica ou backbone; como ilustrado na **Figura 2.7**, o ângulo ϕ é definido em

Figura 2.6: Estrutura secundária do tipo folha- β mista. Da primeira à terceira fita, a folha- β é antiparalela, e da terceira à quinta, a folha- β é paralela (Branden & Tooze, 1999, figura 2.7-b).

relação à rotação em torno do eixo que liga o carbono- α e o nitrogênio do grupo amino ($C_\alpha - N$); e o ângulo ψ em relação à rotação em torno do eixo que liga o carbono- α e carbono do grupo carboxila ($C_\alpha - C'$). Em modelos computacionais, onde a cadeia polipeptídica é representada pelos principais átomos do backbone, é freqüente a utilização dos ângulos diedros (ϕ, ψ) nos potenciais que guiam as simulações.

O mapa de Ramachandran, representado na **Figura 2.8** e chamado assim em homenagem ao primeiro cientista que calculou as regiões (ϕ, ψ) geometricamente permitidas, mostra os pontos obtidos para todos os resíduos —com exceção da glicina (Gly) e da prolina (Pro)— de 310 estruturas de proteínas conhecidas, que se concentram nas três regiões mais escuras. A maior região escura, centralizada, em $\phi \simeq -100$ e $\psi \simeq +140$, corresponde aos resíduos que estão nos segmentos destendidos da cadeia e nas estruturas tipo folhas- β ; as outras duas regiões escuras, centralizadas, aproximadamente, em $(-70; -30)$ e $(+45; +30)$,

Figura 2.7: Cadeia polipeptídica mostrando os ângulos diedros (ϕ, ψ). O ângulo ϕ indica a rotação em torno da ligação entre o carbono- α e o nitrogênio do grupo amino ($C_{\alpha}-N$); o ângulo ψ indica a rotação em torno da ligação entre o carbono- α e carbono do grupo carboxila ($C_{\alpha}-C'$) (adaptado de Richard et al., 1997, p. 43, figura 2.24).

Figura 2.8: Mapa de Ramachandran. Note-se as regiões dos ângulos diedros, (ϕ, ψ) , geometricamente permitidas. A maior região escura, centralizada, aproximadamente, em $(-100, +140)$, corresponde aos resíduos em folhas- β e em segmentos destendidos da cadeia; as outras duas regiões escuras, centralizadas, aproximadamente, em $(-70; -30)$ e $(+45; +30)$, correspondem aos resíduos em hélices- α (Thornton, J. M. (1992), p.60, figura 2.1).

correspondem aos resíduos em hélices- α .

Os loops são estruturas irregulares abundantes, chegando a abranger 1/4 do total dos resíduos das proteínas globulares, e apresentam-se como segmentos da cadeia, de vários comprimentos e formas, conectando estruturas secundárias regulares entre si; os loops de comprimento menores também são chamados turns. Nos loops e turns, os grupos aminos e carboxilas do backbone geralmente não estão unidos entre si por ligações de hidrogênio; isto lhes permite criar pontes de hidrogênio com o solvente, uma vez que esses segmentos estão quase sempre na superfície da molécula. É importante notar que os sítios ativos das proteínas globulares localizam-se, quase que exclusivamente, nessas partes irregulares da cadeia. É interessante o fato de que a estrutura central de todos os anticorpos é muito semelhante; o que diferencia um anticorpo de outro são seus sítios ligantes

e estes são construídos por seis **loops** que variam tanto no tamanho quanto na sequência de aminoácidos. Também é nas regiões dos **loops** que ocorre a maior parte das mutações (Branden & Tooze, 1999).

2.3.3 Estrutura terciária

A estrutura tridimensional da cadeia polipeptídica, que caracteriza seu estado nativo, é também chamada **estrutura terciária**. Em 1957, o grupo de Kendrew resolveu a primeira estrutura tridimensional de uma proteína globular, a da mioglobulina. Dois anos após, foi resolvida a segunda estrutura tridimensional proteica; hoje em dia, já se conhece cerca de 20 mil estruturas de proteínas. A estrutura completa de uma proteína é usualmente documentada através da lista de coordenadas de todos os seus átomos. No Protein Data Bank (PDB) (<http://rutgers.rcsb.org/pdb/>) estão depositadas e editadas todas as listas das coordenadas atômicas das estruturas proteicas resolvidas.

Embora no presente já se conheçam dezenas de milhares de estruturas proteicas, a leitura do código genético tem indicado que muito mais sequências de proteínas ainda não têm suas estruturas tridimensionais resolvidas. Este é um dos apelos mais fortes para que se intensifiquem os esforços na solução do problema do **folding** de proteínas. A principal técnica de resolução da estrutura terciária é a cristalografia por difração de raios X; ela é bastante demorada e a maior dificuldade de sua aplicação é a obtenção dos cristais de proteínas; esta técnica envolve diversas etapas e é praticamente a mesma desde 1957 (Branden & Tooze, 1999). É possível determinar estruturas de proteínas pequenas em solução através da ressonância magnética nuclear (RMN), que consegue localizar as posições dos átomos que apresentam momento magnético nuclear, no caso, o hidrogênio. Para determinar as estruturas de proteínas maiores por RMN fazem-se algumas modificações, introduzindo-se na cadeia polipeptídica isótopos como carbono-13 e/ou nitrogênio-15, que atuam como marcadores (pontos de

referência espacial) (Fersht, 2000).

A análise das estruturas tridimensionais conhecidas de proteínas globulares permite generalizar que a maioria delas parece ser construída por módulos estruturais conhecidos como domínios. Os **domínios estruturais** são entidades geométricas globulares que se supõem enovelarem-se independentemente uns dos outros; e, mais ainda, supõem-se que alguns se enovelam primeiro, ajudando o enovelamento subsequente das partes restantes (Schulz & Schirmer, 1990; Jaenicke, 1999), de acordo com o mecanismo de enovelamento chamado *fold-ing* hierárquico (Baldwin & Rose, 1999). Os domínios estruturais constituem as unidades básicas para a classificação estrutural das proteínas usadas pelos principais métodos: o **SCOP** —Structural Classification of Protein Database, do grupo de Chothia— (<http://scop.mrc-lmb.cam.ac.uk/scop/>) e o **CATH** —Hierarchic Classification of Protein Domain Structures, do grupo de Orengo— (<http://www.biochem.ucl.ac.uk/bsm/catch/>). Tais classificações também utilizam, em conjunto com os domínios, critérios filogenéticos para estabelecer as relações evolucionárias entre as proteínas.

As proteínas globulares pequenas e a maioria das de tamanho médio têm um único domínio; já as proteínas grandes podem ter vários domínios. O SCOP só reconhece uma unidade estrutural como domínio se ela também ocorre individualmente (Murzin et al., 1995), diferentemente do CATH, que usa critérios mais complexos para tal classificação. Isto é um exemplo das possíveis divergências que podem haver entre diferentes métodos de classificação estrutural. O interessante é que todos eles conseguem identificar, entre as dezenas de milhares de proteínas estudadas, somente cerca de 1000 tipos de enovelamento. Também é interessante observar que diferentes proteínas com o mesmo enovelamento freqüentemente apresentam estruturas secundárias (em posições periféricas) e regiões de *loops* que diferem em tamanho e conformação (Murzin et al., 1995).

Tanto o método SCOP quanto o CATH agrupam as proteínas em **classes es-**

truturais, de acordo com suas estruturas secundárias. Como os próprios nomes das classes revelam, a classe “somente- α ” reúne as proteínas constituídas por hélices- α , e a classe “somente- β ” reúne as proteínas constituídas por folhas- β . O CATH agrupa na classe mista as proteínas que apresentam hélices- α e folhas- β ; o SCOP tem duas classes para estas proteínas, dependendo se as estruturas hélices- α e folhas- β alternam-se ou se estão segregadas. Também existem proteínas que não apresentam nem hélices- α e nem folhas- β , constituindo uma classe a parte.

2.4 Seqüência, Estrutura e Função

Mesmo sendo consenso que a estrutura tridimensional de uma proteína globular já está definida na sua seqüência de monômeros, ainda não se conhecem os detalhes de como esta informação é processada. Assim como se observam seqüências altamente homólogas —isto é, bastante semelhantes— correspondendo a uma mesma estrutura, também se observam muitas seqüências significativamente diferentes enovelando-se em uma mesma estrutura (Devos & Valencia, 2000). Experimentos de mutação sugerem que o padrão hidrofóbico-polar (HP) da seqüência é mais determinante do que a própria seqüência em si dos resíduos (Kamtekar et al., 1993; Baldwin & Rose, 1999).

Como mencionado acima, estudos envolvendo dezenas de milhares de estruturas conhecidas de proteínas —cujas coordenadas estão armazenadas no banco de dados Protein Data Bank (PDB)— concluem que, apesar da enorme complexidade estrutural, todas estas estruturas podem ser resumidas a uma combinação de aproximadamente 1000 formas de enovelamento (Murzin et al., 1995; Thornton et al., 1999). Isto significa que apenas uma fração muito pequena das tantas estruturas que seriam possíveis é utilizada na acomodação configuracional das proteínas.

A função de uma proteína não é tão bem definida quanto sua sequência e estrutura tridimensional; diferentes técnicas experimentais mostram apenas aspectos parciais das funções (Devos & Valencia, 2000). Em áreas distintas do conhecimento é usual associar-se forma e função. Acredita-se que a topologia da superfície proteica está intimamente relacionada à sua função (Lee & Richards, 1971). Para Prigogine, a estrutura de uma biomolécula depende da sua função, e, também, do tipo de interação da biomolécula com o ambiente (Prigogine, 1996); conseqüentemente, para ele, estrutura e função não podem ser compreendidos separadamente. Sequência e função relacionam-se de uma maneira bastante complexa (Bork et al., 1998). Constatam-se muitas seqüências com alto grau de homologia correspondendo a uma mesma função; no entanto, também se observam seqüências sem relação alguma apresentando a mesma função. Tudo isto sugere que “seqüência, estrutura e função das proteínas talvez tenham se estabelecido através de um fenômeno evolucionário complexo que não permite simples interpretações” (Devos & Valencia, 2000).

2.5 Interações Presentes nas Proteínas

As proteínas globulares, para serem funcionalmente ativas, devem estar no meio aquoso, em condições intensivas apropriadas de temperatura, pH e pressão. Acredita-se que as estruturas tridimensionais proteicas dependem tanto das interações entre os aminoácidos como também das interações desses com o meio solvente. Mais precisamente, acredita-se que o estado nativo é estabilizado por uma variedade de forças coesivas não-covalentes, incluindo-se ligações de hidrogênio, forças eletrostáticas, empacotamento estérico e efeito hidrofóbico (Nelson et al., 1997).

Os vinte tipos de aminoácidos que ocorrem naturalmente nas proteínas diferem entre si apenas pela cadeia lateral; é esta que lhes confere propriedades

peculiares. Em pH fisiológico, existem dois aminoácidos considerados ácidos (ácido glutâmico e o ácido aspártico) e dois considerados básicos ou alcalinos (lisina e arginina); os outros apresentam carácter neutro. Nas proteínas, quando a cadeia lateral de um aminoácido básico entra em contato com a de um aminoácido ácido, a interação eletrostática que se estabelece entre elas chama-se **ponte salina**, uma vez que os grupos carboxila e amino dessas cadeias laterais ficam ionizados. Estas **interações eletrostáticas** seriam bastante intensas se as proteínas estivessem no vácuo e não na água (onde elas são 80 vezes mais fracas): a constante dielétrica da água é suficientemente alta para oferecer um efeito de blindagem às interações eletrostáticas, que apresentam energia entre $+0,3 \text{ Kcal/mol}$ (repulsiva) a $-5,0 \text{ Kcal/mol}$ (atrativa). Em estudos de desnaturação por alteração de pH, aminoácidos considerados neutros em condições fisiológicas podem se tornar ácidos ou básicos (Schulz & Schirmer, 1990).

Entre dois átomos quaisquer relativamente próximos, separados por uma certa distância d , existe sempre uma fraca interação (na ordem de $0,01$ a $0,1 \text{ Kcal/mol}$), conhecida como **força de van der Waals**, ou **força de dispersão de London**, devido a flutuações na distribuição espacial de suas cargas elétricas. Esta interação é descrita pela parte atrativa do potencial de Lennard-Jones, que está ilustrado na **Figura 2.9**; se a distância entre os átomos é $d > R_m$, o potencial é atrativo e varia com o inverso da distância elevada à sexta potência ($1/d^6$); se, ao contrário, a separação entre eles é muito pequena, $d \ll R_m$, os átomos se repelem de acordo com o princípio de exclusão de Pauli, e o potencial varia com $1/d^{12}$ (Schulz & Schirmer, 1990). O parâmetro R_m corresponde à distância entre os átomos em que o potencial é mínimo, ou seja, em que a força de atração e repulsão estão em equilíbrio (lembrando que força é igual ao negativo do gradiente do potencial, $\vec{F} = -\vec{\nabla} V$).

As **ligações de hidrogênio** são abundantes em proteínas: ocorrem tanto intracadeia como entre a cadeia e o solvente (Daune, 1999). Uma ligação

Figura 2.9: Potencial de Lennard-Jones entre dois átomos separados pela distância d . Se $d > R_m$, o potencial é atrativo; se, ao contrário, a separação entre os átomos é muito pequena, $d \ll R_m$, ocorre repulsão de acordo com o princípio de exclusão de Pauli. R_m é a distância entre os átomos em que o potencial é mínimo (Schulz & Schirmer, 1990, p.29, figura 3.1).

de hidrogênio ocorre sempre que dois átomos eletronegativos compartilham um hidrogênio. Em proteínas, o nitrogênio do grupo amino está ligado covalentemente a um hidrogênio; quando este átomo de hidrogênio se aproxima do oxigênio de um grupo carboxila, esse oxigênio (que é eletronegativo) atrai o hidrogênio de tal forma que estes dois grupos (amino e carboxila) de aminoácidos diferentes ficam unidos por uma ligação de hidrogênio. Nota-se que quaisquer duas partes da molécula, inclusive o **backbone**, que interagem através de ligações de hidrogênio também podem formar ligações de hidrogênio com as moléculas de água. As ligações de hidrogênio são atrativas, com energia em torno de $-4,0$ a $-3,0$ Kcal/mol.

Sabe-se que, em geral, as moléculas de água preferem formar ligações de hidrogênio entre si a estabelecer ligações com qualquer parte das proteínas globulares nelas inseridas. No entanto, as moléculas de água também interagem com as cadeias laterais dos aminoácidos proteicos. Esta interação água-aminoácido é seletiva; existem diferentes graus de afinidade entre as cadeias laterais dos diversos aminoácidos e a água. Isto permite classificar os vinte aminoácidos que ocorrem em proteínas como polares, apolares ou anfífilos. Os aminoácidos po-

lares ou hidrofílicos formam facilmente ligações de hidrogênio com a água, sendo aqueles que têm maior afinidade com ela. Os aminoácidos apolares ou hidrofóbicos não apresentam dipolo elétrico e não formam ligações de hidrogênio com a água, apresentando baixa afinidade com este solvente. Há também os aminoácidos anfifílicos, que apresentam parte da cadeia lateral polar e parte apolar.

A análise das estruturas de proteínas globulares revela que elas são constituídas por uma região central de alta densidade formada, principalmente, pelas regiões hidrofóbicas das cadeias laterais dos aminoácidos, na tentativa de minimizar o contato destas com a água. Entre o caroço hidrofóbico e a água estão, preferencialmente, as regiões hidrofílicas dos aminoácidos. Estas, localizando-se na superfície da molécula, formam pontes de hidrogênio com a água. Este padrão estrutural das proteínas indica que a “força” dominante no processo de enovelamento das proteínas globulares solúveis em água é a hidrofóbica. Contudo, deve ser destacado que não se trata de uma interação intensa entre os resíduos hidrofóbicos (estes interagem através do potencial de Lennard-Jones) e, sim, da tendência das moléculas de água em rearranjar as ligações de hidrogênio entre si, e entre elas e os resíduos polares, a fim de minimizar as perdas interacionais típicas entre água-água.

Existem proteínas que apresentam, na sua seqüência, alguns aminoácidos do tipo cisteína. Quando as cadeias laterais destes aminoácidos se encontram, os grupos *SH* sofrem oxi-redução, perdem o hidrogênio, e os enxofres ligam-se quimicamente formando uma **ponte de sulfeto** (Greighton, 2000). Esta é uma ligação covalente com energia da ordem de 100 Kcal/mol . As proteínas globulares pequenas que se encontram no meio extracelular geralmente apresentam pontes de sulfeto, o que lhes garante maior estabilidade mecânica.

É interessante lembrar que a probabilidade de uma ligação, com energia E , ser desfeita, em favor de outra com energia E' , é geralmente expressa pelo peso

de Boltzmann, $e^{-\Delta E/kT}$, sendo $\Delta E = E' - E$, e, k , a constante de Boltzmann (na temperatura fisiológica, $NkT \simeq 0,6 Kcal/mol$, sendo N o número de Avogadro).

As conformações nativas das proteínas globulares solúveis em água são extremamente sensíveis às alterações físico-químicas do solvente. Quando estas se afastam das condições fisiológicas, as proteínas passam do estado de ordem (nativo) para o estado de desordem (desnaturado), conhecido como **processo de desnaturação**. No outro sentido, quando as condições físico-químicas se reaproximam das fisiológicas, ocorre o processo de folding (enovelamento) ou renaturação das proteínas, isso se o processo de desnaturação for reversível. A variação da energia livre típica entre os estados nativo e desnaturado, de proteínas globulares, está entre, aproximadamente, -5 e $-15 Kcal/mol$, o que equivale a quebra de umas poucas ligações de hidrogênio (Lattman & Rose, 1993). Existem diversos métodos de desnaturação de proteínas:

1. A variação do pH altera os estados de ionização de algumas cadeias laterais das proteínas, desfazendo ligações tipo pontes salinas e quebrando ligações de hidrogênio; usualmente, as proteínas tornam-se menos estáveis na medida em que o pH diminui, diminuindo também a temperatura do pico do calor específico (veja-se **Figura 2.13**) (Nelson & Cox, 2000).
2. O aquecimento da solução proteica aumenta a energia vibracional e rotacional das moléculas; quando a energia térmica kT passa a ser suficientemente grande para romper as interações nativas, ocorre a desnaturação por calor (Privalov, 1992).
3. Por outro lado, o resfriamento da solução proteica, diminuindo a energia vibracional e rotacional das moléculas, interfere no conjunto das diversas interações presentes, podendo promover o rompimento das interações nativas, causando a desnaturação a frio.

4. Agentes desnaturantes, como uréia e guanidina, conseguem romper até mesmo a estrutura da água, aumentando a solubilidade das partes hidrofóbicas da molécula proteica, anulando, dessa forma, a força hidrofóbica, que desempenha papel importante na estabilidade da estrutura nativa. Estes desnaturantes também formam ligações de hidrogênio com os grupos amino e carboxila da cadeia principal, fazendo com que estes grupos quebrem suas pontes de hidrogênio nativas que os unia, desfazendo, assim, as estruturas secundárias (Nelson & Cox, 2000).
5. O aumento da pressão hidrostática também leva à desnaturação devido à incorporação de moléculas de água no interior da molécula proteica; a desnaturação à alta pressão é, entre todos os processos de desnaturação, a que mais retém segmentos estruturais (Zhang et al., 1995; Hummer et al., 1998; Mohana-Borges et al., 1999).

2.6 Aspectos Termodinâmicos do Folding

O trabalho de Anfinsen, em 1961, tratou o problema da **desnaturação de proteínas** como um **processo reversível** (Anfinsen, 1973) sob condições apropriadas, tornando possível, portanto, o seu tratamento termodinâmico. No final da década de 70, Privalov desenvolveu uma técnica de alta sensibilidade, conhecida como calorimetria diferencial de varredura, que tem possibilitado medidas diretas da variação da energia nos processos de enovelamento e desnaturação de proteínas (van Osdol et al., 1991). A determinação de propriedades termodinâmicas (macroscópicas) é importante no sentido de auxiliar o entendimento da energética (microscópica) envolvida no enovelamento de proteínas (potencial intramolecular e interações cadeia-solvente) (Soulages, 1998).

2.6.1 Folding de dois estados e a constante de equilíbrio

Medidas termodinâmicas têm indicado que o processo de enovelamento, para a maioria das proteínas globulares, pequenas e de um único domínio, pode ser adequadamente representado por uma **transição de dois estados**: ou as moléculas estão no estado nativo N, ou estão no estado desnaturado D; se existem estados intermediários, o tempo que as moléculas permanecem nele é desprezível. Muitas simplificações ocorrem quando se considera o **folding** como um processo de dois estados; assim, o enovelamento reversível pode ser visto como uma reação de equilíbrio que interconverte os estados N e D —cada estado obedecendo a uma simples distribuição de microestados que são função da temperatura T , do pH e outras condições do meio solvente (Shortle, 1989). Embora a aproximação de dois estados seja limitada às proteínas menores, ela permite aplicar a termodinâmica do equilíbrio no entendimento das características básicas da reação de enovelamento.

Quando se considera o comportamento de dois estados de uma população de proteínas em meio solvente, a **temperatura de meio ponto** $T_{1/2}$ é aquela em que metade da população das moléculas está enovelada e metade está desnaturada. Para um processo de dois estados, a **constante de equilíbrio aparente**, K_N , é simplesmente a razão entre a concentração de moléculas no estado nativo, $[N]$, e a concentração de moléculas desnaturadas, $[D]$:

$$K_N = \frac{[N]}{[D]}. \quad (2.1)$$

2.6.2 Energia livre de Gibbs

A variação da **energia livre de Gibbs** entre os estados nativo e desnaturado é a medida fundamental da estabilidade de uma proteína, e pode ser escrita como (Shortle, 1989):

$$\Delta G_N = G_N - G_D = -RT \ln K_N, \quad (2.2)$$

onde R é a constante universal dos gases ($8,3 \text{ J/mol.K}$), T é a temperatura absoluta e K_N é a constante de equilíbrio.

Propriedades termodinâmicas, como a variação da energia livre de Gibbs, ΔG_N , entre os estados nativo e desnaturado, podem ser determinadas através de diferentes métodos experimentais, sendo o mais utilizado a calorimetria diferencial por varredura (Shortle, 1989). As proteínas globulares apresentam uma estabilidade marginal, pois como já mencionado, ΔG_N apresenta valores típicos entre -5 e -15 kcal/mol (Lattman & Rose, 1993) —lembrando que $RT \simeq 0,6 \text{ kcal/mol}$ é a energia térmica na temperatura fisiológica ($T = 310 \text{ K}$).

Do ponto de vista biológico, é uma vantagem o fato das proteínas serem marginalmente estáveis, pois os seres vivos não só produzem constantemente as proteínas, como também as degradam constantemente. É interessante perceber que a degradação de uma proteína é tão importante quanto a sua síntese. Algumas proteínas (intracelulares) duram o mesmo tempo que as células, mas outras “vivem” apenas segundos (Bachmair et al., 1986).

A estabilidade de uma molécula proteica, que depende da variação da energia livre de Gibbs, ΔG_N , também pode ser expressa em termos das variações entálpica e entrópica, entre os estados nativo e desnaturado, como se explica a seguir.

Quando um sistema recebe calor, através de um processo à pressão constante, é bem estabelecido que a **variação da entalpia**, ΔH , deste sistema, entre dois estados quaisquer, é equivalente ao **calor**, Q , absorvido por ele. Se, além da pressão, a temperatura também se mantém constante neste processo, a variação da entalpia estará relacionada diretamente à variação das interações atômicas e moleculares; o rompimento ou formação de ligações químicas sempre produz

variações de entalpia (Chagas, 1999, Atkins, 1994). Já a **entropia**, S , de um sistema está relacionada com a capacidade do mesmo adquirir diferentes configurações; conseqüentemente, a variação da entropia, ΔS , entre dois estados significa a variação do número de configurações acessíveis entre estes (Shortle, 1989). Quanto mais graus de liberdade um sistema apresentar maior será a capacidade de adquirir diferentes configurações (Chagas, 1999). A energia livre de Gibbs pode ser expressa como $G = H - TS$ e, assim, quando a temperatura é mantida constante, pode-se escrever a variação da energia livre de Gibbs como:

$$\Delta G = \Delta H - T\Delta S. \quad (2.3)$$

Observa-se que ΔG depende tanto da contribuição entálpica, ΔH —quebra ou formação de ligações químicas—, como da contribuição entrópica, $T\Delta S$ —número de diferentes possibilidades configuracionais.

Neste trabalho adotar-se-á a aproximação de que a variação da energia livre no **processo de enovelamento** é devida, principalmente, ao efeito entrópico. Esta aproximação é justificada tomando-se como base experimental a constatação de que os grupos hidrofílicos estabelecem ligações de hidrogênio tanto com o meio como entre eles, quando no interior da proteína (Rose et al., 1985), não contribuindo, assim, de forma majoritária com a variação da parcela entálpica.

Na maioria das reações biológicas a variação do volume e da pressão são desprezíveis; então, a variação da energia livre de Gibbs é equivalente à variação da **energia livre de Helmholtz**, $F = E - TS$, e a variação da entalpia é equivalente a variação da energia interna, ΔE , assim:

$$\Delta F = \Delta E - T\Delta S. \quad (2.4)$$

As variações espontâneas de ΔF são tais que E tende a ser o menor possível e

TS o maior possível; assim o sistema tenderá a se encontrar no mínimo de F .

2.6.3 Efeito hidrofóbico e área acessível ao solvente

O efeito hidrofóbico ou entrópico pode ser elucidado através de um sistema bastante simples, onde se tem, inicialmente, duas gotas de óleo em água, cada uma com volume V e superfície S . As gotas de óleo são altamente hidrofóbicas e não formam pontes de hidrogênio com a água, que é polar. Na **Figura 2.10 (a)** ilustra-se como as moléculas de água organizam-se ao redor das gotas de óleo de tal forma a maximizarem o número de ligações de hidrogênio entre elas. Considere-se, então, que estas duas gotas de óleo se encontram e se unem, constituindo uma gota maior (**Figura 2.10 (b)**). Nesta nova situação, o número de moléculas de água organizadas é menor do que quando as duas gotas estavam separadas, pois a organização depende da superfície hidrofóbica da gota de óleo que é exposta, que agora é $1,6S$ em vez de $2S$. Pode-se dizer que é o **efeito entrópico** ou **hidrofóbico** que estabiliza o sistema nessa nova situação, onde as gotas de óleo estão unidas e o meio solvente apresenta maior grau de liberdade. Em outras palavras, a configuração das gotas de óleo unidas é mais favorável por que a entropia do sistema água-óleo é maior, e, conseqüentemente, a energia livre é menor.

Em 1959, Kauzmann considerou que o mesmo efeito entrópico que une gotas de óleo dispersas na água também une as partes hidrofóbicas das proteínas, quando em meio aquoso, guardada a diferença de que os aminoácidos em proteínas estão ligados covalentemente aos seus vizinhos na cadeia e não podem se mover independentemente uns dos outros (Kauzmann, 1959; Roseman, 1988). Ele cunhou a expressão “**ligação hidrofóbica**” para descrever o ganho de energia livre que ocorre ao se transferir resíduos apolares do meio aquoso para o interior da proteína (Kauzmann, 1959). Desde então, tornou-se usual considerar o efeito hidrofóbico como o principal responsável por organizar a estrutura

Figura 2.10: Gotas de óleo em água. Em (a), tem-se duas gotas de óleo organizando (diretamente) 16 moléculas de água. Em (b), as duas gotas se uniram, e, agora, o número total de moléculas de água organizadas (diretamente) é 11.

tridimensional das proteínas globulares solúveis em água (Kamtekar et al., 1993; Baldwin, 2002).

Na tentativa de quantificar o efeito hidrofóbico no processo de enovelamento de proteínas, foram desenvolvidas muitas **escalas hidrofóbicas** —que medem a afinidade dos aminoácidos (ou resíduos) pela água. Algumas estão baseadas na análise das estruturas de proteínas conhecidas, quantificando-se a frequência com que cada resíduo está exposto à água ou escondido no interior das moléculas proteicas. No entanto, a validade destas escalas é muito discutida; por exemplo, certas análises das estruturas nativas de proteínas revelam que, em média, tem-se 50% de resíduo hidrofóbicos na superfície exposta ao solvente (Roseman, 1988). Outras escalas hidrofóbicas estão baseadas na medida da variação da energia livre quando se transfere as cadeias laterais dos aminoácidos do meio aquoso para um solvente orgânico, o qual tenta reproduzir o interior das proteínas. Existem diferenças significativas entre estas escalas, pois elas são dependentes de muitos parâmetros, como da escolha do meio orgânico e do método empregado. Em alguns casos, observa-se que aminoácidos, considerados hidrofóbicos em uma escala, aparecem como hidrofílicos em outra (Rose et al., 1985).

As escalas que utilizam a variação da energia livre de transferência estão

Figura 2.11: Área acessível ao solvente. Cada átomo da proteína é representado por uma esfera de raio igual ao seu raio de van der Waals. A área acessível ao solvente é determinada com o auxílio de uma esfera de raio $R = 1,4 \text{ \AA}$ — representando a molécula de água — que “rola” por toda a superfície externa da proteína (Daune, 1999, p.279, Figura 13.1).

fundamentadas no conceito da **área acessível ao solvente**, proposto por Lee e Richards em 1971. Para definir a área acessível ao solvente, de um grupamento atômico, cada átomo deste grupamento deve ser representado por uma esfera de raio igual ao seu raio de van der Waals, como é mostrado na **Figura 2.11**. A área acessível ao solvente constitui uma superfície imaginária que envolve o grupamento atômico, de tal forma que cada ponto desta superfície imaginária é o centro de uma esfera de raio R , representando a molécula do solvente; esta esfera de solvente pode “rolar” por toda a superfície externa do grupamento atômico, sem, no entanto, penetrá-lo. Considerando-se a água como solvente, R vale $1,4 \text{ \AA}$, pois este é o raio de van der Waals para a molécula de água (Lee & Richards, 1971).

Definição da variação da área exposta

Considerando-se um resíduo i , representa-se a sua área exposta ao solvente, quando a proteína está num estado desnaturado (tomado como referência), por A_i^D ; e a média da área exposta ao solvente por $\langle A_i^N \rangle$, quando diversas estruturas nativas de proteínas são consideradas na análise. Define-se, então, a variação da

Figura 2.12: Variação da área (dos aminoácidos) exposta ao solvente (em Å²), em função da variação da energia livre (em cal/mol), quando os aminoácidos são transferidos do meio aquoso para um solvente orgânico (adaptado de Rose et al., 1985, Figura 4-b).

área exposta durante o processo de enovelamento, para este resíduo i , como:

$$\Delta A_i = A_i^D - \langle A_i^N \rangle. \quad (2.5)$$

Rose e colaboradores demonstraram que a variação da área exposta ao solvente dos aminoácidos é proporcional à variação da energia livre quando estes são transferidos do meio aquoso para um solvente orgânico, como ocorre no processo de enovelamento, conforme ilustra a **Figura 2.12** (Rose et al., 1985).

2.6.4 Capacidade térmica

Em calorimetria de proteínas, usa-se a seguinte expressão mecânico-estatística para calcular a **capacidade térmica**, à pressão constante, C_p :

$$C_p = \frac{\partial \langle H(T) \rangle}{\partial T} = \frac{\langle H^2(T) \rangle - \langle H(T) \rangle^2}{kT^2}, \quad (2.6)$$

onde $H(T)$ é a entalpia na temperatura T ; o símbolo $\langle \rangle$ denota média termodinâmica e k é a constante de Boltzmann (Kaya & Chan, 2000).

As interpretações da capacidade térmica, à pressão constante, das proteínas não são tão diretas quanto sua definição (isto é, derivada da entalpia em relação à temperatura). A dificuldade está na atribuição do significado à variação da entalpia na reação de desnaturação reversível de proteínas por alteração da temperatura. Alguns autores (Shortle, 1989; Makhatadze & Privalov, 1990), como será visto abaixo, atribuem à variação da entalpia o calor absorvido pela água devido à hidratação hidrofóbica —que é a exposição dos grupos hidrofóbicos às moléculas de água durante a desnaturação.

A **Figura 2.13** ilustra o comportamento da capacidade térmica, à pressão constante, C_p , em função da temperatura, T , considerando-se três valores diferentes de pH , para uma proteína globular (Privalov, 1997). Observe-se que, na medida que T aumenta, C_p inicialmente também aumenta, mas, depois de atingir um valor máximo, C_p diminui, retornando a valores próximos do valor inicial, porém um pouco mais elevados. Isto evidencia que existe uma variação da capacidade térmica, à pressão constante, entre os estados desnaturado e nativo, definida como $\Delta C_p = C_D - C_N$, onde C_D é a capacidade térmica do estado desnaturado D (correspondendo a temperaturas mais altas) e C_N é a do estado nativo N (correspondendo a temperaturas mais baixas) (Shortle, 1989).

Argumenta-se que o aumento da capacidade térmica com a temperatura, e, entre os estados nativo e desnaturado, é devido: (i) em grande parte ao efeito de hidratação dos grupos que, no estado nativo, estavam no interior da proteína e, no estado desnaturado, passam a contatar com a água; e (ii) ao aumento da liberdade configuracional dos grupos internos da proteína após a ruptura das ligações que mantinham a estrutura nativa (Makhatadze & Privalov, 1990; Privalov & Makhatadze, 1992).

Privalov e Makhatadze, definiram um estado compacto teórico como aquele

Figura 2.13: Capacidade térmica de uma proteína globular, à pressão constante, em função da temperatura, para três pH diferentes: 2,0 (a), 2,5 (b), e 4,5 (c). Note-se, entre as linhas pontilhada e contínua, a diferença entre as capacidades térmicas dos estados desnaturados (U) e nativos (N), ΔC_p (Privalov, 1997, Figura 5).

em que a proteína globular já está bastante compactada, sem, no entanto, apresentar as ligações intramoleculares. Então, calcularam a capacidade térmica para algumas proteínas globulares correspondente a este estado e compararam com esta capacidade térmica do estado nativo. Estes autores se surpreenderam ao encontrar praticamente os mesmos valores para ambos os estados, pois isto significava que a ruptura das ligações não-covalentes intracadeia, que mantêm a estrutura nativa, não contribuía de forma significativa com o aumento da capacidade térmica (Privalov & Makhatadze, 1992). Estudos posteriores confirmam que a contribuição para a capacidade térmica das interações não-covalentes é muito pequena (Gómez et al., 1995).

2.6.5 Entalpia calorimétrica e Entalpia de van't Hoff

A **entalpia calorimétrica**, ΔH_{cal} , é o calor total absorvido em todo o processo de transição da desnaturação por elevação da temperatura, T , obtida pela inte-

gração da área abaixo da curva da capacidade térmica, à pressão constante, C_p , em função da temperatura.

Freqüentemente, dados experimentais de calorimetria de proteínas fornecem a entalpia van't Hoff, que pode ser obtida de diversas maneiras. Uma das fórmulas mais usadas para calcular a **entalpia de van't Hoff**, ΔH_{vH} , é:

$$\Delta H_{vH} = 2T_{mp}\sqrt{kC_p(T_{mp})}. \quad (2.7)$$

T_{mp} é a temperatura de meio ponto e pode ser: (i) a temperatura $T_{\max}$, correspondente ao pico da capacidade térmica, à pressão constante; (ii) a temperatura T_d , correspondente àquela onde a metade do calor total é absorvido; (iii) a temperatura $T_{1/2}$ em que 50% das proteínas se encontram no estado nativo. $C_p(T_{mp})$ é a capacidade térmica, à pressão constante, correspondente à temperatura T_{mp} utilizada; e k é a constante de Boltzmann (Chan, 2000).

Quando o **folding** é de dois estados as três temperaturas de meio ponto são iguais; portanto, as três fornecerão o mesmo valor para a entalpia de van't Hoff. Adicionalmente, para um processo de desnaturação/renaturação de dois estados, as entalpias calorimétrica e de van't Hoff também são numericamente iguais (Kaya & Chan, 2000).

2.7 Aspectos Cinéticos do Folding de Proteínas

A maioria dos estudos teóricos e experimentais sobre o enovelamento de proteínas direciona-se para a solução do chamado **paradoxo de Levinthal**, tentando compreender como uma proteína pode encontrar umas poucas configurações, que constituem a conformação nativa, no meio de um vasto espaço configuracional (Govindarajan & Goldstein, 1995). No final da década de 70, Levinthal chamou atenção para a improbabilidade estatística das proteínas encontrarem a estrutura

nativa por busca aleatória. Por exemplo, ao se considerar que:

(a) cada aminoácido que constitui a proteína possa assumir apenas dois estados configuracionais; (b) a proteína em questão é relativamente pequena, apresenta 110 resíduos na sequência; (c) que a cadeia proteica leva apenas $10^{-15}s$ para visitar uma configuração; e (d) a exploração do espaço das configurações é aleatória; então, esta proteína hipotética levaria um tempo de $2^{110} \times 10^{-15}s \simeq 10^{18}s$ para percorrer todo o espaço configuracional (tempo maior do que a idade do universo).

Sabe-se que as proteínas pequenas e de um único domínio enovelam-se rapidamente, em milissegundos, o que sugere que devem existir caminhos preferenciais para o enovelamento, conduzindo a molécula para o mínimo global (ou local) de energia. Neste trabalho, utiliza-se um potencial estérico-hidrofóbico que, em estudos anteriores, revelou ser capaz de conduzir as cadeias tipo-proteínas à configuração nativa através de duas rotas preferenciais (da Silva et al., 2001) estando de acordo com dados experimentais, obtidos com proteínas reais (Plaxco et al., 1997), como será mostrado mais adiante.

Na década de 90, surgiram diversos trabalhos experimentais em cinética do enovelamento de pequenas proteínas globulares de um único domínio estrutural (Plaxco et al., 2000). Experimentos termodinâmicos já tinham demonstrado que o **folding** destas biomoléculas pode ser descrito por um processo reversível do tipo dois estados (Privalov & Khechinashvili, 1974), ou seja, ou as moléculas estão no estado nativo, ou estão desnaturadas, não existindo estados intermediários detectáveis; os experimentos cinéticos comprovaram tais características e, então, ficou estabelecido que o enovelamento de pequenas proteínas globulares pode ser descrito, com boa aproximação, através de uma **reação de primeira ordem** (Fersht, 1995). No entanto, existem algumas exceções: a apomioglobina e o citocromo-c, por exemplo, apresentam estados intermediários. Estas proteínas se enovelam de acordo com um processo no qual certas estruturas secundárias

nativas se formam rapidamente e estabilizam a cadeia proteica num estado metaestável intermediário, chamado glóbulo fundido (*molten-globule*) (Baldwin & Rose, 1999).

Um dos principais parâmetros cinéticos é a **taxa de enovelamento** k_f (folding rate) e está relacionado com a probabilidade, por unidade de tempo, de uma molécula passar do estado desnaturado para o estado nativo. As proteínas que se enovelam rapidamente, em geral, apresentam k_f na ordem de 10^3 a $10^5 s^{-1}$, enquanto as que se enovelam mais lentamente apresentam k_f em torno de $10^{-3} s^{-1}$; alguns autores relatam que os valores conhecidos de k_f chegam a abranger 10^9 ordens de grandeza (Chan, 1998).

Uma vez que o enovelamento pode ser descrito como uma reação de primeira ordem, a taxa de concentração de moléculas desnaturadas ($\frac{d[D]}{dt}$) é proporcional à concentração destas moléculas (elevada à primeira potência), ou seja:

$$\frac{d[D]}{dt} = -k_f[D], \quad (2.8)$$

onde k_f é a constante de proporcionalidade e $[D]$, a concentração de moléculas no estado desnaturado no tempo t (Chang, 2000). Considerando-se o enovelamento do tipo dois estados, define-se **tempo característico de folding**, t_f , ou simplesmente **tempo de folding**, o tempo necessário para enovelar 63% das moléculas, ou seja, o tempo de folding é o inverso da taxa de enovelamento ($t_f = 1/k_f$) (Fersht, 1995).

Experimentalmente, pode-se obter k_f diretamente da curva exponencial de fluorescência, que também representa a função concentração de moléculas desnaturadas com o tempo, $[D] = [D_0]e^{-k_ft}$, sendo $[D_0]$ a concentração inicial de proteínas no estado desnaturado (Schindler & Schmid, 1996; Plaxco et al., 1997). Os experimentos de fluorescência são adequados quando a proteína estudada apresenta em sua seqüência o triptofano (Trp). Esse aminoácido apresenta a propriedade de emitir radiação de fluorescência após ser excitado adequadamente. Na maioria

Figura 2.14: Intensidade de fluorescência, em função do tempo, para três concentrações finais de desnaturante. A taxa de enovelamento (correspondente à concentração de desnaturante final) pode ser obtida diretamente de cada curva (adaptado de Ferguson et al., 1999, Figura 3-a).

das proteínas globulares, o Trp encontra-se nas partes mais internas das estruturas tridimensionais nativas; então, na medida em que as proteínas se enovelam o sinal fluorescente diminui na mesma proporção, permitindo, então, quantificar diretamente a concentração relativa de proteínas no estado desnaturado a partir da intensidade de fluorescência. A **Figura 2.14** apresenta três curvas da intensidade de fluorescência, em função do tempo, cada qual obtida com uma concentração final de desnaturante.

Sabe-se que as proteínas com enovelamento tipo dois estados apresentam a taxa de enovelamento, k_f , variando exponencialmente com a concentração do desnaturante, ou, em outras palavras, o logaritmo natural de k_f decresce linearmente com o aumento da concentração de desnaturante (Fersth, 2000), conforme mostra a **Figura 2.15**. O valor de k_f apresentado na literatura, na maioria das vezes, é o seu valor intrínseco —obtido por extrapolação na **Figura 2.15**— correspondente à condição sem desnaturante (Ferguson et al., 1999).

Figura 2.15: Logaritmo da taxa de folding, $\ln k_{obs}$, em função da concentração final de desnaturante. O valor intrínseco da taxa de folding é obtido por extrapolação da curva quando a concentração de desnaturante é nula (Ferguson et al., 1999, Figura 4-a).

2.8 Parâmetros Topológicos do Estado Nativo

Há muitos anos, vem-se discutindo o papel dos contatos de curto e de longo alcance na cinética do enovelamento, tomando-se como base dados teóricos obtidos a partir de modelos computacionais simplificados. Go e colaboradores, os pioneiros nos modelos computacionais para o enovelamento de cadeias tipo-proteínas, encontraram que as “proteínas” tendem a se enovelar mais rapidamente quando os contatos de curto alcance entre os resíduos são mais energéticos do que os de longo alcance (Go & Taketomi, 1978) —lembrando-se que um contato entre dois resíduos é dito de curto alcance (ou contato local) quando entre eles existem poucas unidades de monômeros ao longo da seqüência; e, ao contrário, um contato é dito de longo alcance (ou contato não-local), quando os aminoácidos que estão interagindo estão bastante separados ao longo da seqüência. Em contraste, outro modelo computacional para o enovelamento obtém que são as “proteínas” com maior número de contatos de longo alcance as que se enovelam mais rapidamente e com características de processo de dois estados (Abkevich et al., 1995).

O desenvolvimento de técnicas experimentais em cinética química permitiu quantificar a taxa de enovelamento, k_f , de diversas proteínas globulares e pequenas —ou, dito de outra forma, permitiu medir o tempo característico de enovelamento, t_f , pois $t_f = 1/k_f$ (Schönbrunner et al., 1997).

Sabe-se que as proteínas constituídas por hélices- α (que apresentam muitos contatos de curto alcance) enovelam-se mais rápido do que as constituídas por folhas- β (que apresentam mais contatos de longo alcance) (Chan, 1998). Muito provavelmente, inspirados por tais resultados, Plaxco e colaboradores propuseram um parâmetro topológico —chamado ordem de contato— que quantifica, em média, o alcance dos contatos nativos em proteínas (Plaxco et al., 1998). O grupo de Plaxco trabalhou com a hipótese de que a taxa de enovelamento de proteínas, k_f , diminui exponencialmente na medida que a ordem de contato aumenta. Então, para validar o parâmetro proposto, este grupo avaliou a correlação linear existente entre $\ln k_f$ e ordem de contato, encontrando uma forte correlação negativa (Plaxco et al., 1998; Alm & Baker, 1999).

Acreditando-se que a complexidade estrutural do estado nativo das proteínas globulares possa ser capturada, de alguma forma, através de uma equação matemática, outros parâmetros topológicos foram propostos na literatura, a saber, ordem de longo alcance e distância de contato total (Gromiha & Selvaraj, 2001; Zhou & Zhou, 2002). Todos esses parâmetros utilizam o conceito de **contato topológico**: diz-se que dois monômeros i e j estão em contato topológico, constituindo um par em contato, quando não são consecutivos ao longo da cadeia e estão a uma distância espacial de no máximo R_{cut} , que tipicamente apresenta valores entre 4 e 6 Å (Plaxco et al., 1998). Apresenta-se, a seguir, as definições de tais parâmetros topológicos, que também serão aplicadas ao modelo computacional estudado neste trabalho.

2.8.1 Ordem de contato χ

A **ordem de contato**, χ , representa o número médio de monômeros existentes ao longo da cadeia entre os pares de aminoácidos, i e j , em contato topológico, normalizado pelo número total de contatos, n_c , e pelo número total de resíduos da cadeia, n_r , definida como

$$\chi = \frac{1}{n_c} \sum_{\substack{\{n_c\} \\ |i-j| > l_{cut}}} \frac{|i-j|}{n_r}. \quad (2.9)$$

O somatório é realizado sobre todo o conjunto dos contatos $\{n_c\}$, desde que $|i-j| > l_{cut}$, sendo l_{cut} a separação mínima (ou de corte) entre i e j ao longo da cadeia (usualmente $l_{cut} = 2$); considera-se que os pares ij estão em contato topológico quando a distância espacial entre eles, $|\mathbf{r}_i - \mathbf{r}_j|$, é no máximo 6Å (Plaxco et al., 1998).

Proteínas com χ pequena (5%) têm preferencialmente contatos de curto alcance, ou seja, a separação média dos resíduos em contato ao longo da cadeia é pequena; enquanto as que apresentam χ elevada (40%) exibem preferencialmente interações (ou contatos) de longo alcance (Plaxco et al., 2000).

2.8.2 Ordem de longo alcance χ_{LR}

Com o intuito de quantificar o número de contatos de longo alcance —contatos entre monômeros que estão perto no espaço e bastante separados ao longo da sequência— foi proposto o parâmetro **ordem de longo alcance**, χ_{LR} , definido como:

$$\chi_{LR} = \frac{n_c}{n_r}, \quad (2.10)$$

onde n_c é o número total de contatos de longo alcance e n_r é o número de resíduos da cadeia. Considera-se, nesta definição, que dois monômeros i e j constituem

um contato de longo alcance quando a distância entre eles ao longo da cadeia é $|i - j| > 12$ e a distância espacial entre eles é no máximo $R_{cut} = 8\text{\AA}$ (Gromiha & Selvaraj, 2001).

2.8.3 Distância de contato total χ_T

O parâmetro **distância de contato total**, χ_T , está fundamentado na definição da separação média, $\bar{S}$, ao longo da cadeia. Esta é a função geratriz de todas as outras definições e qualificações de “contatos” e pode ser calculada através de:

$$\bar{S} = \frac{2}{n_r(n_r - 1)} \sum_{|i-j|>0} |i - j|, \quad (2.11)$$

onde $\sum_{|i-j|>0} |i - j|$ significa o somatório das separações ao longo da sequência entre todos os monômeros i e j ; e $\frac{n_r(n_r-1)}{2}$ é o número total de contatos (Zhou & Zhou, 2002).

Não é difícil mostrar que a forma geral da separação média também pode ser dada por:

$$\bar{S} = \frac{2}{n_r(n_r - 1)} \frac{1}{2} \sum_{i=1}^{n_r-1} i(i+1) = \frac{1}{3}(n_r + 1). \quad (2.12)$$

Evidentemente, a grandeza $\bar{S}$ não depende da conformação da cadeia. Mas ao se introduzir uma distância espacial de corte, R_{cut} , e uma separação mínima ao longo da cadeia, a separação de corte, l_{cut} , entre os resíduos i e j em contato, na equação da separação média ao longo da sequência (**Eq.2.11**), obtém-se a definição da **distância de contato total**, χ_T :

$$\chi_T = \frac{1}{n_r} \sum_{\substack{\{n_c\} \\ |i-j|>l_{cut}}} \frac{|i - j|}{n_r}, \quad (2.13)$$

que, agora, é função da conformação da cadeia. Observe-se que o somatório deve

ser feito para todo o conjunto dos pares em contato $\{n_c\}$, onde $|\mathbf{r}_i - \mathbf{r}_j| > R_{cut}$ e $|i - j| > l_{cut}$ (Zhou & Zhou ,2002).

É interessante notar que, quando se escolhe um mesmo l_{cut} e um mesmo R_{cut} para todos os três parâmetros topológicos, pode-se obter χ_T a partir do produto da χ e χ_{LR} .

O parâmetro distância de contato total, χ_T , difere do parâmetro ordem de contato, χ , apenas por um pré-fator: em χ_T a separação relativa ao longo da seqüência é normalizada pelo número total de monômeros e em χ essa separação relativa é normalizada pelo número total de contatos, o que faz com que a ordem de contato represente a média da separação ao longo da seqüência (uma vez que o somatório é sobre todos os contatos), o que não ocorre com χ_T ; esta é a razão deste último ser chamado distância de contato “total” (Zhou & Zhou ,2002).

Nota-se que para o caso especial de se ter um conjunto de proteínas com o mesmo número de resíduos, n_r , e o mesmo número total de contatos, n_c , comum nos modelos em rede, os parâmetros ordem de contato, χ , e distância de contato total, χ_T , diferem apenas por uma constante. Por esta razão, no presente trabalho, as relações encontradas entre χ e outras grandezas também se aplicam à χ_T , pois, no modelo em rede adotado, tem-se $n_r = 27$ e $n_c = 28$ para todas as estruturas que representam configurações nativas.

2.8.4 Correlação entre $\ln k_f$ e Parâmetros Topológicos

Dados experimentais revelam uma certa tendência da taxa de enovelamento de proteínas, k_f , diminuir na medida em que o número de contatos de longo alcance aumenta. Plaxco e colaboradores levantaram a hipótese de que k_f diminui exponencialmente com o aumento da ordem de contato, χ , ou, em outras palavras, de que o logaritmo natural da taxa de enovelamento, $\ln k_f$, diminui linearmente com o aumento de χ . Para validar o parâmetro topológico proposto por eles,

verificaram, então, a correlação linear existente entre $\ln k_f$ e χ (Plaxco et al., 1998). Todos os outros trabalhos que vêm se desenvolvendo nesta linha também usam este mesmo critério de validação (Alm & Baker, 1999; Gromiha & Selvaraj, 2001; Zhou & Zhou, 2002).

Mostram-se na **Figura 2.16**, obtidos para um grupo de 28 proteínas globulares e pequenas (Zhou & Zhou, 2002), os gráficos de dispersão entre o logaritmo natural da taxa de enovelamento, $\ln k_f$, e os parâmetros topológicos: (a) ordem de contato, χ ; (b) ordem de longo alcance, χ_{LR} ; e (c) distância de contato total, χ_T . Comparando-se visulamente os três gráficos (A, B e C), pode-se dizer que a ordem de contato, χ , é a que menos se correlaciona com o $\ln k_f$ (o coeficiente de correlação é $-0,74$), onde se nota que as proteínas com maior χ (acima de 30%) são as que mais se afastam da reta ajustada. A ordem de longo alcance, χ_{LR} , apresenta uma correlação um pouco maior do que a de χ (o coeficiente de correlação é $-0,81$), e a distância de contato total, χ_T , é a que se correlaciona melhor com o $\ln k_f$ (o coeficiente de correlação é $-0,88$). A correlação entre χ_T e o $\ln k_f$ é maior do que aquela entre χ e o $\ln k_f$ porque a χ_T , ao ser normalizada pelo número total de monômeros, preserva mais informação do que a χ ; esta última, ao ser normalizada pelo número total de contatos, torna-se uma média e perde informação estrutural. O parâmetro topológico χ_T revela que o somatório da separação relativa ao longo da seqüência, entre monômeros em contato, é um ingrediente importante na cinética do enovelamento de proteínas.

Neste trabalho, faz-se um estudo da distribuição destes parâmetros topológicos, no conjunto de todas as conformações maximamente compactas do modelo adotado, e verifica-se a relação entre estes parâmetros e algumas propriedades cinéticas e termodinâmicas obtidas através de experimentos computacionais para mais de cinquenta cadeias tipo-proteínas.

Figura 2.16: Logaritmo natural da taxa de folding ($\ln k_f$), em função dos parâmetros topológicos: (A) ordem de contato, (B) ordem de longo alcance, e (C) distância de contato total (adaptado de Zhou & Zhou, 2002, Figura 1).

2.9 O Potencial Estatístico Inter-Resíduos

Em 1985 já existia no Protein Data Bank (PDB) um número razoável de estruturas tridimensionais proteicas determinadas experimentalmente. Isto permitiu a Miyazawa e Jernigan construir uma matriz 20×20 que representa o potencial estatístico de contato entre todos os pares de aminoácidos. Estes autores determinaram a frequência de contato entre os aminoácidos em um conjunto de estruturas proteicas, considerando dois aminoácidos em contato aqueles que não são consecutivos ao longo da cadeia e que estão separados espacialmente por uma distância de até $6,5\text{\AA}$ (Miyazawa & Jernigan, 1985). Usando a aproximação da mistura aleatória, eles calcularam a energia de contato, ou potencial estatístico de contato, U_{ij} , conforme a relação:

$$U_{ij} = -kT \ln f_{ij}, \quad (2.14)$$

onde k é a constante de Boltzmann, T é a temperatura absoluta; i e j são os aminoácidos em contato; e f_{ij} a frequência de contato normalizada entre i e j . É interessante observar que, mesmo considerando que este potencial tenha sido derivado de dados experimentais, ele não representa exatamente o potencial verdadeiro; este último ainda não se conhece (Mirny & Shakhovich, 1996). A normalização da frequência de contato f_{ij} depende da frequência com que o par de aminoácidos i e j se encontra num estado de referência hipotético, no qual a energia entre o par seria zero. Em tal estado de referência, supõe-se que os aminoácidos e as moléculas do solvente estão misturados aleatoriamente, formando contatos independentes e uniformemente distribuídos. No entanto, os aminoácidos em proteínas estão covalentemente ligados, determinando uma sequência específica. A equação acima pode ser usada quando se supõe que cada par de aminoácido é independente dos outros pares da molécula. Entretanto, uma das características mais marcantes das proteínas é que suas estruturas são altamente compactas, de tal forma que o espaço ocupado por pares de aminoá-

cidos pode constituir uma forte restrição para a posição ocupada por outro par. Apesar dessas aproximações, existem evidências favoráveis ao uso do potencial estatístico inter-resíduo (Thomas & Dill, 1996).

A matriz 20×20 de Miyazawa e Jernigan (matriz MJ) foi estudada por outros pesquisadores que utilizaram o método geral de análise de matrizes por decomposição em autovalores e autovetores. Isto lhes permitiu reconstruir a matriz real e simétrica, $N \times N$, através do resultado geral:

$$M_{ij} = \sum_{\alpha=1}^N \lambda_{\alpha} V_{\alpha,i} V_{\alpha,j}, \quad (2.15)$$

onde M_{ij} é o elemento da matriz na linha i e coluna j , λ_{α} é o α -ésimo autovalor e $V_{\alpha,i}$ é o autovetor correspondente para o i -ésimo aminoácido (Li et al., 1997). Li e colaboradores descobriram, assim, que existem apenas dois autovalores significativos, indicando que, apesar das diversas interações presentes nas proteínas, a interação entre os resíduos responsável pela estrutura tridimensional nativa pode ser atribuída a somente duas forças, a saber: a força de interação hidrofóbica e a força de dissociação (demixing). Esses autores chegaram, então, a uma equação simplificada para a matriz MJ:

$$M_{ij} = h_i + h_j - C_2(q_i - q_j)^2/2 \quad (2.16)$$

onde

$$h_i = C_o/2 + C_1 q_i + (C_2/2) q_i^2. \quad (2.17)$$

O grupo de Li percebeu que a parcela $-C_2(q_i - q_j)^2 = \omega_{ij}$ tem uma estreita similaridade com a teoria da solubilidade de Hildebrand, desenvolvida em 1933, a qual é válida para moléculas de líquidos simples, misturadas aleatoriamente, que obedecem o potencial de Lennard-Jones.

A expressão seguinte para ω_{ij} está relacionada com o **Princípio da Segregação** e quantifica a variação da energia devido a mistura de dois tipos de moléculas, dada por:

$$\omega_{ij} = 2M_{ij} - M_{ii} - M_{jj}. \quad (2.18)$$

Pode-se dizer que ω_{ij} é a energia necessária para quebrar os contatos ii (que apresenta energia $M_{ii} = E_{ii}$) e jj (que apresenta energia $M_{jj} = E_{jj}$) e formar dois pares em contato ij (cada um com energia $M_{ij} = E_{ij}$). O Princípio da Segregação diz, então, que, para qualquer par ij , deverá ser válido $E_{ii} + E_{jj} \leq 2E_{ij}$, para que as diferentes partes interagentes, i e j , se separem.

Como o contato ii na água está relacionado com a segregação dos aminoácidos do tipo i na água, o grupo de Li partiu da suposição de que h_i estivesse relacionado com a hidrofobicidade do aminoácido i . De fato, encontrou-se que h_i tem grande correlação com as escalas hidrofóbicas publicadas na literatura. Mais ainda, Li e colaboradores verificaram que o termo linear da equação $M_{ij} = h_i + h_j - C_2(q_i - q_j)^2/2$ é dominante —o valor numérico de $h_i + h_j$ é bem maior do que o termo quadrático. Conseqüentemente, a diferença de energia entre diversas estruturas nativas é principalmente resultante do termo linear $(h_i + h_j)$, revelando que “a força hidrofóbica é a força dominante que direciona o folding de proteínas” (Li et al., 1997).

Este trabalho demonstrará que o **potencial hidrofóbico** ($E_{ij} = h_i + h_j$) é capaz de conduzir a cadeia a estruturas compactas, e eventualmente até à nativa, mas, no entanto, não fornece estabilidade suficiente. Faz-se, então, a suposição de que atributos geométricos dos monômeros da cadeia e especificidades interacionais entre eles desempenham o papel extra na **Eq.2.16**, resultando num **potencial estéreo-hidrofóbico** que, como será demonstrado, é eficiente em reproduzir aspectos importantes do enovelamento de cadeias tipo-proteínas (da Silva et al., 2001; Tarragó et al., no prelo).

2.10 Modelos Computacionais para o Folding

Este tópico tem como objetivo identificar procedimentos comuns aos modelos em rede que descrevem o enovelamento proteico, para, desta forma, situar o modelo utilizado neste trabalho num contexto mais amplo. Antecipa-se que não se intenciona descrever os diversos modelos apresentados na literatura; estes estão descritos em artigos de revisão como “Principles of Protein folding: a perspective from simple exact models” (Dill et al., 1995); “Theory of Protein Folding: the energy landscape perspective” (Onuchic et al., 1997); “Protein Folding Theory: from lattice to all-atom models” (Mirny & Shakhnovich, 2001); e “Protein Folding: a perspective for biology, medicine and biotechnology” (Yon, 2001), entre outros.

Qualquer modelo para o enovelamento de proteínas envolve grandes simplificações, tendo em vista a complexidade dos atributos estereoquímicos dos aminoácidos e, conseqüentemente, de suas interações —mesmo os mais sofisticados, ditos modelos de alta resolução. A representação mais elaborada de uma estrutura proteica é aquela que considera todos os seus átomos; no entanto, ainda não é possível tratar computacionalmente o **folding** através de tal representação, pois isto exigiria tempo de simulação impraticável. Os modelos computacionais para o enovelamento de proteínas utilizam representações altamente simplificadas, tanto da geometria como das interações presentes na cadeia polipeptídica. Estas limitações devem ser levadas em conta quando se comparam dados experimentais, relativos ao **folding** de proteínas, com dados teóricos obtidos computacionalmente.

Os **modelos computacionais** para o enovelamento de proteínas contemplam sempre:

1. um espaço conformacional que preserva a conectividade da cadeia e a representação do volume excluído;

Figura 2.17: Representação da cadeia polipeptídica na configuração nativa, utilizada nos primeiros trabalhos do grupo de Go; cada “conta” corresponde a um monômero, e as ligações covalentes entre eles são simbolizadas pelos traços em destaque (Go & Taketomi, 1978, Figura 1).

2. um conjunto de regras, geralmente através de uma função potencial, que descreve as interações microscópicas entre as partes constituintes da cadeia (Kaya & Chan, 2000).

O primeiro modelo computacional para o **folding** de proteínas foi desenvolvido na década de 70 (Taketomi et al., 1975) e é conhecido atualmente como **modelo de Go**. Neste, o espaço conformacional é uma rede bidimensional onde os monômeros da cadeia podem ocupar somente seus vértices e as ligações covalentes entre eles são representadas pelas arestas —o que o caracteriza como um modelo em rede. A **Figura 2.17** mostra a representação da cadeia polipeptídica no estado nativo utilizada nos primeiros trabalhos de Go e colaboradores (Taketomi et al., 1975; Go & Taketomi, 1978), onde cada “conta” corresponde a um monômero e as ligações covalentes entre eles correspondem aos traços em destaque.

De maneira geral, as regras adotadas nos **modelos de Go** são tais que:

- (a) cada vértice está desocupado ou ocupado por apenas um monômero —regra do volume excluído;
- (b) a probabilidade da cadeia adquirir uma nova configuração no processo de enovelamento é máxima toda vez que a energia do sistema diminui, e é proporcional ao peso de Boltzmann ($e^{-\Delta E/kT}$), toda vez que a energia do sistema aumenta —seguindo o algoritmo de Metropolis (Metropolis et al., 1953) e de acordo com a hipótese de que a estrutura nativa de uma macromolécula corresponde ao mínimo global de energia;
- (c) somente as interações entre pares presentes na estrutura nativa contribuem para a energia do sistema —esta regra implica, obviamente, o conhecimento prévio da configuração nativa (Go & Taketomi, 1978).

Todos os modelos deste tipo apresentam a função energia dependente da configuração em que a cadeia se encontra. A maioria dos modelos em rede tem como função energia uma Hamiltoniana, $H(\{r_k\})$, que considera somente os pares em contato; assim, de forma genérica, uma particular estrutura $\{r_k\}$ apresenta energia (de contato) dada por:

$$H(\{r_k\}) = \sum_{i,j} E_{i,j} \Delta(d_{i,j}), \quad (2.19)$$

onde E_{ij} é a energia de contato do par de monômeros i e j ; e $\Delta(d_{i,j}) = 1$, se os monômeros i e j são **primeiros vizinhos** —isto é, se eles ocupam vértices adjacentes na rede, sem serem contudo consecutivos ao longo da seqüência— e $\Delta(d_{ij}) = 0$, de outra forma —isto é, se os monômeros i e j não são primeiros vizinhos (Gutin et al., 1996; Pereira de Araújo, 1999; Micheletti et al., 1999). Desta maneira, a função energia é dada pela soma das energias dos pares de

monômeros em contato. Em alguns modelos, considera-se que certos contatos têm energia zero, e, por isto, são ditos contatos “não operativos”.

No primeiro artigo do grupo de Go —sobre simulação computacional do enovelamento de cadeias tipo-proteínas— foram avaliadas três regras distintas quanto às interações entre pares de monômeros em contato que devem ser consideradas operativas (neste caso, significando aquelas com energias atrativas): (i) somente as nativas; (ii) qualquer interação entre pares em contato; ou (iii) todas as nativas mais algumas não-nativas. Foi encontrado que, quando se considera apenas as interações presentes na estrutura nativa como as interações operativas, o processo de enovelamento é do tipo dois estados, e, portanto, está de acordo com o que acontece com as proteínas (pequenas e de um único domínio) (Taketomi et al., 1975). Por isto, atualmente se diz que um potencial é do tipo Go quando apenas as interações nativas são consideradas atrativas. No segundo artigo, este grupo questionou o papel das interações de curto e longo alcance no processo de enovelamento, concluindo que, quando as interações de curto alcance são mais energéticas do que as de longo alcance, o processo de enovelamento é mais rápido (Go & Taketomi, 1978).

Estudos recentes, utilizando o potencial de Go —inclusive um com bastante detalhamento, que representa cada aminoácido da proteína pelo seu carbono- α , C_α , e conseqüentemente, a proteína pelos seus átomos de C_α conectados entre si e podendo ocupar qualquer posição no espaço contínuo (Koga & Takada, 2001)— revelam-se bem sucedidos quanto à descrição da cinética do enovelamento; também revelam que a topologia do estado nativo é um fator determinante no mecanismo do **folding**. Isto sugere que as interações nativas desempenham um papel muito mais importante do que as não-nativas; no entanto, em termos de predição de estrutura a partir da seqüência, claramente o potencial de Go não é uma opção, pois presuppõe o conhecimento prévio da estrutura nativa (Alm & Baker, 1999).

Apesar das proteínas reais serem escritas com um alfabeto de 20 letras, a maioria dos modelos para o **folding** utiliza seqüências de aminoácidos escritas com alfabetos reduzidos (Romero et al., 1999). No entanto, sabe-se que modelos que se utilizam de alfabetos mais variados (isto é, com maior número de tipos de monômeros) apresentam maior possibilidade de as cadeias tipo-proteínas enovelarem-se em uma única conformação de menor energia, favorecendo o enovelamento mais rápido e uma maior estabilidade (Dill et al., 1995).

Diversos modelos em rede, inclusive o estudado neste trabalho, apresentam como primeiro passo a escolha prévia da estrutura alvo que será considerada como a nativa; esta estratégia é chamada por alguns autores de **folding** inverso. O segundo passo é a definição da seqüência de monômeros que leva à estrutura nativa. Os modelos diferenciam-se pelo número de tipos de monômeros considerados e pela energia de interação entre eles, mas praticamente todos os potenciais que determinam as interações entre dois monômeros i e j quaisquer, apresentados na literatura, seguem o princípio da segregação (Li et al., 1997) pela desigualdade, ou seja: $E_{ii} + E_{jj} < 2E_{ij}$ (Li et al., 1996; Micheletti et al., 1999; Rossi et al., 2000) indicando que além da força hidrofóbica —que, como será demonstrado mais adiante, satisfaz o princípio da segregação pela igualdade—, existe uma outra componente energética nestes potenciais. Neste sentido, os potenciais apresentados por Pereira de Araújo (1999) e por da Silva et al. (2001) são exceções, pois satisfazem o princípio da energia da segregação pela igualdade. No entanto, o potencial desenvolvido por da Silva et al. (2001), apesar de numericamente corresponder ao potencial hidrofóbico, é, de fato, um potencial estéreo-hidrofóbico, pois considera, juntamente com a energia hidrofóbica, um conjunto de restrições geométricas ou estéricas. Neste trabalho, exploram-se diversos aspectos deste potencial estéreo-hidrofóbico; demonstra-se sua eficiência quanto ao sucesso em enovelar cadeias tipo-proteínas e em garantir características cinéticas e termodinâmicas observadas nas proteínas globulares.

Muitos modelos, após definirem a composição da cadeia, geram aleatoriamente muitas seqüências e selecionam, por exemplo:

- (a) aquela que apresenta a estrutura nativa previamente escolhida como seu estado de mínima energia, não degenerado (isto é, só existe esta configuração com este mínimo de energia) (Socci et al., 1997);
- (b) aquela que maximiza, em valor absoluto, a diferença entre a energia da configuração nativa e a média das energias das configurações não nativas; ou, em outras palavras, aquela que otimiza o **gap** de energia ($Z = \frac{E_n - \langle E \rangle}{\sigma}$) entre o estado nativo (E_n) e a média das energias das conformações alternativas ($\langle E \rangle$), medido em unidades do desvio padrão (σ) da distribuição de energia dos estados não-nativos, conhecido como método Z-score (Abkevich et al., 1995; Gutin et al., 1996; Vendrusculo et al., 2000).

O presente trabalho utiliza-se da metodologia desenvolvida anteriormente por da Silva e colaboradores (da Silva et al., 2001; da Silva, 2002), a qual classifica a seqüência de uma determinada estrutura nativa (ou estrutura alvo) usando um alfabeto de dez letras, de acordo com uma sintaxe baseada na regra “mais (menos) hidrofóbico dentro (fora)” (Dill et al., 1995; Pereira de Araújo, 1999) e considerações geométricas. Esta sintaxe apresenta um aspecto importante que a diferencia das demais: a classificação da seqüência não depende apenas da estrutura nativa considerada (estrutura-alvo), mas também de aspectos geométricos mais abrangentes que foram identificados heurísticamente através de um conjunto representativo de diversas configurações maximamente compactas.

Em geral, os modelos em rede representam a estrutura nativa de uma cadeia tipo-proteína por uma estrutura maximamente compacta, fundamentados na constatação de que as proteínas globulares apresentam estruturas nativas altamente compactadas. No entanto, a escolha de tal configuração nativa parece

ser completamente arbitrária, apesar de contemplar a propriedade de ser maximamente compacta; na maioria das vezes, os modelos nada têm a dizer sobre o fato de as estruturas nativas das proteínas globulares apresentarem um número relativamente pequeno (em torno de 1000 tipos) de enovelamento (Pereira de Araújo, 1999). Tal observação conduziu à elaboração do conceito de **desenhabilidade** de uma estrutura nativa. Quando um número relativamente grande de seqüências diferentes conduzem a uma mesma estrutura nativa, diz-se que tal estrutura é bastante desenhável (Li et al., 1996; Micheletti et al., 1999).

Considerando a importância da configuração nativa, as 103346 configurações maximamente compactas —que se obtêm ao dobrar (de todas as maneiras possíveis) uma cadeia de 27 monômeros em uma rede cúbica— foram estudadas por alguns pesquisadores:

- (i) Govindarajan e Goldstein estudaram as 103346 configurações maximamente compactas quanto à otimização de um parâmetro energético, $\mathcal{R}$, definido pela razão entre a média das diferenças de energia entre os estados aleatórios e o nativo, e a largura da distribuição da energia das configurações aleatórias, o que, na prática, é equivalente ao método denominado Z-score. A estrutura “mais otimizada” é aquela que apresenta o maior valor de $\mathcal{R}$ e pode ser vista no lado direito da **Figura 2.18**, que mostra a distribuição do valor de $\mathcal{R}$ otimizado para o conjunto de todas as configurações compactas $3 \times 3 \times 3$ (Govindarajan & Goldstein, 1995).
- (ii) Li e colaboradores verificaram qual o número de seqüências HP, isto é, seqüências que apresentam somente dois tipos de monômeros, hidrofóbicos (H) e polares (P) (com as energias de contato entre os resíduos fixadas em $E_{HH} = -2, 3$, $E_{HP} = -1, 0$ e $E_{PP} = 0$), que podem determinar cada uma das 103346 configurações maximamente compactas. Eles encontraram que estas configurações apresentam desenhabilidades bastante diferenciadas,

Figura 2.18: Distribuição de $\mathcal{R}$, indicando a “otimização energética” para o conjunto das 103346 configurações compactas (CSA), de acordo com o potencial utilizado por Govindarajan & Goldstein (1995). A CSA mas “otimizada” está mostrada no lado direito da figura (Govindarajan & Goldstein, 1995, Figura 1).

sendo a estrutura mais desenhável aquela mostrada na **Figura 2.19** (Li et al., 1996).

- (iii) da Silva e colaboradores estudaram um subconjunto —constituído por cerca de vinte estruturas— das 103346 estruturas maximamente compactas, escolhido de forma a cobrir a distribuição dos diferentes valores de dois parâmetros topológicos —a saber, ordem de contato e número de estruturas secundárias na forma de grandes manivelas, denominadas estruturas STTS— e desenvolveram uma sintaxe que incluiu a classificação dos monômeros quanto ao nível de hidrofobicidade e um conjunto de regras, aplicáveis em todas as 103346 estruturas. Encontraram que estruturas alvos com ordens de contato relativamente pequenas ($< 0,35$) tendem a apresentar maior desenhabilidade e tempo característico de enovelamento pequeno (da Silva et al., 2001). O presente trabalho parte da metodolo-

Figura 2.19: Estrutura CSA mais “desenhável”, entre as 103346 configurações maximamente compactas, de acordo com o grupo de Li (Li et al., 1996, Figura 1).

gia desenvolvida por estes autores e explora características cinéticas e termodinâmicas de um subconjunto mais amplo, constituído por mais de 40 configurações maximamente compactas.

Apesar das simplificações envolvidas, acredita-se que os modelos em rede conseguem representar, de forma satisfatória, diversos aspectos do problema real do enovelamento proteico. Assim, é possível testar uma ampla variedade de concepções a respeito das proteínas, desde o desenho de suas seqüências, a cinética do processo de enovelamento, o cálculo de energia livre, entre outras (Vendrusculo et al., 2000). É importante observar, no entanto, que a habilidade de reproduzir propriedades observadas em proteínas é necessária, mas não suficiente, para a validação das características microscópicas presumidas nos modelos; por outro lado, se as propriedades de um modelo estão em desacordo com os dados experimentais, fica claramente indicada a sua deficiência (Kaya & Chan, 2000).

No próximo capítulo, apresentam-se os detalhes do modelo estéreo-hidrofóbico utilizado neste trabalho.

2.11 Referências Bibliográficas

Abkevich, V.I. et al. (1995) Impact of Local and Non-Local Interactions on Thermodynamics and Kinetics of Protein Folding. *J. Mol. Biol.*, **252**: 460-471

Alm, E. & Baker D. (1999) Matching Theory and Experiment in Protein Folding. *Curr. Opin. Struct. Biol.*, **9**: 189-196

Andrade, M. A. et al. (1999) Functional Classes in the Three Domains of Life. *J. Mol. Evol.*, **49**: 551-557

Anfinsen, C. B. (1973) Principles that Govern the Folding of Protein Chains. *Science* **181**: 223-230

Atkins, P. W. (1994) **The Elements of Physical Chemistry.** (4.ed.) Oxford University Press Inc., New York.

Bachmair, A. et al. (1986) In Vivo Half-Life of a Protein is a Function of its Amino-Terminal Residue. *Science*, **234**: 179-186

Baker, D. (2000) A Surprising Simplicity to Protein Folding. *Nature*, **405**: 39-42

Baldwin, R. L. (2002) Making a Network of Hydrophobic Clusters. *Science*, **295**: 1657-1658

Baldwin, R. L. & Rose, G. D. (1999) Is Protein Folding Hierarchic? I. Local Structure and Peptide Folding. *Trends Biochem. Sci.*, **24**:26-33

Berg, M. A. et al. (2000) Prospects in Computational Molecular Medicine: A Millennial Mega-Project on Peptide Folding. *J. Mol. Struct. (Theochem)*, **500**: 5-58

Bonneau, R. et al.(2002) Contact Order and Ab Initio Protein Structure Prediction. *Protein Science*, **11**: 1937-44

Bork, P. et al. (1998) Predicting Function: From Genes to Genomes and Back. *J. Mol. Biol.*, **283**: 707-725

Branden, C. & Tooze, J. (1999) **Introduction to Protein Structure.** (2.ed.) Garland Publishing, New York

Chagas, A. P. (1999) **Termodinâmica Química: Fundamentos, Métodos e Aplicações**. Editora da UNICAMP, Campinas

Chan, H. S. (1998) Matching Speed and Locality. *Nature*, **392**: 761-763

Chan, H. S. (2000) Modeling Protein Density of States: Additive Hydrophobic Effects are Insufficient for Calorimetric Two-State Cooperativity. *Proteins: Struct. Funct. Genet.*, **40**: 543-571

Chang, R. (2000) **Physical Chemistry for the Chemical and Biological Sciences**. University Science Books, Sausalito, CA.

Daune, M. (1999) **Molecular Biophysics: Structures in Motion**. Oxford University Press Inc., New York

Devos, D. & Valencia, A. (2000) Practical Limits of Function Prediction. *Proteins: Struct. Funct. Genet.*, **41**: 98-107

Dill, K. A. et al. (1995) Principles of Protein Folding: A Perspective from Simple Exact Models. *Protein Sci.*, **4**: 561-602

Ferguson, N. et al. (1999) Rapid Folding with and without Populated Intermediates in the Homologous Four-Helix Proteins Im7 and Im9. *J. Mol. Biol.*, **286**: 1597-1608

Ferrara, F. & Caffisch, A. (2001) Native Topology or Specific Interactions: What is More Important for Protein Folding? *J. Mol. Biol.*, **306**: 837-850

Fersht, A. R. (1995) Optimization of Rates of Protein Folding: The Nucleation-Condensation Mechanism and its Implications. *Proc. Natl. Acad. Sci. USA*, **92**: 10869-10873

Fersht, A. (2000) **Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding**. (3.ed.) W.H. Freeman and Company, New York.

Garrett, R. H. & Grisham, C. M. (1995) **Biochemistry**. Saunders College Publishing, USA.

Go, N. & Taketomi, H. (1978) Respective Roles of Short- and Long-Range

Interactions in Protein Folding. *Proc. Natl. Acad. Sci. USA.*, **75**: 559-563

Goldberg, S.(1998) **Descomplicando ... Bioquímica**. (2.ed.) Artmed, Porto Alegre.

Gómez, J. et al. (1995) The Heat Capacity of Proteins. *Proteins: Struct. Funct. Genet.*, **22**: 404-412

Govindarajan, S. & Goldstein, R. (1995) Searching for Foldable Protein Structures Using Optimized Energy Functions. *Biopolymers*, **36**: 43-51

Greighton, T. E. (2000) Protein Folding Coupled to Disulphide-Bond Formation. In **Mechanisms of Protein Folding**. (2.ed.) (R. H. Pain Editor) Oxford University Press Inc., New York.

Grob β ,M. (1996) Linguistic Analysis of Protein Folding. *FEBS Lett.*, **390**: 249-252

Gromiha, M. M. & Selvaraj, S. (2001) Comparison between Long-range Interactions and Contact Order in Determining the Folding Rate of Two-State Proteins: Applications of Long-Range Order to Folding Rate Prediction. *J. Mol. Biol.*, **310**: 27-32

Gutin, A. M. et al. (1996) Chain Length Scaling of Protein Folding Time. *Phys. Rev. Lett.*, **77**: 5433-5436

Hummer, G. et al. (1998) The Pressure Dependence of Hydrophobic Interactions is Consistent with the Observed Pressure Denaturation of Proteins. *Proc. Natl. Acad. Sci. USA*, **95**: 1552-1555

Jaenicke, R. (1999) Stability and Folding of Domain Proteins. *Progress in Biophysics & Molecular Biology*, **71**: 155-241

Kaya, H. & Chan, H. S. (2000) Polymer Principles of Protein Calorimetric Two-State Cooperativity. *Proteins: Struct. Funct. Genet.*, **40**: 637-661

Kamtekar, S. et al. (1993) Protein Design by Binary Patterning of Polar and Nonpolar Amino Acids. *Science*, **262**: 1680-85

Kauzmann, W. (1959) Some Factors in the Interpretation of Protein Denat-

uration. *Adv. Protein Chem.*, **14**: 1-63

Koga, N. & Takada, S. (2001) Roles of Native Topology and Chain-Length Scaling in Protein Folding: a simulation study with a Go-like model. *J. Mol. Biol.*, **313**: 171-180

Lattman, E. E. & Rose, G. D. (1993) Protein Folding —What's the Question? *Proc. Natl. Acad. Sci. USA*, **90**, 439-441

Lee, B. & Richards, F. M. (1971) The Interpretation of Protein Structures: Estimation of Static Accessibility. *J. Mol. Biol.*, **55**: 379-400

Li, H. et al. (1996) Emergence of Preferred Structures in a Simple Model of Protein Folding. *Science*, **273**: 666-669

Li, H. et al. (1997) Nature of Driving Force for Protein Folding: A Result from Analyzing the Statistical Potential. *Phys.Rev.Lett.*, **79**: 765-768

Makhatadze, G.I. & Privalov, P.L. (1990) Heat Capacity of Proteins: I. Partial Molar Heat Capacity of Individual Amino Acid Residues in Aqueous Solutions —Hydration Effect. *J. Mol. Biol.*, **213**: 375-384

Metropolis, N. et al. (1953) Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.*, **21**: 1087-1092

Micheletti, C. et al. (1999) A Comparative Study of Existing and New Desing Techniques for Protein Models. *J. Chem. Phys.*, **110**: 9730-9738

Miyazawa, S. & Jernigan, R.L. (1985) Estimation of Effective Interresidue Contact Energies from Protein Crystal Structures. *Macromolecules*, **18**: 534-552

Mirny, L. A. & Shakhnovich, E. (1996) How to Derive a Protein Folding Potential? A New Approach to an Old Problem. *J. Mol. Biol.*, **264**: 1164-79

Mirny, L. & Shakhnovich, E. (2001) Protein Folding Theory: From Lattice to All-Atom Models. *Annu. Rev. Biophys. Biomol. Struct.*, **30**: 361-396

Mohana-Borges, R. et al. (1999) Folding of a Pressure-Denatured Model Protein. *Proc. Natl. Acad. Sci. USA*, **96**: 7888-7893

Murzin, A. G. et al. (1995) SCOP: A Structural Classification of Proteins Database for the Investigation of Sequences and Structures. *J. Mol. Biol.*, **247**: 536-540

Nelson, J.C. et al. (1997) Solvophobicity Driven Folding of Nonbiological Oligomers. *Science*, **277**: 1793-1796

Nelson, D. L. & Cox, M. M. (2000) **Lehninger Principles of Biochemistry**. (3.ed.) Worth Publishers, New York.

Onuchic, J. N. et al. (1997) Theory of Protein Folding: The Energy Landscape Perspective. *Annu. Rev. Phys. Chem.*, **48**: 545-600

Okamoto, Y. (1998) Protein Folding Problem as Studied by New Simulation Algorithms. *Recent Res. Devel. in Pure & Applied Chem.*, **2**: 1-23

Pereira de Araújo, A. F. (1999) Folding Protein Models with a Simple Hydrophobic Energy Function: The Fundamental Importance of Monomer Inside/Outside Segregation. *Proc. Natl. Acad. Sci.*, **96**: 12482-12487

Plaxco, K. et al. (1997) A Comparison of the Folding Kinetics and Thermodynamics of Two Homologous Fibronectin Type III Modules. *J. Mol. Biol.*, **270**, 763-770

Plaxco, K. W. et al. (1998) Contact Order, Transition State Placement and the Folding Rates of Single Domain Proteins. *J. Mol. Biol.*, **277**: 985-994

Plaxco et al. (2000) Topology, Stability, Sequence, and Length: Defining the Determinants of Two-State Protein Folding Kinetics. *Biochemistry*, **39**: 11177-11183

Prigogine, I. (1996) **O Fim das Certezas**. (2.ed.) Fundação Editora da UNESP, São Paulo.

Privalov, P. (1997) Thermodynamics of Protein Folding. *J. Chem. Thermodynamics*, **29**: 447-474

Privalov, P. L. (1992) Physical Basis of the Stability of the Folded Conformations of Proteins. In: **Protein Folding** (Thomas E. Creighton Editor). W. H.

Freeman and Company, USA.

Privalov, P.L. & Khechinashvili, N. N. (1974) A Thermodynamic Approach to the Problem of Stabilization of Globular Protein Structure: A Calorimetric Study. *J. Mol. Biol.*, **86**: 665-684

Privalov, P.L. & Makhatadze, G.I. (1992) Contribution of Hydration and Non-Covalent Interactions to the Heat Capacity Effect on Protein Unfolding. *J. Mol. Biol.*, **224**: 715-723

do Reis, L. M. (1995) Folding de Proteínas: Simulação Monte Carlo de um Modelo em Rede com Potencial Randômico. Dissertação de Mestrado, FFCLRP, Departamento de Física e Matemática, Universidade de São Paulo.

Richard, M. et al. (1997) Proteins I: Composition and Structure. In **Text-book of Biochemistry with clinical correlations**. (4.ed.) Wiley-Liss, New York.

Romero, P. et al. (1999) Folding Minimal Sequences: The Lower Bound for Sequence Complexity of Globular Proteins. *FEBS Letters*, **462**: 363-367

Rose et al. (1985) Hydrophobicity of Amino Acid Residues in Globular Proteins. *Science*, **229**: 834-838

Roseman, M. A. (1988) Hydrophilicity of Polar Amino Acid Side-Chains is Markedly Reduced by Flanking Peptide Bonds. *J. Mol. Biol.* **200**: 513-522

Rossi, A. et al. (2000) A Novel Iterative Strategy for Protein Design. *J. Chem. Phys.*, **112**: 2050-2055

Schindler, T. & Schmid, F. X. (1996) Thermodynamic Properties of an Extremely Rapid Folding Reaction. *Biochemistry*, **35**: 16833-16842

Schonbrun, J. et al. (2002) Protein Structure Prediction in 2002. *Curr. Opin. Struct. Biol.*, **12**: 348-54

Schönbrunner, N. et al. (1997) Folding of the Disulfide-Bonded β -Sheet Protein Tendamistat: Rapid Two-State Folding Without Hydrophobic Collapse. *J. Mol. Biol.*, **268**: 526-538

Shortle, D. (1989) Probing the Determinants of Protein Folding and Stability with Amino Acid Substitutions. *J. Biol. Chem.*, **264**: 5315-5318

Schulz, G. E. & Schirmer, R. H. (1990) **Principles of Protein Structure**. (9.ed.) Springer-Verlang, New York.

da Silva, R. A. et al. (2001) Deterministic Folding: The role of Entropic Forces and Steric Specificities. *J. Chem. Phys.*, **114**: 4235-4242

da Silva, R. A. (2002) **Enovelamento Determinístico de Macromoléculas**. Tese de Doutorado, FFCLRP, Departamento de Física e Matemática, Universidade de São Paulo.

Socci, N.D. et al. (1997) Exploring the Protein Folding Funnel Landscape. *Physica D*, **107**: 366-382

Soulages, J. L. (1998) Chemical Denaturation: Potential Impact of Undetected Intermediates in the Free Energy of Unfolding and m-Values Obtained from a Two-State Assumption. *Biophysical Journal*, **75**:484-492

Taketomi, H. et al. (1975) Studies on Protein Folding, Unfolding and Fluctuations by Computer Simulation. *Int. J. Peptide Protein Res.*, **7**: 445-459

Tarragó, M. E. P. et al. (2003) Steric Constraints as Folding Coadjuvant. *Phys. Rev. E*, no prelo.

Thomas, P. D. & Dill, K. A. (1996) Statistical Potentials Extracted from Protein Structures: How Accurate are They? *J. Mol. Biol.*, **257**: 457-469

Thornton, J. M. et al. (1999) Protein Folds, Functions and Evolution. *J. Mol. Biol.*, **293**: 333-342

Thornton, J. M. (1992) Protein Structures: The End Point of the Folding Pathway. In: **Protein Folding** (Thomas E. Greighton Editor). W. H. Freeman and Company. USA

Vendruscolo, M. et al. (2000) Comparison of Two Optimization Methods to Derive Energy Parameters for Protein Folding: Perceptron and Z score. *Proteins: Struct. Funct. Genet.*, **41**: 192-201

van Osdol et al. (1991) Multifrequency Calorimetry of the Folding/Unfolding Transition of Cytochrome C. *Biophysical Journal*, **59**: 48-54

Yon, J. M. (2001) Protein Folding: A Perspective for Biology, Medicine and Biotechnology. *Braz. J. Med. Biol. Res.*, **34**: 419-435

Zhang, J. et al. (1995) NMR Study of the Cold , Heat, and Pressure Unfolding of Ribonuclease A. *Biochemistry*, **34**: 8631-8641

Zhou, H. & Zhou, Y. (2002) Folding Rate Prediction Using Total Contact Distance. *Biophysical Journal*, **82**: 458-463

Capítulo 3

O Modelo Estéreo-Hidrofóbico

3.1 Introdução

Como já mencionado no capítulo anterior, estudos teóricos da matriz de Miyazawa e Jernigan (matriz MJ) —que representa o potencial estatístico de contato, ou, em outras palavras, a energia de contato entre os pares de aminoácidos (Miyazawa & Jernigan, 1985)— consideram as forças hidrofóbicas ou entrópicas como as dominantes no processo que direciona a cadeia para a conformação nativa (Li et al., 1997). Foi demonstrado que cada elemento M_{ij} da matriz MJ pode ser reproduzida pela simples equação $M_{ij} = h_i + h_j - C(q_i - q_j)^2$ (**Eq.2.16**), onde h_i é o nível hidrofóbico do resíduo i , e o termo quadrático, $-C(q_i - q_j)^2$, está diretamente relacionado com a energia de dissociação dos resíduos i e j . No entanto, a energia de interação entre i e j pode ser aproximada pela soma $h_i + h_j$, uma vez que este termo linear é dominante sobre o termo quadrático na equação acima (Li et al., 1997). Isto significa que, a partir do momento em que se tem uma escala hidrofóbica $\{h_i\}$ adequada, um potencial intra-cadeia efetivo do tipo hidrofóbico pode ser escrito como $E_{ij} = h_i + h_j$. Observa-se que, quando a energia é escrita desta forma, o princípio da segregação de energia, $2E_{ij} - E_{ii} - E_{jj} > 0$, é satisfeito marginalmente através da igualdade, ou seja,

$2E_{ij} - E_{ii} - E_{jj} = 0$; note que, nesse caso, é possível escrever E_{ij} como a média de E_{ii} e E_{jj} , ou seja, $E_{ij} = (E_{ii} + E_{jj})/2$.

Um dos objetivos deste trabalho foi testar o potencial hidrofóbico, dado por $E_{ij} = h_i + h_j$, através de simulações Monte Carlo e encontrou-se que, apesar deste ser eficiente em conduzir a cadeia à conformação nativa, ele não reproduz diversas propriedades esperadas de proteínas, como a estabilidade do estado nativo. Adotou-se, então, um potencial estéreo-hidrofóbico, $E_{ij}^* = h_i + h_j + c_{ij}$, que, concomitantemente com a energia hidrofóbica, considera um conjunto adequado de regras $\{c_{ij}\}$ do tipo restrições de caroço duro (hard-core). No próximo capítulo, demonstra-se que as restrições estéricas desempenham papel significativo no processo de enovelamento, afetando os caminhos para o folding e ajudando a estabilizar a cadeia no estado nativo, como introduzido em trabalhos anteriores (da Silva et al., 2001; da Silva, 2002).

3.2 O Design da Estrutura Nativa

Nesta seção, são definidos alguns parâmetros topológicos da estrutura nativa de uma molécula tipo-proteína e os critérios de classificação da sua sequência de monômeros.

3.2.1 Topologia do estado nativo

Uma proteína globular, no seu estado funcional, apresenta-se na forma de um glóbulo altamente compactado, e seu correspondente nos modelos computacionais é, em geral, representada por uma estrutura maximamente compacta. Existem 103346 maneiras diferentes de dobrar-se, na forma maximamente compacta, uma cadeia de 27 contas em um espaço configuracional tipo rede tridimensional. Cada uma destas estruturas define um cubo com dimensões $3 \times 3 \times 3$ e é chamada configuração Compact Self-Avoiding ou simplesmente **CSA**. No

presente trabalho, a cadeia polipeptídica é representada por uma sequência de 27 monômeros, ou contas, que ocupam os vértices de uma rede cúbica. Em princípio, a estrutura nativa pode ser representada por qualquer uma das 103346 CSA possíveis para uma cadeia com 27 monômeros.

Utilizando-se da metodologia desenvolvida em trabalhos anteriores (Marques dos Reis, 1995), geraram-se as 103346 configurações maximamente compactas que podem abrigar uma cadeia de 27 monômeros em rede cúbica. Inspirado em diversos trabalhos experimentais que relacionam propriedades cinéticas com parâmetros topológicos do estado nativo de proteínas (Plaxco et al., 1998; Alm & Baker, 1999; Zhou & Zhou, 2002), desenvolveu-se um programa que conta o número n_c de pares de monômeros em contato topológico —isto é, quais os pares de monômeros que ocupam posições vizinhas na rede, sem estarem conectados ao longo da sequência— e quantos monômeros $|i - j|$ existem entre eles ao longo da sequência, para, então, através das definições apresentadas anteriormente, calcular, para cada CSA:

- (a) o parâmetro ordem de contato, χ , que representa a separação média ao longo da sequência entre dois monômeros em contato (vide subseção **2.8.1**), dado por:

$$\chi = \frac{1}{n_c} \sum_{\substack{\{n_c\} \\ |i-j| > l_{cut}}} \frac{|i - j|}{n_r}; \quad (3.1)$$

- (b) o parâmetro ordem de longo alcance, χ_{LR} , que quantifica o número relativo de contatos de longo alcance (vide subseção **2.8.2**), dado por:

$$\chi_{LR} = \frac{n_c}{n_r}, \quad (3.2)$$

sendo $n_r = 27$, o número total de resíduos da cadeia. Considera-se, neste trabalho, que dois monômeros i e j constituem um contato de longo alcance quando a distância entre eles ao longo da cadeia é $|i - j| \geq 12$.

Quando a cadeia tipo-proteína está enovelada na configuração considerada nativa, de acordo com metodologia desenvolvida recentemente (da Silva et al., 2001), cada monômero assume um dos três atributos geométricos: “*S*”, se as suas duas ligações covalentes estão alinhadas (Straight line); “*T*”, se as duas ligações covalentes formam um ângulo reto (Turn); e, “*E*”, se o monômero for da ponta da cadeia (End), como se mostra nas **Figuras 3.1** e **3.2**. Particularmente, uma estrutura compacta do tipo CSA pode apresentar um número ϕ de estruturas tipo secundárias, chamadas “estruturas *STTS*”, onde cada uma delas contém quatro monômeros consecutivos na forma *S*, *T*, *T* e *S*, todos no mesmo plano, formando uma “manivela”. A **Figura 3.1** mostra uma configuração compacta (CSA 870) com o maior número ϕ de estruturas *STTS* possível ($\phi = 8$); e, em contraposição, representa-se na **Figura 3.2** uma configuração (CSA 99865) que não apresenta nenhuma estrutura do tipo *STTS* ($\phi = 0$). Também, calculou-se, para cada uma das 103346 CSA, o número ϕ de estruturas secundárias tipo *STTS*; o capítulo seguinte, na seção 4.1, mostrará como todas as 103346 CSA se distribuem segundo χ , χ_{LR} e ϕ .

3.2.2 Design da estrutura nativa

Empregou-se a metodologia desenvolvida por da Silva e colaboradores (2001) para a classificação das seqüências das cadeias tipo-proteínas. Inicia-se escolhendo-se uma estrutura maximamente compacta (CSA) que representará a estrutura nativa. Nesta, os monômeros são identificados por um índice, i , que representa o número de seus contatos com o meio, definindo um conjunto $\{R_i\}$, onde $i = 0, 1, 2$ e 3 ; o índice i também pode ser interpretado como sendo proporcional à área da superfície acessível ao solvente na estrutura nativa.

Qualquer configuração maximamente compacta do tipo $3 \times 3 \times 3$, como a mostrada na **Figura 3.3**, por exemplo, apresenta:

Figura 3.1: Configuração CSA 870, que apresenta menor ordem de contato ($\chi = 0,2381$), baixo valor para ordem de longo alcance ($\chi_{LR} = 0,2143$) e maior número de estruturas *STTS* ($\phi = 8$). Cada monômero assume um dos três atributos geométricos: “S”, se as suas duas ligações covalentes estão alinhadas (Straight line); “T”, se as suas duas ligações covalentes formam um ângulo reto (Turn); e, “E”, se o monômero for da ponta da cadeia (End).

Figura 3.2: Configuração CSA 99865, que, dentre todas as 103346 configurações CSA, apresenta maior valor de ordem de contato ($\chi = 0,4947$), maior valor de ordem de longo alcance ($\chi_{LR} = 0,7143$) e menor número de estruturas *STTS* ($\phi = 0$). Cada monômero é classificado de acordo com seus atributos geométricos: “S”, se as suas duas ligações covalentes estão alinhadas (Straight line); “T”, se as suas duas ligações covalentes formam um ângulo reto (Turn); e, “E”, se o monômero for da ponta da cadeia (End).

Figura 3.3: Exemplo de uma configuração maximamente compacta (CSA), onde se mostram as posições das quatro classes de monômeros, de acordo com o número de contatos que estabelecem com o meio solvente.

- (a) um monômero $r(0)$ no centro do cubo;
- (b) seis monômeros pertencentes ao grupo $R_1 = \{r(1, j)\}$ que ocupam os centros das faces do cubo;
- (c) doze monômeros pertencentes ao grupo $R_2 = \{r(2, j)\}$ que ocupam os meios das arestas;
- (d) oito monômeros do tipo $r(3)$ que ocupam os vértices do cubo.

A seguir, apresenta-se a regra, ou sintaxe, para se estabelecer a sequência da cadeia tipo-proteína, dada a sua estrutura nativa. Os elementos de R_1 e R_2 são especificados de acordo com a vizinhança topológica, ou seja, identificando-se quais os monômeros que ocupam as posições vizinhas na rede, sem serem vizinhos ao longo da cadeia. Quanto aos elementos do grupo R_1 : será um

monômero $r(1, 1)$, se o número de vizinhos topológicos na forma S é maior ou igual ao número de vizinhos topológicos na forma T ; de outra maneira, será $r(1, 2)$.

Quanto aos elementos do grupo R_2 , será um monômero:

1. $r(2, 1)$, se seus dois vizinhos topológicos são monômeros do tipo $r(1, 1)$ e $r(1, 1)$;
2. $r(2, 2)$, se seus dois vizinhos topológicos são monômeros do tipo $r(1, 1)$ e $r(1, 2)$;
3. $r(2, 3)$, se seus dois vizinhos topológicos são monômeros do tipo $r(1, 1)$ e $r(3)$;
4. $r(2, 4)$, se seus dois vizinhos topológicos são $r(1, 2)$ e $r(1, 2)$, ou $r(1, 2)$ e $r(3)$, estando $r(3)$ na forma T ;
5. $r(2, 5)$, se seus dois vizinhos topológicos são $r(1, 2)$ e $r(3)$, estando $r(3)$ na forma E ; e finalmente,
6. $r(2, 6)$, se seus dois vizinhos topológicos são $r(3)$.

Nas simulações computacionais, adotou-se representar a classificação dos monômeros por letras, de tal forma que:

- (a) R_0 contém o elemento $r(0)$, simbolizado por r ;
- (b) R_1 contém os elementos $r(1, 1)$ e $r(1, 2)$; representados pelas letras a e h , respectivamente;
- (c) R_2 contém os elementos $r(2, 1)$, $r(2, 2)$, $r(2, 3)$, $r(2, 4)$, $r(2, 5)$ e $r(2, 6)$, correspondentes às letras b , g , f , i , e , e d , respectivamente; e, finalmente,
- (d) R_3 contém o elemento $r(3)$, simbolizado pela letra c .

Supõe-se neste modelo que a hidrofobicidade de um resíduo é inversamente proporcional ao número de contatos que o mesmo estabelece com o solvente, quando a cadeia se encontra na estrutura nativa. Assim, aos dez monômeros acima, são atribuídas as seguintes hidrofobicidades, dadas em unidades de kT :

$$R_0: r(0) \rightarrow -2, 1$$

$$R_1: r(1, 1) \text{ e } r(1, 2) \rightarrow (-2, 0 \text{ e } -1, 9)$$

$$R_2: r(2, 1), r(2, 2), \dots, r(2, 6) \rightarrow (-1, 2; -1, 1; -1, 0; -1, 0; -0, 9 \text{ e } -0, 8)$$

$$R_3: r(3) \rightarrow +0, 8.$$

No presente trabalho, com esta sintaxe em mãos, escolhe-se a estrutura alvo que se deseja estudar, de acordo com suas características topológicas —como ordem de contato, ordem de longo alcance e número de estruturas do tipo *STTS*— e, então, determina-se a seqüência de sua cadeia. Isto significa que as características geométricas da configuração nativa estão, de certa forma, codificadas na sua seqüência, o que se acredita ocorrer com as proteínas reais.

3.3 Energia Hidrofóbica

No modelo estudado neste trabalho, assume-se que o efeito hidrofóbico gera a força dominante que direciona o enovelamento das proteínas (Li et al., 1997). Assim, uma questão básica a ser respondida é: até que ponto o perfil hidrofóbico da seqüência de aminoácidos de uma proteína é responsável pelo processo que leva a cadeia a se conformar na estrutura nativa?

3.3.1 Modelo hidrofóbico

Para estudar e tentar responder a esta questão, é conveniente, pelo menos no início, isolar o efeito hidrofóbico dos demais fatores possíveis. Para tanto, considera-se que os N resíduos, que constituem a cadeia polipeptídica, diferenciam-se quimicamente apenas pelos seus níveis de hidrofobicidade, que são medidos pelas

energias hidrofóbicas específicas; então, o conjunto dos L tipos de monômeros possíveis é representado por $\{\gamma_k\} = \{\gamma_1; \gamma_2; \dots \gamma_L\}$. Considerando-se, ainda, que a sequência desta cadeia proteica seja representada por $\{s_k\} = \{s_1; s_2; \dots s_N\}$, pode-se associar a cada unidade s_k um nível de hidrofobicidade γ_k de tal forma que $\gamma_k = \gamma_k(s_k)$.

É importante também considerar que as interações cadeia-solvente são de curto alcance (ou seja, só são diferentes de zero quando a distância entre os monômeros e as moléculas do solvente for muito pequena) e governadas pela natureza polar ou hidrofóbica dos resíduos, e que as interações intra-resíduos e solvente-solvente são do tipo potencial de caroço duro (interações do tipo *hard-core*).

É bem estabelecido que a energia hidrofóbica, $H = H(\{s_k\}, \{n_k\})$, da sequência, $\{s_k\}$, é linearmente dependente da área da superfície acessível ao solvente de cada grupamento atômico (Lee & Richards, 1971; Chothia, 1974), onde H pode ser expressa pelo conjunto do número de moléculas do solvente, $\{n_k\}$, em contato com cada unidade k da molécula proteica; isto é:

$$H = \Upsilon_0 + \sum_{\{k\}} n_k \gamma_k, \quad (3.3)$$

onde Υ_0 é uma constante e o somatório é sobre todas as N unidades da cadeia. Faz-se γ_k ser positivo para os aminoácidos hidrofóbicos, pois, assim, H aumenta quando aumenta o número de contatos cadeia-solvente; e, ao contrário, faz-se γ_k ser negativo para os aminoácidos polares. As coordenadas de cada conformação particular da cadeia $\{r_k\}$ determina o conjunto dos números $\{n_k\}$, que representam os contatos cadeia-solvente, ao mesmo tempo que especificam os contatos intracadeias, isto é, os contatos entre a unidade k e as outras m unidades da cadeia, que serão chamadas $p, q, \dots t$. Note-se que n_k conta o número de contatos cadeia-solvente, enquanto m indica o número de contatos intracadeia.

Para analisar a variação do número de contatos da cadeia com a água, devido à alteração conformacional, considera-se que:

- (a) inicialmente, a cadeia se encontra numa estrutura $\{\mathbf{r}_k\}_0$ com energia correspondente $H = H_0$, como mostra a **Figura 3.4 (a)**.
- (b) Supõe-se, então, que um resíduo k particular é removido de seu ambiente original, alterando a energia da cadeia para $H' = H_0 - n_k \gamma_k + (n_p \gamma_p + n_q \gamma_q + \dots + n_t \gamma_t)$; pois, conseqüentemente, são removidos n_k contatos entre o monômero k e as moléculas de solvente, enquanto novos contatos, $\mathcal{N} = (n_p + n_q + \dots + n_t)$, entre as moléculas de água e as m unidades $p, q, \dots, t$, são estabelecidos, ocupando o espaço deixado pela unidade k , como é esquematizado na **Figura 3.4 (b)**.
- (c) Agora, supõe-se que o resíduo k é recolocado novamente na cadeia, em um sítio diferente, estabelecendo n_k'' contatos novos com as moléculas do solvente, e removendo $\mathcal{N}'' = (n_u + n_v + \dots + n_w)$ contatos entre as moléculas do solvente e as m'' unidades $u, v, \dots, w$; isto conduz a cadeia para um estado energético dado por $H'' = H' + n_k'' \gamma_k - (n_u \gamma_u + n_v \gamma_v + \dots + n_w \gamma_w)$; ver **Figura 3.5**.

A variação total de energia é $\Delta H = H'' - H_0$ entre as configurações final (H'') e original (H_0), resultando em:

$$\Delta H = (n_k'' - n_k) \gamma_k + (n_p \gamma_p + \dots + n_t \gamma_t) - (n_u \gamma_u + \dots + n_w \gamma_w), \quad (3.4)$$

onde as duas somas $(n_p \gamma_p + n_q \gamma_q + \dots + n_t \gamma_t)$ e $(n_u \gamma_u + n_v \gamma_v + \dots + n_w \gamma_w)$ têm, respectivamente, m e m'' termos. Observar que a energia de ligação covalente não foi alterada; de fato, este processo virtual de remover e retornar uma ou mais unidades ao sistema corresponde à geração de uma nova configuração da cadeia, em que se conservam todas as ligações covalentes.

Figura 3.4: Representação esquemática dos contatos que os monômeros p , q e t estabelecem com e, virtualmente, sem o resíduo k na cadeia. Em (a), o resíduo k estabelece n_k contatos com as moléculas de água e, também, estabelece contatos com os resíduos p , q e t . Em (b), tem-se m novos contatos entre as unidades p , q e t e as moléculas de água, pois o resíduo k foi removido.

Figura 3.5: Representação esquemática dos contatos que os monômeros u , v e z estabelecem com o resíduo k na nova configuração. O resíduo k tem, agora, n_k'' contatos com as moléculas do solvente e m'' contatos com os resíduos u , v e z .

3.3.2 Equivalência entre energia hidrofóbica e energia de contato entre os monômeros

A energia hidrofóbica envolve essencialmente a interação entre o solvente e os elementos da cadeia proteica. Sua estimativa é baseada em dados experimentais (Nozaki & Tanford, 1971) envolvendo variações entálpicas e entrópicas, sendo, assim, importante no processo dinâmico do folding.

A energia hidrofóbica do sistema, tratada anteriormente (na subseção **3.3.1**), pode ser escrita, em certas condições, através do formalismo da energia de contato entre os monômeros; o objetivo deste tópico é identificar quando isto é possível. A energia de contato entre os monômeros é:

$$E = \varepsilon_0 + \sum_{\{i,j\}} \varepsilon_{i,j} \Delta(d_{ij}), \quad (3.5)$$

onde ε_0 é uma constante e $\varepsilon_{i,j}$ é a energia de contato entre os monômeros i e j ; $\Delta(d_{ij}) = 1$, se a distância, d_{ij} , entre os monômeros (por exemplo, entre seus centros de massa) satisfaz $d_{ij} \leq D$, e $\Delta(d_{ij}) = 0$ de outra forma; D é a distância de corte previamente definida. Para a mesma configuração inicial, $\{\mathbf{r}_k\}_0$, a energia da cadeia é $E = E_0$; realizando-se exatamente as mesmas transformações configuracionais anteriores (como em **3.3.1**), isto é, removendo o monômero k de seu ambiente original, onde ele faz m contatos com os monômeros $p, q, \dots, t$, a energia se altera para $E' = E_0 - (\varepsilon_{kp} + \varepsilon_{kq} + \dots + \varepsilon_{kt})$, e retornando este monômero novamente para o sistema, em um sítio alternativo, ele estabelece m'' novos contatos com os monômeros $u, v, \dots, w$, correspondendo à energia $E'' = E' + (\varepsilon_{ku} + \varepsilon_{kv} + \dots + \varepsilon_{kw})$. A variação de energia resultante é $\Delta E = E'' - E_0$, e pode ser escrita como:

$$\Delta E = -(\varepsilon_{kp} + \varepsilon_{kq} + \dots + \varepsilon_{kt}) + (\varepsilon_{ku} + \varepsilon_{kv} + \dots + \varepsilon_{kw}), \quad (3.6)$$

onde a primeira e a segunda soma têm, respectivamente, m e m'' termos. A

energia de contato, $\varepsilon_{i,j}$, é vista como um potencial de campo médio, pois ela não inclui detalhes locais das interações cadeia-solvente. Então, se $\varepsilon_{i,j}$ representa somente a interação de energia hidrofóbica, ela deve ser função somente de γ_i e γ_j , ou seja, $\varepsilon_{ij} = f(\gamma_i, \gamma_j)$. No entanto, a representação de $f(\gamma_i, \gamma_j)$ por uma série em γ_i e γ_j deve ser truncada nos termos lineares, como sugere a **Eq.2.16**, resultando em $\varepsilon_{i,j} = A + B\gamma_i + C\gamma_j$, onde A , B e C são constantes. A simetria $\varepsilon_{i,j} = \varepsilon_{j,i}$ implica $B = C$; então, fixando $B = -\eta$ e $A = 0$, a **Eq.3.6** pode ser reescrita como:

$$\Delta E = (m - m'')\eta\gamma_k + \eta(\gamma_p + \dots \gamma_t) - \eta(\gamma_u + \dots \gamma_w). \quad (3.7)$$

As duas representações da energia hidrofóbica, nas **Eq.3.4** e **Eq.3.7**, mostram notáveis diferenças: a primeira (**Eq.3.4**) especifica detalhes locais sobre as interações cadeia-solvente através dos conjuntos $(n_p, n_q, \dots n_t)$ e $(n_u, n_v, \dots n_w)$, enquanto a segunda (**Eq.3.7**) representa tais interações de maneira uniforme através de todo o sistema, como indica o parâmetro η . A discrepância entre as variações da energia configuracional deve ser creditada à diversidade de tamanhos e formas dos monômeros, que fica explícita na primeira formulação, mas é considerada em média na segunda.

Outra diferença importante é que na **Eq.3.4** o número de contatos cadeia-solvente $(n_p, n_q, \dots n_t)$ e $(n_u, n_v, \dots n_w)$ pode variar, predispondo o sistema para futuras mudanças configuracionais, sem, no entanto, mover partes da cadeia. Este tipo de movimento não é considerado na aproximação da energia de contato, mas pode ser importante na dinâmica da evolução configuracional.

Entretanto, faz-se a pergunta: será que existe um sistema no qual ambas as representações coincidem? A resposta é sim. As duas representações tornam-se exatamente equivalentes se o sistema (constituído pela cadeia mais solvente) é descrito através de um modelo em rede que apresenta um número de coordenação, Z , fixo, com todos os seu sítios ocupados, ou pelos monômeros da cadeia, ou

pelas moléculas do solvente. (O número de coordenação Z de um sítio da rede representa o número de ligações ou contatos possíveis para um monômero ou molécula do solvente, que ocupa tal sítio.) Para demonstrar esta equivalência, considera-se um modelo em rede apropriado no qual as N unidades da cadeia estão restritas a ocupar sítios consecutivos e diferentes de uma rede regular 3-D; todos os outros sítios da rede, chamados daqui para frente de sítios do solvente, são preenchidos por moléculas do solvente. As unidades da cadeia são consideradas todas do mesmo tamanho, representadas por contas com o mesmo volume $\mathcal{V}$, e cada sítio do solvente contém um certo número fixo η de moléculas de água que é o suficiente para ajustar-se em $\mathcal{V}$. Conseqüentemente, na **Eq.3.7**, o número n_c de contatos entre um particular monômero c e somente um sítio de solvente é igual a η , isto é, $n_c = \eta$, para qualquer c . Também, pelo fato da coordenação da rede, Z , ser constante para todos os sítios, o número de sítios da rede que a unidade k contata antes e depois do seu movimento deve ser constante, isto é: $n_k/\eta + m = n_k''/\eta + m''$. Usando estas igualdades na **Eq.3.4**, tem-se:

$$\Delta H = (m - m'')\eta\gamma_k + \eta(\gamma_p + \dots \gamma_t) - \eta(\gamma_u + \dots \gamma_w), \quad (3.8)$$

o que é exatamente igual à **Eq.3.4** para ΔE . Este resultado será explorado na seção seguinte, onde se adicionam as especificidades estéricas intermonômeros.

3.4 Especificidades Estéricas

A simplificação introduzida pela aproximação da energia de contato, através de um potencial puramente hidrofóbico, $E_{ij} = h_i + h_j$, apesar de garantir que a cadeia encontre sua estrutura nativa, não garante estabilidade configuracional ao sistema, como será mostrado mais adiante. Adota-se, então, neste trabalho, o potencial estéreo-hidrofóbico, $E_{ij}^* = h_i + h_j + c_{ij}$, que considera um conjunto de restrições $\{c_{ij}\}$ do tipo caroço duro (hard-core), juntamente com a energia

h	-2.1	-2.0	-1.9	-1.2	-1.1	-1.0	-1.0	-0.9	-0.8	+ 0.8
especificidades estéricas	 RAH	 RAH BGF	 RAH GIE	 A	 A H	 A C	 H C	 H C	 D C	 FIEDC
classes	0	1		2						3

Figura 3.6: Representação do conjunto das restrições estéricas. Na linha superior, mostram-se as energias hidrofóbicas (h) dos dez monômeros, que estão identificados pelas letras R, A, etc. As especificidades estéricas estão indicadas esquematicamente e representam quais os pares de monômeros que podem fazer contato topológico durante o processo de enovelamento. Indica-se, também, as quatro classes de monômeros cujos índices (0, 1, 2 e 3) correspondem ao número de contatos que estes estabelecem com o meio, quando a cadeia está na configuração nativa (adaptado de Tarragó et al., 2003, figura 1).

hidrofóbica, introduzido recentemente na literatura (da Silva et al., 2001; da Silva, 2002).

O **potencial estéreo-hidrofóbico** assume que as interações numa rede regular ocorrem:

- (i) entre pares de resíduos que são vizinhos mais próximos, através de um conjunto de ajustes e impedimentos estéricos $\{c_{i,j}\}$, que representam, de certa forma, a diversidade das especificidades estéricas encontrada no conjunto dos aminoácidos naturais;
- (ii) entre resíduos e moléculas do solvente, que é numericamente equivalente, neste contexto, à energia de contato intermonômeros $E_{i,j} = -(h_i + h_j)$.

O conjunto de restrições estéricas $\{c_{i,j}\}$ foi desenvolvido heurísticamente, e determina quais são os pares de monômeros que podem fazer contato topológico,

como indicado na **Figura 3.6**, seja qual for a configuração em que a cadeia se encontre (da Silva et al., 2001; da Silva, 2002; Tarragó et al., 2003).

3.5 Formalismo Entrópico para o Folding

As interações energéticas cadeia-solvente, solvente-solvente e cadeia-cadeia, em conjunto, são difíceis de serem tratadas analiticamente. Por outro lado, experimentalmente, as medidas energéticas correspondem às variações da energia livre que mesclam energia interna e entropia. Assim, é desejável um tratamento especial que possa utilizar diretamente dados experimentais.

3.5.1 Entropia e energia livre

Considere-se um sistema isolado, de volume V_0 , energia E_0 , e constituído por N_0 moléculas de solvente, e uma única macromolécula. A este sistema está associado um número total de estados, $\Gamma_0 = \Gamma_0(E_0, N_0, V_0)$, e, portanto, entropia $S_0 = k \ln \Gamma_0$. A questão é descrever a conformação da cadeia que pode, em princípio, assumir qualquer configuração que seus vínculos covalentes e estéricos permitem. Esta formulação corresponde ao *ensemble* microcanônico, e, assim, qualquer um dos estados, $\{\alpha\}$, do sistema como um todo, a princípio, possui a mesma probabilidade de ocorrência, a saber, $p_\alpha = 1/\Gamma_0$ —postulado fundamental da mecânica estatística. Por outro lado, sabe-se empiricamente que, no equilíbrio macroscópico, uma proteína no estado nativo apresenta-se numa determinada conformação, constituída por um conjunto relativamente pequeno de configurações intercambiantes. Surge, então, a pergunta: por que (em condições fisiológicas adequadas) as configurações da conformação nativa da proteína aparecem com mais frequência do que qualquer uma das outras configurações?

De fato, o postulado fundamental da mecânica estatística não deve implicar que cada configuração, c , da cadeia, em particular, tenha a mesma probabili-

dade de ocorrência, pois, de acordo com o que se verifica experimentalmente, a proteína, no estado nativo, apresenta-se em uma conformação bem definida. Para compreender melhor este aparente paradoxo, é essencial envolver todo o sistema, evitando considerar apenas a proteína. Assim, cada estado do sistema como um todo considera todas as moléculas do solvente e a proteína. Supõem-se, então, duas configurações distintas da cadeia proteica, a saber, c_1 e c_2 , tal que, quando a cadeia está em c_1 , o meio solvente apresenta mais possibilidades configuracionais do que quando a cadeia está em c_2 . Desta forma, ao longo do tempo, a configuração c_1 aparece com mais frequência porque, nessa situação, o meio complementar (o solvente) tem mais alternativas, lembrando-se que a probabilidade de cada estado α do sistema como um todo é $p_\alpha = 1/\Gamma_0$.

Então, na formulação do *ensemble* microcanônico, as chances de ocorrência de cada configuração c da macromolécula proteica é ponderada através do número de configurações que restam para o solvente —sistema complementar, isto é, $p'_c = \Gamma'_c/\Gamma_0$, onde Γ'_c é o número de configurações possíveis ao meio solvente para a mesma configuração c da cadeia. Termodinamicamente, pode-se descrever a situação subdividindo o sistema como um todo em duas partes: um subsistema envolvendo a proteína e um certo número, n , de moléculas do solvente, ambos encerrados num certo volume V , restando para o sistema complementar $N' = N_0 - n$ moléculas de solvente contidas num volume $V' = V_0 - V$. Também associa-se ao subsistema uma energia E , e, assim, resta ao sistema complementar uma energia $E' = E_0 - E$. Se, desta forma, o subsistema é considerado como que “congelado” na configuração c , resta ao sistema complementar Γ' estados, reduzindo a entropia do sistema como um todo para $S' = k_B \ln \Gamma'$, sendo, então, a variação de entropia $\Delta S = S_0 - S'_0$ dada por

$$\Delta S = -k \ln \frac{\Gamma'}{\Gamma_0}, \quad (3.9)$$

onde se usa a definição estatística da entropia, $S = k \ln \Gamma$. No entanto, substituindo-se, na relação acima, a expressão $p_c = \Gamma'/\Gamma_0$, que quantifica a probabilidade da

configuração c ocorrer, tem-se:

$$p_c = \exp(-\Delta S/k), \quad (3.10)$$

o que significa dizer que a configuração da cadeia que menos (ou mais) reduz a entropia do sistema complementar é aquela que ocorre com maior (ou menor) probabilidade. Note-se que as equações finais são exatamente as clássicas (Goodstein, 1975); porém, para interpretá-las adequadamente, é preciso lembrar que não se trata aqui de um sistema homogêneo.

Se o sistema não é homogêneo, como se pode escrever $\Delta S = \Delta S(E, n, V)$? Para se obter esta relação de forma genérica, é preciso interpretar o significado e identificar os componentes de cada um dos termos: E , n , e V . Quanto à energia E , estão envolvidas as interações intramoleculares (cadeia-cadeia), as interações entre as n moléculas do solvente e a macromolécula, entre as n moléculas do solvente entre si, e, finalmente, as interações entre as n moléculas do solvente e as outras N' moléculas do mesmo. Acontece que não se conhece diretamente estes valores, pois todas as interações ocorrem simultaneamente, e uma influencia a outra. Neste sentido, o que se tem são valores experimentais como a energia livre de solvatação e a energia livre de transferência, havendo assim a necessidade de se saber relacionar estas energias livres com todas as componentes de E e n . (O volume V é uma grandeza importante quando há variação de pressão, mas este caso não será particularmente discutido aqui.)

Um caso especial que ilustra a aplicação desse formalismo é aquele resultante quando se considera as interações entre a cadeia e o meio solvente mais importantes do que as outras. Assim, no sistema proteína mais solvente, estabelece-se que as interações, envolvendo variação de energia, ocorrem somente entre cada monômero e o meio solvente, e, então, pode-se utilizar dados experimentais para o cálculo de ΔS . A relação $\Delta F = \Delta E - T\Delta S$ (energia livre de Helmholtz, **Eq.2.4**) permite escrever $\Delta S = -\Delta F/T$, uma vez que a interação monômero-

monômero não é considerada, o que equivale a se estabelecer $\Delta E = 0$; desta forma, a energia livre de transferência (quantificada nas escalas de hidrofobicidades) pode ser usada diretamente para se estimar ΔS . Esta aproximação se justifica, pelo menos, para o processo de enovelamento em si, pois as interações energéticas intracadeia se tornam importantes somente na configuração nativa.

3.5.2 Probabilidade de transição entre duas configurações

A vantagem de se considerar somente as interações solvente-cadeia, desprezando-se todos os outros tipos de interações (cadeia-cadeia, solvente-solvente), nas simulações Monte Carlo, reside no fato de se conseguir usar diretamente os valores das medidas das energias livres de transferência na probabilidade de transição entre duas configurações.

Imagine-se a macromolécula, ou cadeia proteica, livre em uma solução com muitíssimas partículas de solvente. A probabilidade da cadeia encontrar-se numa configuração, c , é, então, associada ao peso estatístico:

$$p_c = \exp -[(S_0 - S'_c)/k], \quad (3.11)$$

onde S_0 é a entropia do sistema como um todo e S'_c é a entropia do meio quando a cadeia está “congelada” na configuração c . Como S_0 é uma constante, pode-se, então, escrever:

$$p_c = A \exp(S'_c/k), \quad (3.12)$$

onde A é uma constante independente de c , e é determinada pelas condições de normalização da probabilidade: $\sum_{\{c\}} p_c = 1$.

A probabilidade da cadeia transicionar da configuração c para c' pode ser obtida a partir da condição de equilíbrio detalhado que diz:

$$p_c T_{cc'} = p_{c'} T_{c'c}, \quad (3.13)$$

onde p_c é a probabilidade da cadeia estar em c ; $T_{cc'}$ é a probabilidade da cadeia transicionar da configuração c para a c' ; $p_{c'}$ é a probabilidade da cadeia estar em c' ; e $T_{c'c}$ é a probabilidade da cadeia transicionar da configuração c' para a c .

Se a entropia do meio aumenta com a transição da configuração c para c' , então se considera $T_{cc'} = 1$ e, conseqüentemente, $T_{c'c} = \frac{P_c}{P_{c'}}$, fornecendo:

$$T_{c'c} = \frac{A \exp(S'_c/k)}{A \exp(S'_{c'}/k)} \quad (3.14)$$

ou

$$T_{c'c} = \exp(S'_c - S'_{c'}/k) = \exp(-\Delta S'/k), \quad (3.15)$$

onde $\Delta S' = (S'_c - S'_{c'})$.

Mostra-se, a seguir, que a probabilidade de transição acima também pode ser escrita em termos da variação da energia livre. Escrevendo-se a entropia como $S = \frac{(E-F)}{T}$, tem-se que sua variação entre duas configurações é dada por $\Delta S'_{cc'} = [(E'_c - F'_c) - (E'_{c'} - F'_{c'})]/T$ ou $-\Delta S'_{cc'} = [(E'_c - E'_{c'}) - (F'_c - F'_{c'})]/T$. É preciso lembrar que se supõe neste trabalho que a variação da energia interna é nula, isto é, as interações monômero-monômero são ignoradas, com exceção da interação tipo caroço-duro ou do volume excluído (que é infinitamente repulsiva quando os monômeros tentam ocupar o mesmo sítio). Logo, se a variação da energia interna da macromolécula, entre duas configurações, é zero, a variação da energia do meio correspondente também é zero, isto é, $E'_c - E'_{c'} = 0$; conseqüentemente, $-\Delta S'_{cc'} = [-(F'_c - F'_{c'})]/T = (F'_{c'} - F'_c)/T$. Desta forma, tem-se a variação de entropia escrita em função da variação da energia livre do meio, que pode ser medida experimentalmente, e assim utilizada nas simulações computacionais.

Conseqüentemente, a probabilidade de transição da configuração c para a c' pode ser escrita em termos da energia livre de transferência, que é a própria energia hidrofóbica, pois as interações presentes mais importantes são entre os monômeros e a água; isto é:

Figura 3.7: Movimentos configuracionais permitidos. Os movimentos que deslocam um monômero de cada vez são o movimento de canto (corner move) e os movimentos de ponta (end moves); os movimentos que deslocam duas contas de cada vez são chamados de movimentos de manivela (crankshaft moves) (Socci & Onuchic, 1994; Figura 2).

$$T_{cc^0} = \exp(\Delta F'/kT), \quad (3.16)$$

onde $\Delta F' = F'_{c^0} - F'_c$.

3.6 Movimentos Configuracionais Permitidos

Na década de 70, foi demonstrado que a dinâmica do processo de enovelamento de cadeias em modelos em rede, onde a condição do volume excluído impõe que duas contas (monômeros) não podem ocupar o mesmo local na rede, depende dos possíveis movimentos configuracionais permitidos (Hilhort & Deutch, 1975). O ideal é usar um conjunto de movimentos que permita que a cadeia visite todas as configurações possíveis, o que garante a ergodicidade do sistema. Por esta razão, para uma cadeia de 27 contas em rede cúbica, como a adotada neste trabalho, o conjunto de determinados movimentos relativamente simples, que inclui o deslocamento de um ou dois monômeros por vez, é eficiente em reproduzir praticamente todas as configurações. Apesar de se ter conhecimento

que este conjunto de movimentos não conduz a configurações que apresentam nós, ele continua sendo considerado, na prática, como ergódico, pois as cadeias das proteínas reais não apresentam nós (Socci & Onuchic, 1994). A saber, os movimentos que deslocam apenas uma conta de cada vez são: (a) movimento de canto (corner move) e (b) movimento de ponta (end move); o movimento que desloca duas contas de cada vez é o (c) movimento de manivela, como ilustrados na **Figura 3.7**.

3.7 Simulações Monte Carlo

Neste trabalho, estuda-se o comportamento de moléculas proteicas em solução, via simulação Monte Carlo, em duas situações distintas: a de enovelamento (ou folding) e a de desnaturação por calor (ou unfolding). Em ambas, descreve-se explicitamente uma única cadeia em solução, e registra-se sistematicamente, em momentos independentes, as grandezas de interesse. Isto resulta em uma coleção de dados estatisticamente independentes, correspondentes ao que se obteria em uma única medida que levasse em consideração muitíssimas moléculas —como é feito na prática.

Nas simulações de folding, a uma dada temperatura T , parte-se da cadeia tipo-proteína “desnaturada” e registra-se o “tempo MC” ou “**tempo de folding**” necessário para a cadeia visitar, pela primeira vez, a configuração nativa; adota-se como uma unidade de tempo 8100 passos de simulação Monte Carlo. Nas simulações computacionais do processo de folding destacam-se as propriedades cinéticas e, no outro sentido da reação, no unfolding, evidenciam-se as propriedades termodinâmicas. No folding, a configuração inicial da cadeia corresponde a um estado desnaturado, representado por qualquer configuração aleatória, desde que os elementos da cadeia não apresentem nenhum contato topológico. Já, nas simulações de desnaturação, em cada temperatura, T , a cadeia encontra-se inicialmente enovelada na configuração nativa, onde estão

presentes os 28 contatos topológicos nativos. Em intervalos regulares, registra-se a energia de contato, E , da configuração (aqui representando a energia hidrofóbica), o número total de contatos topológicos e o número de contatos topológicos nativos. Também, conta-se o número de vezes em que a cadeia visita a configuração nativa.

É importante salientar que, neste estudo, a **temperatura** T é, de fato, a temperatura reduzida, dada por

$$T = \frac{k\theta}{k\theta_0},$$

onde θ é a temperatura propriamente dita e $k\theta_0$, expressa em unidades arbitrárias, é a energia de referência.

3.7.1 Atividade configuracional

A atividade configuracional da cadeia proteica é monitorada nas simulações através do número total de contatos e do número de contatos nativos entre os pares de monômeros e, também, pela frequência com que o sistema visita a configuração nativa. Tais parâmetros são definidos como:

- (a) **número relativo médio de contatos:** é a soma do número de contatos em cada configuração amostrada, dividido pelo número de configurações amostradas, normalizado pelo número máximo de contatos entre os vizinhos topológicos (ou seja, dividido por 28);
- (b) **número relativo médio de contatos nativos:** é a soma do número de contatos nativos em cada configuração amostrada, dividido pelo número de configurações amostradas, normalizado também pelo número máximo de contatos nativos;
- (c) **frequência relativa na configuração nativa:** é o número de vezes em que a cadeia se encontra na configuração nativa (CSA alvo), dividido pelo número total das conformações amostradas.

3.7.2 Cálculo da capacidade térmica

Experimentalmente, as proteínas são levadas à desnaturação através da variação das grandezas intensivas, como pH (pela introdução de uma determinada concentração de desnaturante na solução proteica) ou temperatura (no caso, desnaturação por calor). Se o processo é reversível, para iniciar a reação do folding é suficiente reconstituir as condições fisiológicas iniciais.

Neste modelo, cada vértice da rede cúbica está ocupado ou por uma molécula do solvente ou por um monômero da cadeia. Isto conduz à identidade matemática entre a energia hidrofóbica da cadeia, devido à interação monômero-solvente, e a energia de interação entre os pares de monômeros em contato topológico, o que pode ser considerado como energia interna da cadeia, como demonstrado anteriormente na subseção **3.3.2**. Assim, a energia hidrofóbica, E , de uma configuração c é a soma das energias dos pares ij em contato topológico, formalmente escrita por:

$$E = \varepsilon_0 + \sum_{\{i,j\}} \varepsilon_{i,j} \Delta(d_{ij}), \quad (3.17)$$

onde ε_0 é uma constante, $\varepsilon_{i,j}$ é a energia de contato entre os monômeros “ i ” e “ j ”; e $\Delta(d_{ij}) = 1$, se a distância d_{ij} entre os monômeros satisfaz $d_{ij} = 1$, e $\Delta(d_{ij}) = 0$, de outra forma. Se o valor da constante ε_0 é considerado nulo, a energia de uma cadeia totalmente aberta (desnaturada) será zero; no entanto, salienta-se que os resultados independem do valor desta constante, pois o que interessa no processo de enovelamento ou desnaturação não é o valor absoluto da energia, e, sim, sua variação.

Calcula-se, então, em cada temperatura, T , a variação de energia, que corresponde à capacidade térmica à pressão constante da “proteína em solução”, através da definição apresentada em **2.7.4**, que corresponde à:

$$C/k = \frac{1}{k} \frac{\partial \langle E(T) \rangle}{\partial T} = \frac{\langle E^2(T) \rangle - \langle E(T) \rangle^2}{k^2 T^2}, \quad (3.18)$$

onde $E(T)$ é a energia na temperatura T ; o símbolo $\langle \rangle$ denota média termodinâmica e k é a constante de Boltzmann (Kaya & Chan, 2000).

Observa-se que, no modelo adotado, a energia é dada em termos de unidades reduzidas, isto é, em unidades de kT ; logo C/k é adimensional. Como será visto mais adiante, a temperatura $T \simeq 1,0$ é aquela em que o processo de enovelamento ocorre mais rapidamente para algumas estruturas nativas, e, também, em que a frequência com que a cadeia visita a configuração nativa é maior; logo, grosso modo, pode-se associar a esta temperatura $T \simeq 1,0$ à temperatura fisiológica.

3.7.3 Distância ponta-a-ponta e raio de giração

Estudos de macromoléculas comumente descrevem a atividade conformacional através da distância ponta-a-ponta e do raio de giração. A **distância ponta-a-ponta** (d_{pp}) de uma proteína é definida como a distância entre os centros de massa dos dois aminoácidos das extremidades da cadeia, e o seu **raio de giração** (R_G) é definido como a raiz da média, ao quadrado, das distâncias dos átomos (ou dos centros de massa dos grupos de átomos) ao centro de massa da macromolécula. Para uma cadeia com n monômeros de massas iguais e l ligações entre eles ($l = n - 1$), o quadrado do raio de giração é dado por:

$$R_G^2 = \frac{1}{l+1} \sum_{i=1}^l R_{gi}^2, \quad (3.19)$$

onde R_{gi}^2 é o quadrado da distância do i -ésimo monômero ao centro de massa da macromolécula. Neste caso, o quadrado do raio de giração da macromolécula também pode ser obtido por:

$$R_G^2 = \frac{1}{n^2} \sum_{j>i}^l r_{ij}^2, \quad (3.20)$$

onde r_{ij} é a distância entre os monômeros i e j (Cantor & Schimmel, 1980).

No modelo adotado neste trabalho, quando a cadeia está no seu estado considerado nativo, ela estará em alguma das 103346 configurações maximamente compactas (CSA). Tem-se, então, que:

- (a) a distribuição dos valores possíveis para a dpp das CSA cobre cinco valores: $1,41 = \sqrt{2}$; $2,0$; $2,45$; $2,83$ e $3,46 = \sqrt{12}$;
- (b) a distribuição dos monômeros ao redor do centro de massa de qualquer uma das 103346 CSA é a mesma, pois há sempre oito monômeros nos vértices, doze nos meios das arestas e seis nos centros das faces; assim, qualquer CSA apresenta o mesmo raio de giração, $R_G = 1,41$.

3.8 Referências Bibliográficas

Alm, E. & Baker D. (1999) Matching Theory and Experiment in Protein Folding. *Curr. Opin. Struct. Biol.*, **9**: 189-196

Cantor, C.R. & Schimmel, P.R.(1980) **Biophysical Chemistry. Part III: The Behavior of Biological Macromolecules**. W.H. Freeman and Company, New York.

Chothia, C. (1974) Hydrophobic Bonding and Accessible Surface Area in Proteins. *Nature*, **248**: 338-339

Goodstein, D. L. (1975) **States of Matter**. Prentice-Hall, Inc., New Jersey.

Hilhorst, H. J. & Deutch, J. M. (1975) Analysis of Monte Carlo Results on the Kinetics of Lattice Polymer Chains with Excluded Volume. *J. Chem. Phys.*, **63**:5153-5161

Kaya, H. & Chan, H. S. (2000) Polymer Principles of Protein Calorimetric Two-State Cooperativity. *Proteins: Struct. Funct. Genet.*, **40**: 637-661

Lee, B. & Richards, F. M. (1971) The Interpretation of Protein Structures: Estimation of Static Accessibility. *J. Mol. Biol.*, **55**: 379-400

Li, H. et al. (1997) Nature of Driving Force for Protein Folding: A Result from Analyzing the Statistical Potential. *Phys.Rev.Lett.*, **79**: 765-768

Marques dos Reis, L. (1995) **Folding de Proteínas: Simulação Monte Carlo de um Modelo em Rede com Potencial Randômico**. Dissertação de Mestrado,FFCLRP, Departamento de Física e Matemática, Universidade de São Paulo.

Miyazawa, S. & Jernigan, R.L. (1985) Estimation of Effective Interresidue Contact Energies from Protein Crystal Structures. *Macromolecules*, **18**: 534-552

Nozaki, Y. & Tanford, C. (1971) The Solubility of Amino Acids and Two Glycine Peptides in Aqueous Ethanol and Dioxane Solutions. *J. Biol. Chem.*, **246**: 2211-2217

Plaxco, K. W. et al. (1998) Contact Order, Transition State Placement and the Folding Rates of Single Domain Proteins. *J. Mol. Biol.*, **277**: 985-994

da Silva, R. A. et al. (2001) Deterministic Folding: The Role of Entropic Forces and Steric Specificities. *J. Chem. Phys.*, **114**: 4235-4242

da Silva, R. A. (2002) **Enovelamento Determinístico de Macromoléculas**. Tese de Doutorado, FFCLRP, Departamento de Física e Matemática, Universidade de São Paulo.

Tarragó, M. E. P. et al. (2003) Steric Constraints as Folding Coadjuvant. *Phys. Rev. E*, no prelo.

Socci, N. D. & Onuchic, J. N. (1994) Folding Kinetics of Proteinlike Heteropolymers. *J. Chem. Phys.*, **101**: 1519-1528

Zhou, H. & Zhou, Y. (2002) Folding Rate Prediction Using Total Contact Distance. *Biophysical Journal*, **82**: 458-463

Capítulo 4

Resultados e Discussão

4.1 Distribuição das CSA Segundo Parâmetros Topológicos

Com o objetivo de utilizar critérios topológicos para a escolha da configuração maximamente compacta (CSA), que representará a configuração nativa ou configuração alvo neste trabalho, as 103346 configurações maximamente compactas (CSA), que constituem um cubo $3 \times 3 \times 3$, foram caracterizadas em função da ordem de contato, χ , da ordem de longo alcance, χ_{LR} , e do número de estruturas $STTS$, ϕ . As distribuições apresentadas a seguir revelam que a maioria destas estruturas apresenta: ordem de contato, χ , ao redor da média da distribuição, situando-se entre 0,30 e 0,35 (**Figura 4.1**); ordem de longo alcance, χ_{LR} , aproximadamente, entre 0,25 e 0,40 (**Figura 4.2**) e número de estruturas $STTS$, ϕ , pequeno ou nulo (**Figura 4.3**). A maioria dos estudos em rede, que utiliza cadeia tipo-proteína com 27 monômeros, apresenta resultados baseados no comportamento de uma única (ou muito poucas) configuração CSA, escolhida ao acaso entre as 103346 configurações maximamente compactas. Isto significa que esta CSA escolhida ao acaso provavelmente terá as características topológicas como acima descritas. Este trabalho propõe uma nova abordagem para o

Figura 4.1: Distribuição das 103346 configurações CSA em função da ordem de contato, χ , que fornece o número relativo médio de monômeros existentes entre os pares (de monômeros) em contato topológico.

problema do enovelamento de proteínas, ao levar em consideração os aspectos geométricos das configurações escolhidas como nativas.

Em analogia com as estruturas das proteínas globulares, os padrões geométricos do tipo *STTS*, das configurações CSA, podem ser associados às estruturas secundárias regulares do tipo hélices- α . Plaxco e colaboradores verificaram, para um conjunto de proteínas globulares e pequenas, que existe uma forte correlação negativa entre o número relativo de hélices- α e ordem de contato de uma estrutura nativa, com coeficiente de correlação igual a $-0,72$ (Plaxco et al., 1998). Na **Figura 4.4** constata-se uma forte correlação linear negativa entre o valor máximo da ordem de contato, χ_{max} , e o maior número de estruturas do tipo *STTS* que uma configuração CSA pode apresentar (coeficiente de correlação $-0,98$); verifica-se, por exemplo, que as configurações que têm número alto de estruturas do tipo *STTS* (sete ou oito) apresentam ordem de contato menor do que $0,32$.

Na **Figura 4.5** mostra-se a correlação linear entre a ordem de contato e a ordem de longo alcance, para as 103346 configurações CSA; cujo coeficiente de correlação é $0,82$.

Figura 4.2: Distribuição das 103346 configurações CSA, em função da ordem de longo alcance, χ_{LR} , ou seja, o número relativo de contatos topológicos que estão separados, ao longo da seqüência, por onze ou mais monômeros.

Figura 4.3: Distribuição das 103346 configurações CSA em função do número ϕ de estruturas secundárias do tipo *STTS*.

Figura 4.4: Valor máximo da ordem de contato que ocorre para cada número ϕ de estruturas *STTS*; a reta ajustada apresenta coeficiente de correlação $-0,98$.

Figura 4.5: Ordem de longo alcance em função da ordem de contato para as 103346 configurações CSA; o coeficiente de correlação linear é $0,82$.

Figura 4.6: Capacidade térmica, C/k , em unidades da constante de Boltzmann, em função da temperatura, T , para a cadeia 870; note-se que em $T = T' \simeq 0,9$ ocorre um ponto fora da tendência da curva.

4.2 Estudo Termodinâmico do Modelo Estéreo-hidrofóbico

4.2.1 Resultados para a cadeia 870

Inicialmente, mostram-se os resultados termodinâmicos obtidos nas simulações de desnaturação por calor da cadeia 870. Esta última é assim chamada, porque sua sequência foi obtida a partir da configuração CSA 870 (mostrada na **Figura 3.1**), que apresenta menor ordem de contato, $\chi = 0,2381$, maior número de *STTS* possível, $\phi = 8$, embora não apresente ordem de longo alcance, χ_{LR} , nula (que é o menor valor possível para este parâmetro), mas, sim, $\chi_{LR} = 0,2143$.

A capacidade térmica do sistema, C/k , em função da temperatura, T , — determinada por simulação Monte Carlo a partir do teorema da flutuação e dissipação, a saber, $C/k = \beta^2[\langle H^2(T) \rangle - \langle H(T) \rangle^2]$ — mostrada na **Figura 4.6**, permite identificar uma temperatura na qual as simulações equivalentes, mas independentes, apresentam resultados diferentes. Esta temperatura será identificada por T' , e, neste caso, ocorre em $T \simeq 0,9$; note-se o ponto fora da

tendência da curva C/k , ao redor de $T = 0,9$. O máximo da capacidade térmica, $C_{max}/k \simeq 81$, ocorre ao redor da temperatura $T_{max} = 1,5$, sendo a “largura da curva à meia altura” aproximadamente 0,4.

As **Figuras 4.7** e **4.8** mostram detalhes da atividade configuracional da cadeia 870, nas simulações de desnaturação por calor, em função da temperatura, T . A **Figura 4.7** mostra o número de contatos topológicos entre os monômeros, mais precisamente, em termos do número médio relativo do total de contatos e do número médio relativo dos contatos nativos; a **Figura 4.8** mostra a frequência relativa com que a cadeia ocupa a configuração nativa. Tendo-se em mente que, nas simulações de desnaturação, em cada temperatura T , parte-se da cadeia na configuração nativa, observa-se que:

- (i) em $T' \simeq 0,9$ ocorrem perturbações em todas as grandezas medidas: a capacidade térmica, os valores relativos médios do total de contatos e dos contatos nativos, e a frequência relativa na configuração nativa não se comportam de acordo com a tendência das respectivas curvas;
- (ii) para temperaturas entre 0,5 e T' praticamente todos os contatos formados são os nativos, uma vez que o número relativo médio do total de contatos é praticamente igual ao o número relativo médio dos contatos nativos;
- (iii) nas temperaturas entre 0,5 e T' a frequência relativa na configuração nativa cai de 0,2 a aproximadamente 0,1;
- (iv) em $T = T_{max}$ ocorre inflexão nas curvas que representam o número relativo médio do total de contatos e o número relativo médio de contatos nativos;
- (v) em $T = T_{max}$ a frequência relativa na configuração nativa cai, praticamente, a zero.

Apesar de, nesta seção, dar-se ênfase aos dados termodinâmicos, mostram-se nas figuras seguintes, por questão de completeza, também os aspectos dinâmicos do comportamento configuracional. Na **Figura 4.9**, mostra-se a evolução

Figura 4.7: Número relativo médio de contatos topológicos, para a cadeia 870, em função da temperatura, T ; os quadrados representam o número relativo médio do total de contatos, e os círculos, o número relativo médio dos contatos nativos. Note-se que, em $T = T' \simeq 0,9$, ocorrem flutuações.

Figura 4.8: Frequência relativa com que a cadeia 870 visita a configuração nativa, em função da temperatura, T ; note-se que, em $T = T' \simeq 0,9$, ocorre um ponto fora da tendência da curva.

Figura 4.9: Energia configuracional instantânea para a cadeia 870 em função do tempo de simulação, em $T = 0,83$.

da energia para a cadeia 870 em função do “tempo de simulação”, lembrando que cada unidade de tempo corresponde a 8100 passos Monte Carlo. Note-se que, nas simulações de desnaturação por calor, sempre se parte da configuração nativa em cada temperatura, T , cuja energia é $E^{(n)} = -59,8$. A **Figura 4.9** corresponde à temperatura $T = 0,83$, onde a cadeia visita com muita frequência a configuração nativa; a **Figura 4.10** corresponde à evolução da energia para a temperatura $T = 0,93$, onde se observa que todas as grandezas termodinâmicas apresentaram flutuações; e, finalmente, a **Figura 4.11** apresenta a evolução da energia configuracional para a temperatura onde ocorre o valor máximo da capacidade térmica, $T_{max} \simeq 1,5$, onde pode ser visto que a cadeia visita muitas configurações com energias diferentes, desde a nativa ($E^{(n)} = -59,8$) até aquelas bastante distendidas com energias maiores do que $E = -10$.

Figura 4.10: Evolução da energia para a cadeia 870 na temperatura $T = 0,93$, em que ocorrem descontinuidades nos gráficos de C/k e dos diversos parâmetros que quantificam a atividade configuracional em função da temperatura, T .

Figura 4.11: Energia da configuração em função do tempo de simulação, para a cadeia 870, na temperatura do pico do calor específico, $T_{max} = 1,5$.

Figura 4.12: Capacidade térmica, C/k , em função da temperatura, T , para a cadeia 99865; note-se que, em $T = T' \simeq 1,3$ ocorre um ponto fora da tendência da curva.

4.2.2 Resultados para a cadeia 99865

Como um exemplo contrastante ao anterior, em relação aos parâmetros topológicos tratados aqui, apresentam-se, a seguir, os resultados das simulações de desnaturação por calor, obtidos para a cadeia 99865, cuja estrutura nativa, mostrada na **Figura 3.2**, corresponde à configuração CSA 99865, que apresenta maior ordem de contato, $\chi = 0,4947$, maior ordem de longo alcance, $\chi_{LR} = 0,7143$, e menor número de *STTS*, $\phi = 0$.

A curva da capacidade térmica, C/k , em função da temperatura, T , mostrada na **Figura 4.12**, permite, também, identificar uma temperatura, $T' \simeq 1,3$, ao redor da qual ocorrem perturbações (como será visto mais adiante, esta região de perturbação é característica do modelo). Observe-se que o máximo da capacidade térmica ($C_{max}/k \simeq 55$) ocorre ao redor da temperatura $T_{max} = 1,8$, sendo a “largura à meia altura” desta curva igual a 0,8. Em relação à cadeia 870, cujos resultados foram apresentados anteriormente, note-se que T' e T_{max} são maiores e também a curva da capacidade térmica para a cadeia 99865 se apresenta menos localizada: o máximo da capacidade térmica, C_{max}/k , é menor e a curva que a

Figura 4.13: Número relativo médio do total de contatos (quadrados) e número relativo médio de contatos nativos (círculos) em função da temperatura, T , para a cadeia 99865; note-se que, em $T = T' \simeq 1, 3$, ocorrem flutuações.

Figura 4.14: Frequência relativa na configuração nativa, para a cadeia 99865, em função da temperatura, T ; note-se que, em $T = T' \simeq 1, 3$, ocorre brusca descontinuidade na curva.

Figura 4.15: Distância ponta-a-ponta (quadrados) e raio de giração (círculos) para as cadeias 870 (símbolos cheios) e 99865 (símbolos vazios), em função da temperatura, T .

representa é mais larga.

Os resultados que quantificam a atividade configuracional da cadeia 99865 estão mostrados nas **Figuras 4.13** e **4.14**, onde se nota que:

- (i) para temperaturas menores do que a temperatura $T' \simeq 1,3$, praticamente todos os contatos são nativos;
- (ii) no intervalo de temperatura que satisfaz $T < T'$, a frequência relativa na configuração nativa varia de aproximadamente 0,4 a 0,2;
- (iii) em torno da temperatura T' ocorrem descontinuidades nas curvas que representam o número relativo médio do total de contatos, o número relativo médio de contatos nativos e a frequência relativa na configuração nativa em função de T , sendo que esta última cai bruscamente a valores próximos a zero.

A **Figura 4.15** mostra, comparativamente, a distância ponta-a-ponta (quadrados) e o raio de giração (círculos) para as cadeias 870 (pontos cheios) e para a 99865 (pontos vazios); observe-se que, nas respectivas temperaturas T' , onde

ocorrem as perturbações anteriormente notadas, a distância ponta-a-ponta para ambas as cadeias apresenta visível descontinuidade.

4.2.3 Influência das condições iniciais

Diferentes seqüências de movimentos da cadeia

Nas simulações de desnaturação por calor de uma dada cadeia tipo-proteína, parte-se sempre da mesma configuração nativa em cada temperatura, T ; conseqüentemente, “diferentes condições iniciais” neste contexto implicam diferentes “histórias”, isto é, diferentes seqüências de movimentos da cadeia. Mostram-se, a seguir, os resultados das simulações de desnaturação por calor, partindo-se de doze condições iniciais diferentes, para a cadeia 11025, que apresenta como configuração nativa a CSA 11025, cuja ordem de contato é $\chi = 0,2460$, ordem de longo alcance, $\chi_{LR} = 0,250$, e número de estruturas $STTS$, $\phi = 0$. Observa-se na **Figura 4.16** que, no intervalo de temperatura entre $T' \simeq 0,8$ e $T_{max} \simeq 1,6$, ocorrem flutuações na capacidade térmica, C/k , dependentes das condições iniciais; o mesmo pode ser visto nas **Figuras 4.17** e **4.18**: ocorrem flutuações tanto nos números dos contatos topológicos total e nativo como da frequência relativa com que a cadeia visita a configuração nativa, no intervalo entre T' e T_{max} , mas cerca de T' a flutuação relativa é significativamente maior.

Influência da configuração inicial na curva da capacidade térmica

Neste tópico, ilustra-se, de forma comparativa, o efeito da configuração inicial no comportamento da capacidade térmica em função da temperatura. Usam-se duas configurações iniciais bem diferentes: a configuração CSA nativa e uma configuração aleatória, sem nenhum contato topológico. A **Figura 4.19** mostra a capacidade térmica, C/k , em função da temperatura, T , para a cadeia 99865; os pontos cheios correspondem à cadeia inicialmente na configuração CSA nativa, e os pontos vazios, à cadeia inicialmente na configuração aleatória. Note-se que,

Figura 4.16: Capacidade térmica, C/k , em função da temperatura, T , para a cadeia 11025, usando-se doze condições iniciais diferentes; parte-se sempre da CSA nativa.

Figura 4.17: Atividade configuracional, em função da temperatura, T , para a cadeia 11025, usando-se doze condições iniciais diferentes; parte-se sempre da CSA nativa. Os pontos cheios indicam o número relativo médio do total de contatos, e os pontos vazios, o número relativo médio dos contatos nativos.

Figura 4.18: Frequência relativa de visitas à configuração nativa, em função da temperatura, T , para a cadeia 11025, usando-se doze condições iniciais diferentes; parte-se sempre da CSA nativa.

Figura 4.19: Capacidade térmica, C/k , em função da temperatura, T , para a cadeia 99865, partindo-se da configuração nativa (pontos cheios), e partindo-se de uma configuração aleatória sem nenhum contato topológico (pontos vazios).

Figura 4.20: CSA de Goldstein, que é a CSA “mais otimizada” energeticamente, segundo Govindarajan e Goldstein (1995), e CSA de Li, que é a “mais desenhável”, segundo Li e colaboradores (1996); ambas apresentam número de estrutura *STTS*, $\phi = 0$.

quando se inicia a simulação com a cadeia em uma configuração aleatória, ocorre uma grande flutuação nos valores de C/k em temperaturas $T < T'$ (círculos vazios), $T' \simeq 1, 3$, indicando que a cadeia fica aprisionada, ao longo da simulação, em mínimos locais de energia.

4.2.4 Resultados para as cadeias de Goldstein e de Li

Na seção 2.9 foram apresentadas duas configurações CSA que outros autores encontraram como a “mais otimizada” (Govindarajan & Goldstein, 1995) —que será chamada “CSA de Goldstein”— e a “mais desenhável” (Li et al., 1996) —que será chamada “CSA de Li”. Os parâmetros topológicos destas configurações são: ordem de contato, $\chi = 0,2989$, e ordem de longo alcance, $\chi_{LR} = 0,2142$,

para a CSA de Goldstein; e ordem de contato, $\chi = 0,4021$, e ordem de longo alcance, $\chi_{LR} = 0,4643$, para a CSA de Li; ambas as configurações CSA apresentam número de estruturas *STTS*, $\phi = 0$ e são mostradas na **Figura 4.20**. Obtiveram-se, então, as seqüências das duas cadeias tipo-proteínas, correspondentes a estas configurações CSA, de acordo com a sintaxe utilizada neste estudo, e as energias dos estados nativos: $E^{(n)} = -58,8$, para a CSA de Goldstein, e $E^{(n)} = -67,6$, para a CSA de Li. Observando-se que, para estas duas cadeias, **as simulações MC de desnaturação partiram sempre de configurações aleatórias**, mostram-se, a seguir, os resultados da capacidade térmica, C/k (**Figura 4.21**), e da atividade configuracional (**Figura 4.22**), em função da temperatura, T . Também pode ser visto na **Figura 4.23**, o tempo de folding, que significa o “tempo” de simulação necessário para a cadeia encontrar, pela primeira vez, a configuração CSA nativa; cada unidade de tempo de simulação é igual a 8100 passos MC. O intervalo de temperatura utilizado nestas simulações vai de $T = 0,4$ a $T = 2,5$. Observe-se que a curva C/k para a cadeia de Goldstein, apresenta menos pontos que não seguem a tendência da curva do que a correspondente para a cadeia de Li; acontece que a cadeia de Goldstein encontra pela primeira vez a CSA nativa em $T = 0,7$, enquanto que a cadeia de Li encontra a CSA nativa, pela primeira vez, em $T = 1,0$; note-se que é somente a partir destas temperaturas, em ambos os casos, que as curvas C/k apresentam-se “mais bem comportadas”. Desta forma, é a partir destas temperaturas — $T = 0,7$, para a cadeia de Goldstein, e $T = 1,0$, para a cadeia de Li— que os resultados da atividade configuracional devem ser considerados (**Figura 4.22**). Observe-se na **Figura 4.22** que a cadeia de Goldstein apresenta estabilidade configuracional considerável, num certo intervalo de temperatura, enquanto que a cadeia de Li não a apresenta, pois a frequência com que esta última visita a configuração nativa é muito próxima de zero (triângulos vazios).

Figura 4.21: Capacidade térmica, C/k , em função da temperatura, T , partindo-se sempre de uma configuração aleatória, para a cadeia de Goldstein (quadrados cheios) e para a cadeia de Li (círculos vazios).

Figura 4.22: Atividade configuracional, em função da temperatura, T , partindo-se sempre de uma configuração aleatória, para a cadeia de Goldstein e para a cadeia de Li. Os quadrados representam o valor relativo médio do total de contatos; os círculos, o valor relativo médio dos contatos nativos; e, os triângulos, a frequência com que a CSA nativa é visitada.

Figura 4.23: Tempo de folding, em função da temperatura, T , para a cadeia de Goldstein (quadrados cheios) e para a cadeia de Li (círculos vazios); tempo próximo de zero indica que a cadeia enovela rapidamente na configuração CSA nativa, e tempo igual a 100 mil, significa que a cadeia não encontrou a configuração nativa, no intervalo de tempo considerado.

4.2.5 Correlação entre quantidades termodinâmicas e parâmetros topológicos

A desnaturação por calor de dezenas de cadeias tipo-proteínas foi considerada neste trabalho. Estas correspondem a configurações nativas caracterizadas pelos mais diversos valores dos parâmetros topológicos, como ordem de contato, ordem de longo alcance e número de estruturas *STTS*. Investigaram-se as possíveis correlações entre os atributos topológicos das configurações nativas e os parâmetros termodinâmicos das cadeias correspondentes. Encontrou-se significativa correlação linear negativa entre a energia da configuração nativa e os parâmetros topológicos como ordem de contato (coeficiente de correlação igual a $-0,78$) e ordem de longo alcance (coeficiente de correlação igual a $-0,74$); veja-se a **Figura 4.24 (a)** e **(b)**, respectivamente. Pode-se dizer, então, que as configurações nativas, com valores baixos para ordem de contato e para ordem de longo alcance, tendem a apresentar energias maiores, em torno de $E^{(n)} = -60$, comparadas às com ordem de contato e ordem de longo alcance mais altas, que

Figura 4.24: Energia da configuração nativa em função da ordem de contato, χ , e da ordem de longo alcance, χ_{LR} , para as mais de 40 configurações CSA estudadas.

apresentam energias em torno de $E^{(n)} = -67$, cerca de 10% menores.

Os gráficos da **Figura 4.25 (a) e (b)** mostram a dispersão dos pontos que representam o valor máximo da capacidade térmica (C_{max}/k) e a “largura à meia altura” da curva capacidade térmica, C/k (em função da temperatura, T), em função da ordem de contato, χ . Observe-se que existe uma tendência da curva C/k , em função da temperatura, T , apresentar-se “mais baixa” e “mais larga” na medida em que a ordem de contato, χ , da configuração nativa, aumenta. No entanto, a grande dispersão dos dados — em (a) o coeficiente de correlação linear é $-0,34$, e, em (b), é $0,46$ — indica que mais de um parâmetro topológico pode influenciar de forma significativa nas características termodinâmicas da cadeia.

A título de curiosidade, mostra-se na **Figura 4.26** a correlação entre a ordem de longo alcance e a ordem de contato para o conjunto das configurações CSA estudadas; o coeficiente de correlação linear é $0,96$. Este último indica que as configurações CSA escolhidas apresentam estes dois parâmetros topológicos mais fortemente correlacionados do que quando se considera o conjunto das 103346 configurações CSA, cujo coeficiente de correlação é $0,82$.

Figura 4.25: Valor máximo da capacidade térmica, (a), e “largura à meia altura”, (b), em função da ordem de contato, χ , da configuração nativa, para mais de 40 configurações CSA estudadas. Em (a) o coeficiente de correlação linear é $-0,34$ e em (b), é $0,46$; a baixa correlação observada em ambos os casos pode indicar que vários fatores topológicos devem ser considerados em conjunto.

Figura 4.26: Ordem de longo alcance em função da ordem de contato para o conjunto das configurações CSA estudadas.

Figura 4.27: Temperatura T , onde o número relativo médio do total de contatos é 0,5, em função da temperatura T_{max} , onde ocorre o pico da capacidade térmica, para as configurações CSA estudadas; o coeficiente de correlação linear é 0,71.

Figura 4.28: Temperatura, T_{max} , onde ocorre o pico da capacidade térmica, em função da ordem de contato, χ , da configuração nativa, para mais de 40 configurações CSA estudadas.

A **Figura 4.27** mostra os valores da temperatura, T , em que o número relativo médio do total de contatos é 0,5, em função da temperatura onde ocorre o valor máximo da capacidade térmica, T_{max} ; e a **Figura 4.28** apresenta a temperatura do máximo da capacidade térmica, T_{max} , em função da ordem de contato da configuração nativa, ambas para as mais de 40 configurações CSA estudadas.

4.2.6 Discussão dos aspectos termodinâmicos dos resultados

As simulações de desnaturação por calor foram realizadas de forma a cobrir as distribuições dos parâmetros topológicos, principalmente a ordem de contato, χ , e o número de estruturas $STTS$, ϕ . Os resultados das simulações, para as cadeias que apresentam configurações nativas com baixo valor de χ e elevado valor de ϕ , são claramente bastante diferenciados daqueles obtidos para as cadeias correspondentes a configurações com χ elevado e ϕ pequeno, evidenciando que o potencial estérico-hidrofóbico adotado neste trabalho é capaz de relacionar o comportamento termodinâmico da cadeia tipo-proteína com os atributos topológicos da configuração nativa correspondente.

De uma maneira geral, observa-se que existe uma temperatura, T' , em torno da qual ocorrem perturbações em todas as grandezas termodinâmicas monitoradas nas simulações, significando que estas são dependentes das condições iniciais. Aqui, entende-se por condição inicial a “história” dos movimentos da cadeia, que é determinada pela sequência dos números aleatórios, uma vez que, em cada temperatura, T , a cadeia parte sempre da configuração nativa. Para sistemas em equilíbrio, as grandezas termodinâmicas não dependem de tal sequência de números aleatórios; conseqüentemente, esta dependência em relação à história dos movimentos da cadeia indica que o sistema não está em equilíbrio, nesta dada temperatura. Para temperaturas menores do que T' , a energia do

sistema não varia de forma significativa, e a cadeia encontra-se em configurações que estão próximas, em energia, da configuração nativa. Nestas temperaturas baixas, é como se o sistema estivesse fortemente atraído pela configuração nativa —quase que congelado—, pois praticamente todos os contatos observados são os nativos. Apesar da energia térmica, kT , ser considerada pequena neste intervalo, ela é suficiente para atuar de forma diferenciada nas cadeias correspondentes a configurações nativas com diferentes parâmetros topológicos. De fato, quando se analisa o comportamento das cadeias em temperaturas baixas, aquelas que correspondem a configurações nativas com **maior valor de ordem de contato**, χ , **e menor valor do número de estruturas STTS**, ϕ , permanecem na configuração nativa duas vezes mais, em média, do que as cadeias que apresentam configurações nativas com **menor χ e maior ϕ** . Este resultado indica que as cadeias correspondentes às estruturas nativas com **maior ordem de contato e menor número de estruturas STTS** tendem a apresentar maior rigidez em relação à exploração do espaço conformacional em torno da estrutura nativa: ou a cadeia está na configuração maximamente compacta (nativa) ou em outra configuração muito próxima à nativa, de tal forma que retorna a esta última com bastante frequência. Por outro lado, as cadeias que correspondem às configurações nativas com **ordem de contato χ pequena e número ϕ de estruturas STTS grande** são mais maleáveis e exploram um espaço conformacional bem mais vasto em torno da configuração nativa, mesmo em baixas temperaturas. Este comportamento sugere que tais cadeias orbitam em torno da configuração nativa e de muitas outras configurações com energias próximas àquela da conformação nativa.

Mais pode ser dito a respeito da atividade configuracional analisando a forma com que a frequência relativa na configuração nativa vai a zero. Para as cadeias que apresentam configurações nativas com **maior ordem de contato**, χ , **e menor número de estruturas STTS**, ϕ , na temperatura T' , ocorrem

grandes descontinuidades nas curvas que representam os valores relativos médios do total de contatos e dos contatos nativos, e da frequência relativa na configuração nativa, sendo que esta última tende a zero em T' . Tem-se, então, para estas cadeias, uma transição de fase marcada por T' , que separa dois regimes de atividade configuracional com características diferenciadas:

1. para $T < T'$, as cadeias se encontram congeladas, e kT é pequeno demais para explorar configurações diversificadas;
2. para $T > T'$, kT é suficientemente grande para tirar a cadeia da configuração nativa e conduzi-lá por muitas outras configurações agora acessíveis.

As cadeias que apresentam configurações nativas com **menor valor de ordem de contato**, χ , e **maior número de estruturas** $STTS$, ϕ , apresentam a curva da frequência relativa na configuração nativa como uma função contínua da temperatura, desde $T = 0,5$ até a temperatura do pico da capacidade térmica (T_{max}), sendo que esta última é a temperatura em que, de fato, a frequência na nativa cai a zero; com exceção da temperatura T' , onde ocorre descontinuidade dependente da semente (fato que será explorado mais adiante). Tem-se, então, uma transição marcada por T_{max} que separa duas regiões de atividade configuracionais distintas.

A temperatura em que ocorre o máximo da capacidade térmica, T_{max} , para todas as configurações estudadas, marca a inflexão nas curvas do número relativo médio do total de contatos, do número relativo médio de contatos nativos, da distância ponta-a-ponta e do raio de giração, indicando mudanças na atividade configuracional, separando dois regimes conformacionais.

4.3 Cinética do modelo estéreo-hidrofóbico

Nas simulações do processo de folding parte-se sempre, em cada temperatura, de uma configuração aleatória —diversas condições iniciais significam configu-

rações aleatórias diferentes, sendo em todas elas o número de contatos topológicos nulo— onde:

1. mediu-se o **tempo de folding**, definido como o tempo necessário para a cadeia visitar, pela primeira vez, a configuração nativa, lembrando-se que uma unidade de tempo de simulação é, neste trabalho, igual a 8100 passos Monte Carlo;
2. contou-se a **freqüência relativa na configuração nativa**, dada pelo número relativo de vezes em que a cadeia visita esta configuração;
3. verificou-se quantas **configurações CSA alternativas** a cadeia visitou durante a simulação;
4. levantou-se o **histograma da distribuição da energia** em cada temperatura;
5. registrou-se a evolução da **energia em função do tempo de simulação**.

4.3.1 Resultados para a cadeia 870

A cadeia 870, cuja seqüência de monômeros corresponde à configuração CSA 870, escolhida como alvo (nativa), apresenta menor ordem de contato, $\chi = 0,2381$, maior número de estruturas STTS, $\phi = 8$, e ordem de longo alcance, $\chi_{LR} = 0,2143$. A **Figura 4.29** mostra o tempo de folding em função da temperatura para uma dada condição inicial, onde o tempo próximo de zero significa que a cadeia encontra rapidamente a configuração nativa, e tempo igual a 100 mil significa que a cadeia não encontra a configuração nativa nesta “janela de tempo” considerada (equivalente à $8,1 \times 10^8$ passos MC); observa-se que existe um intervalo de temperatura entre aproximadamente 0,8 e 2,2 em que o tempo de folding é relativamente muito pequeno.

Figura 4.29: Tempo de folding, em função da temperatura, T , para a cadeia 870. Tempo próximo de zero significa que a cadeia encontra rapidamente a configuração nativa, e tempo igual a 100 mil significa que a cadeia não enovela na configuração nativa neste intervalo considerado.

A **Figura 4.30** mostra como varia a frequência relativa com que a cadeia visita a configuração nativa em função da temperatura, T ; observe-se que em, $T = 0,63$, a cadeia não enovela para esta dada condição inicial, indicando que a mesma fica presa em armadilhas energéticas.

As **Figuras 4.31** e **4.32** mostram os histogramas das distribuições da energia nas temperaturas $T = 1,0$ e $T = 2,4$, respectivamente, para a cadeia 870, cuja energia na configuração nativa é $E^{(n)} = -59,8$. Atente-se que em $T = 1,0$, a cadeia visita outras configurações com energias preferencialmente muito próximas à da nativa, cujo valor satisfaz $E < -59$, enquanto que, em $T = 2,4$, a cadeia visita diversas configurações bastante distendidas, com energias muito maiores.

Da **Figura 4.33** à **4.37**, mostra-se a energia da cadeia 870 em função do tempo de simulação para cinco temperaturas diferentes, ilustrando-se como se

Figura 4.30: Frequência relativa da cadeia 870 na sua configuração nativa (CSA 870), em função da temperatura, T . Em $T = 0,63$, onde a frequência é zero, a cadeia não enovela no intervalo de tempo considerado.

Figura 4.31: Distribuição da energia, em $T = 1,0$, para a cadeia 870, cuja energia na configuração nativa é $E^{(n)} = -59,8$.

Figura 4.32: Distribuição da energia, em $T = 2, 4$, para a cadeia 870.

processa a busca da configuração nativa. Para temperaturas baixas, como na **Figura 4.33**, correspondente à temperatura $T = 0,43$, a cadeia fica presa em diferentes conformações e lentamente evolui em direção à configuração nativa ($E^{(n)} = -59,8$). Na **Figura 4.34**, em $T = 0,63$, a cadeia fica aprisionada, neste caso, em configurações com energias entre aproximadamente $E = -50$ e -45 , não conseguindo encontrar a configuração nativa no intervalo de tempo considerado. Já, na **Figura 4.35**, correspondente à $T = 1,0$, a cadeia oscila, principalmente, entre a energia da configuração nativa e energias aproximadamente iguais à $E = -50$; e, na **Figura 4.36**, a cadeia encontra-se na temperatura do pico da capacidade térmica, onde a flutuação de energia é máxima, indo desde a energia da configuração nativa, $E^{(n)} = -59,8$, até aproximadamente $E = -10$. Na **Figura 4.37**, a temperatura $T = 3,0$ revela-se alta o suficiente para garantir que a cadeia, na maioria das vezes, permaneça com energia oscilando entre aproximadamente $E = -25$ e $E = 0$; eventualmente, níveis de energias muito mais baixos são atingidos, mas nesta temperatura a estabilidade na configuração nativa é muito reduzida.

Figura 4.33: Energia, em função do tempo MC, no processo de envelhecimento, para a cadeia 870, na temperatura $T = 0,43$.

Figura 4.34: Energia, em função do tempo MC, para a cadeia 870, na temperatura, $T = 0,63$, em que não ocorre envelhecimento, nesta dada condição inicial.

Figura 4.35: Energia, em função do tempo MC, para a cadeia 870, em $T = 1,0$, onde a cadeia visita, com bastante frequência, a configuração nativa ($E^{(n)} = 59,8$).

Figura 4.36: Energia, em função do tempo MC, para a cadeia 870, na temperatura onde ocorre o pico da capacidade térmica, $T_{max} = 1,5$.

Figura 4.37: Energia, em função do tempo MC, para a cadeia 870, em uma temperatura bastante alta, $T = 3,0$, em que a configuração nativa praticamente não é visitada.

4.3.2 Resultados para a cadeia 99865

A cadeia 99865, cuja sequência de monômeros corresponde à configuração nativa CSA 99865, que apresenta maior ordem de contato, $\chi = 0,4947$, maior ordem de longo alcance, $\chi_{LR} = 0,7143$, e menor número de estruturas *STTS*, $\phi = 0$. As **Figuras 4.38** e **4.39** mostram, respectivamente, o tempo de folding e a frequência relativa na configuração nativa, em função da temperatura, T , para sete condições iniciais diferentes. Lembre-se que tempo próximo de zero indica que a cadeia encontra rapidamente a configuração nativa, e tempo igual a 100 mil (equivalente à $8,1 \times 10^8$ passos MC) significa que a cadeia não se enovela na configuração nativa, neste intervalo de tempo considerado. Além do mais, para esta cadeia, o tempo de folding, para uma mesma temperatura, pode mudar várias ordens de grandeza, sendo severamente dependente da condição inicial; e, no pequeno intervalo de temperatura, entre aproximadamente 1,4 e 2,0, a cadeia encontra mais rapidamente seu estado nativo. Na **Figura 4.39**, observa-se que a

Figura 4.38: Tempo de folding, em função da temperatura, T , considerando sete condições iniciais diferentes, para a cadeia 99865. Tempo próximo de zero significa que a cadeia encontra rapidamente a configuração nativa, e tempo igual a 100 mil significa que a cadeia não se enovela na configuração nativa, neste intervalo considerado. (Comparar com as Figuras 4.29 e 4.56.)

freqüência relativa com que a cadeia visita a configuração nativa é dependente da condição inicial, marcadamente no intervalo de temperatura aproximadamente entre $T = 1,0$ e $1,5$.

A cadeia 99865 visita três configurações maximamente compactas com diferentes valores para o número de contatos nativos, como é mostrado na **Figura 4.40**. Tem-se uma configuração CSA com 16 contatos nativos e outra com 22, além da CSA nativa (com 28 contatos nativos).

Os histogramas da distribuição da energia para a cadeia 99865, nas temperaturas $T = 1,2$ e $T = 1,8$, são mostrados nas **Figuras 4.41** e **4.42**, respectivamente. A temperatura $T = 1,2$ corresponde à menor temperatura em que a cadeia é capaz de encontrar a configuração nativa, para esta dada configuração inicial, e, $T = 1,8$, corresponde à temperatura em que ocorre o máximo da capacidade térmica. A energia da configuração nativa é $E^{(n)} = -63,7$; a cadeia

Figura 4.39: Frequência relativa com que a cadeia visita a configuração nativa, em função da temperatura, T , para sete condições iniciais diferentes, correspondente à cadeia 99865.

Figura 4.40: Frequência absoluta em três configurações maximamente compactas (CSA) visitadas pela cadeia 99865, em $T = 1, 2$; a que corresponde à configuração nativa apresenta número de contatos nativos igual a 28.

Figura 4.41: Distribuição da energia para a cadeia 99865 na menor temperatura em que a cadeia encontra a configuração nativa, $T = 1,2$, para uma dada condição inicial.

Figura 4.42: Distribuição da energia para a cadeia 99865 na temperatura do pico da capacidade térmica, $T = 1,8$.

Figura 4.43: Energia da configuração, em função do tempo MC, para a cadeia 99865, em $T = 1, 2$; a energia da configuração nativa é $E^{(n)} = -63, 7$.

Figura 4.44: Energia, em função do tempo MC, para a cadeia 99865 na temperatura do pico da capacidade térmica, $T_{\max} = 1, 8$.

não visita nenhuma outra configuração (além da nativa) com energia igual ou menor do que esta.

O comportamento da energia da cadeia 99865, em função do tempo de simulação, pode ser visto nas **Figuras 4.43 e 4.44**, para $T = 1, 12$, e para $T = 1, 8$, respectivamente.

4.3.3 Comparando o comportamento cinético de duas cadeias com mesma ordem de contato

Apresentam-se nesta subseção alguns resultados das simulações de enovelamento para duas cadeias cujas seqüências correspondem a configurações nativas com mesma ordem de contato, $\chi = 0,3307$, mesma ordem de longo alcance, $\chi_{LR} = 0,3571$, e número ϕ de estruturas *STTS* diferentes. A cadeia 57164 corresponde à seqüência da CSA 57164, que possui $\phi = 6$, valor máximo de estruturas *STTS* que ocorre para esta ordem de contato, e a cadeia 18380, corresponde à seqüência da CSA 18380, que apresenta $\phi = 0$. Na **Figura 4.45** mostra-se o tempo de folding, em função da temperatura, T , para estas duas cadeias, onde se observa que a cadeia correspondente à CSA com $\phi = 6$ (quadrados cheios) apresenta um intervalo de temperatura entre aproximadamente 0,9 e 2,1, em que o tempo para enovelar é bastante pequeno; isto não ocorre para a cadeia correspondente à CSA cujo $\phi = 0$ (círculos vazios).

Na **Figura 4.46** mostra-se a frequência relativa na configuração nativa, em função da temperatura, T , para essas duas cadeias, onde pode ser visto que a cadeia 57164, correspondente à CSA com $\phi = 6$, apresenta estabilidade considerável na configuração nativa (quadrados cheios), enquanto a cadeia 18380, cuja CSA nativa tem $\phi = 0$, praticamente não permanece na configuração nativa em nenhuma temperatura (círculos vazios).

A atividade configuracional dessas duas cadeias com mesmas χ e χ_{LR} é bem diferenciada; enquanto a cadeia 57164 (cuja CSA nativa tem $\phi = 6$) enovela-

Figura 4.45: Tempo de folding para as cadeias 57164 e 18380, cujas configurações nativas apresentam mesma ordem de contato ($\chi = 0, 3307$), mesma ordem de longo alcance ($\chi_{LR} = 0, 3571$) e número ϕ de estruturas *STTS* diferentes.

Figura 4.46: Frequência relativa com que as cadeias visitam as configurações nativas, em função da temperatura, T , para as cadeias 57164 e 18380, que apresentam configurações nativas com mesma ordem de contato ($\chi = 0, 3307$), mesma ordem de longo alcance ($\chi_{LR} = 0, 3571$), e número ϕ de estruturas *STTS* diferentes.

Figura 4.47: Número de vezes que as cadeias 57164 e 18380 visitam configurações maximamente compactas (CSA), em $T = 1,03$. As CSA estão identificadas pelo número de contatos nativos. Note-se que para o caso de $\phi = 6$ somente a CSA nativa é visitada.

se apenas na sua CSA nativa, a cadeia 18380 (cuja CSA nativa tem $\phi = 0$), além de eventualmente encontrar a sua CSA nativa, visita mais sete outras configurações maximamente compactas, todas, porém, com baixa estabilidade. A **Figura 4.47** mostra o número absoluto de vezes que cada configuração maximamente compacta (CSA) é visitada, sendo estas identificadas pelo número de contatos topológicos em comum com a CSA nativa, chamados números de contatos nativos. Obviamente, quando este é 28, trata-se da CSA nativa.

4.3.4 Estudo do tempo de enovelamento para 600 cadeias tipo-proteína

Estudou-se o comportamento do tempo de enovelamento da cadeia 870 (cuja CSA nativa apresenta ordem de contato, $\chi = 0,2381$, e número de estruturas *STTS*, $\phi = 8$) para 600 configurações aleatórias diferentes, sendo todas com-

Figura 4.48: Decaimento do número de cadeias desnaturadas, cujo estado nativo corresponde à CSA 870, em função do tempo, no processo de enovelamento de 600 “moléculas”. A curva cheia corresponde a um ajuste do tipo decaimento exponencial de segunda ordem e a pontilhada, a um de primeira ordem.

pletamente desnaturadas (sem nenhum contato topológico), o que corresponde a um experimento de folding para uma coleção de 600 cadeias tipo-proteínas não interagentes. Lembre-se que tempo de folding é definido aqui como o tempo de simulação necessário para a cadeia encontrar, pela primeira vez, a configuração nativa. Isto foi feito para algumas temperaturas, e em todas elas se observa que a cinética do enovelamento, embora possa ser considerada aproximadamente como um processo do tipo dois-estados —onde os dados se ajustam a uma curva exponencial decrescente de primeira ordem, de acordo como a curva tracejada na **Figura 4.48**— é de fato um processo que segue dois caminhos preferenciais para o folding, cujos dados se ajustam perfeitamente à uma curva exponencial decrescente de segunda ordem, de acordo com a curva cheia da **Figura 4.48**. Este comportamento já tinha sido constatado anteriormente para este potencial estéreo-hidrofóbico (da Silva et al., 2001; da Silva, 2002). Neste caso particular, em que foram utilizadas $N_0 = 600$ cadeias, em condições iniciais diferentes,

foi possível estimar as duas constantes de tempo de folding t_f e encontrou-se $t_f^1 = 455 \pm 9$ e $t_f^2 = 2684 \pm 146$, significando que, a cada intervalo de tempo característico t_f^i , uma fração aproximadamente igual a 37% das proteínas desnaturadas, seguindo a rota i de enovelamento, encontrou a configuração nativa.

4.3.5 Estudo do tempo de enovelamento de diversas cadeias tipo-proteínas

O estudo do tempo de enovelamento de várias cadeias tipo-proteínas foi realizado numa mesma temperatura, $T = 1,02$, que se mostrou adequada por ser suficientemente alta para garantir que a maioria das cadeias escapassem das armadilhas energéticas e topológicas, na busca do estado nativo, e, ao mesmo tempo, por ser suficientemente baixa para conferir à cadeia a máxima estabilidade no estado nativo. Nesta subseção, chama-se **tempo de enovelamento**, ou tempo de folding, a média dos três menores tempos MC de simulação em que a cadeia encontra, pela primeira vez, a configuração nativa; para cada seqüência tipo-proteína partiram-se de quinze condições iniciais diferentes. Simulou-se o folding de mais de 40 cadeias cujas configurações nativas correspondentes apresentam os mais diversos valores de ordem de contato, ordem de longo alcance e número de estruturas *STTS*; no entanto, mostram-se os resultados apenas para as cadeias que se enovelaram no intervalo de tempo menor do que 100 mil (o que é equivalente a $8,1 \times 10^8$ passos MC, conforme convencionou-se anteriormente). Nas **Figuras 4.49 e 4.50** mostram-se o tempo de folding (em escala logarítmica), em função da ordem de contato, χ , e da ordem de longo alcance, χ_{LR} , respectivamente, onde os correspondentes coeficientes de correlação linear são 0,73 e 0,69. Tal correlação, em ambos os casos, permite dizer que existe uma tendência do tempo de enovelamento em aumentar exponencialmente com os parâmetros ordem de contato e ordem de longo alcance. As cadeias correspondentes a con-

Figura 4.49: Tempo MC de enovelamento, para diversas cadeias tipo-proteínas, em função da ordem de contato, na temperatura $T = 1,02$; o coeficiente de correlação é 0,73.

figurações nativas com ordem de contato e ordem de longo alcance pequenas ($\chi < 0,32$ e $\chi_{LR} < 0,30$) tendem a se enovelar rapidamente. Na **Figura 4.51** mostra-se o tempo de folding (em escala logarítmica) em função do número, ϕ , de estruturas *STTS* da configuração nativa correspondente; observe-se que existe uma tendência do tempo de folding da cadeia em diminuir na medida em que o número ϕ aumenta; o coeficiente linear da reta ajustada é $-0,69$.

4.3.6 Discussão geral dos aspectos cinéticos dos resultados

De uma maneira geral, pode-se dizer que as cadeias com estados nativos caracterizados por ordem de contato e ordem de longo alcance pequenas, e número de estruturas *STTS* elevado, apresentam um intervalo relativamente amplo de temperatura (entre aproximadamente 0,9 e 2,1) onde o tempo de folding é bastante pequeno e independente das condições iniciais (como pode ser visto na **Figura**

Figura 4.50: Tempo MC de enovelamento, para diversas cadeias tipo-proteínas, em função da ordem de longo alcance, na temperatura $T = 1,02$; o coeficiente de correlação é 0,69.

Figura 4.51: Tempo MC de enovelamento, para diversas cadeias tipo-proteínas, em função do número de estruturas *STTS*, na temperatura $T = 1,02$; o coeficiente de correlação é $-0,67$.

4.29 e, também, na **Figura 4.56** mais adiante). Estas cadeias tendem a visitar apenas uma configuração maximamente compacta que corresponde à CSA nativa, sendo a frequência na mesma significativa desde a menor temperatura em que a cadeia se enovela, até a temperatura T_{max} , correspondente ao pico da capacidade térmica.

Por outro lado, as cadeias correspondentes às configurações CSA com ordem de contato e ordem de longo alcance elevadas e número de estruturas *STTS* pequeno, apresentam um intervalo de temperatura relativamente pequeno (entre aproximadamente 1,5 e 1,8) onde o tempo para a configuração nativa ser encontrada, pela primeira vez, é menor do que 20 mil e é fortemente dependente das condições iniciais, conforme **Figura 4.38**. Adicionalmente, estas cadeias tendem a visitar mais de uma configuração maximamente compacta, além da nativa, sendo a frequência nesta última dependente das condições iniciais e tendendo a zero em temperaturas menores do que a temperatura correspondente ao pico da capacidade térmica.

As distribuições de energia, correspondentes às menores temperaturas em que ocorre o enovelamento, apresentam-se mais dispersas para as cadeias cujas configurações nativas têm ordem de contato, χ , e ordem de longo alcance, χ_{LR} , grandes e número ϕ de estruturas *STTS* pequeno, do que para as cadeias com χ e χ_{LR} pequenos e ϕ grande. Comparando o comportamento da evolução da energia em função do tempo de simulação, para as primeiras temperaturas em que ocorre o enovelamento, percebe-se que as cadeias com estados nativos caracterizados por χ pequena e ϕ elevado (**Figura 4.33**) apresentam a energia diminuindo gradualmente em direção à da configuração nativa, ao longo da simulação, enquanto que as cadeias com estados nativos caracterizados por χ elevada e ϕ pequeno, apresentam, inicialmente, a energia evoluindo em direção à energia da configuração nativa, mas, ao encontrá-la, não permanece nesta por muito tempo, oscilando entre diversos valores (**Figura 4.43**).

4.4 Importância das Restrições Estéricas

4.4.1 Estudo termodinâmico da cadeia 870

Tomando-se como exemplo a cadeia 870, cuja configuração nativa é a CSA 870, que apresenta ordem de contato $\chi = 0,2381$ e número de estruturas $STTS$, $\phi = 8$, mostram-se a seguir os resultados das simulações da desnaturação por calor, partindo-se da configuração nativa em cada temperatura T , usando-se o **potencial estéreo-hidrofóbico** —aquele que considera, além da energia hidrofóbica, as restrições estéricas, onde a energia inter-monômero é $E_{ij}^* = h_i + h_j + c_{ij}$ — e o **potencial hidrofóbico** —onde as restrições estéricas não são consideradas e a energia de contato inter-monômero é $E_{ij} = h_i + h_j$. A **Figura 4.52** mostra a capacidade térmica, C/k , em função da temperatura, obtida com o potencial estéreo-hidrofóbico (círculos cheios) e com o potencial hidrofóbico (círculos vazios). A **Figura 4.53** mostra a atividade configuracional, em função da temperatura, com o potencial estéreo-hidrofóbico (símbolos cheios) e com o potencial hidrofóbico (símbolos vazios); os quadrados representam a média do número relativo do total de contatos, e, os círculos, a média dos valores relativos dos contatos nativos; os triângulos, a frequência relativa da cadeia na configuração nativa.

Por sua vez, a **Figura 4.54** mostra a atividade configuracional, em função da temperatura, monitorada em termos da distância ponta-a-ponta (quadrados) e do raio de giração (círculos), obtida com o potencial estéreo-hidrofóbico (símbolos cheios) e com o potencial hidrofóbico (símbolos vazios).

Para compreender, com mais detalhes, a atividade configuracional, a baixa temperatura, usando-se o **potencial hidrofóbico**, mostra-se, na **Figura 4.55**, a evolução, ao longo do tempo de simulação, do número total de contatos e do número de contatos nativos, para a temperatura $T = 0,63$.

Figura 4.52: Capacidade térmica, C/k , em função da temperatura, T , com o potencial estéreo-hidrofóbico (círculos cheios) e com o potencial hidrofóbico (círculos vazios), para a cadeia 870.

Figura 4.53: Atividade configuracional, em função da temperatura, T , considerando-se o potencial estéreo-hidrofóbico (símbolos cheios) e o potencial hidrofóbico (símbolos vazios). Os quadrados representam o valor relativo médio do total de contatos; os círculos, o valor relativo médio dos contatos nativos; e os triângulos, a frequência relativa na configuração nativa, para a cadeia 870.

Figura 4.54: Atividade configuracional da cadeia 870, em função da temperatura, T , obtida com os potenciais estéreo-hidrofóbico (símbolos cheios) e hidrofóbico (símbolos vazios), onde os quadrados representam a distância ponta-a-ponta, e os círculos, o raio de giração.

Figura 4.55: Evolução do número de contatos total (linha escura) e nativos (linha cinza), para a temperatura $T = 0,63$, usando-se o potencial hidrofóbico; fica evidente a grande maleabilidade do glóbulo: o total de contatos permanece acima de 80% durante toda a simulação, enquanto os contatos nativos caem “bruscamente”.

4.4.2 Estudo cinético da cadeia 5365

Com o objetivo de avaliar o papel das restrições estéricas na cinética do enovelamento, foram executadas simulações de folding, para diversas cadeias, usando-se os dois potenciais diferentes: (i) o **potencial estéreo-hidrofóbico** que vem sendo estudado neste trabalho, lembrando-se que este considera, além da energia hidrofóbica, um conjunto de restrições estéricas que impedem que certos monômeros estabeleçam contato; e (ii) o **potencial hidrofóbico**, aquele que considera somente a energia de interação hidrofóbica. Mostram-se a seguir os resultados para a cadeia 5365, que apresenta como configuração nativa a CSA 5365, cuja ordem de contato é $\chi = 0,2619$, ordem de longo alcance, $\chi_{LR} = 0,2143$, e número de estruturas STTS, $\phi = 8$. As **Figuras 4.56 e 4.57** mostram o tempo para esta cadeia encontrar, pela primeira vez, a configuração nativa, em função da temperatura, considerando-se o **potencial estéreo-hidrofóbico** (para 30 condições iniciais diferentes) e o **potencial hidrofóbico** (para 15 condições iniciais diferentes), respectivamente. Veja-se que, no caso do potencial estéreo-hidrofóbico, o tempo para que a configuração nativa seja encontrada mantém-se muito pequeno (próximo de zero) num intervalo de temperatura relativamente grande, entre 1,0 e 2,2. Por outro lado, o tempo para a cadeia encontrar a configuração nativa, correspondente ao potencial hidrofóbico, apresenta-se reduzido num intervalo de temperatura mais estreito, entre $T \simeq 1,1$ e 1,3.

Na **Figura 4.58** mostra-se a frequência com qua a cadeia encontra a configuração nativa, obtida com os potenciais estéreo-hidrofóbico (símbolos cheios) e hidrofóbico (símbolos vazios), onde se observa que o primeiro garante à cadeia estabilidade na configuração nativa, contrastando com o potencial hidrofóbico que definitivamente não consegue assegurar tal estabilidade. Observando-se o número de contatos nativos das configurações CSA visitadas (**Figura 4.59**), percebe-se que o potencial hidrofóbico (colunas vazias) permite que a cadeia

Figura 4.56: Tempo MC de enovelamento da cadeia 5365, em função da temperatura, T , para 30 condições iniciais diferentes, usando-se o potencial estéreo-hidrofóbico. Note-se que, no intervalo $1 < T < 2$, o tempo de folding é praticamente independente das condições iniciais.

Figura 4.57: Tempo MC de enovelamento da cadeia 5365, em função da temperatura, T , para 15 condições iniciais diferentes, usando-se o potencial hidrofóbico.

Figura 4.58: Frequência relativa com que a cadeia 5365 encontra a configuração nativa, em função da temperatura, T ; os pontos cheios correspondem ao potencial estereo-hidrofóbico, e os vazios, ao potencial hidrofóbico.

Figura 4.59: Frequência absoluta com que a cadeia 5365 visita uma configuração CSA; esta última é identificada pelo número de contatos nativos. As colunas vazias correspondem ao potencial hidrofóbico, e a coluna cheia, ao potencial estereo-hidrofóbico.

Figura 4.60: Energia configuracional, em função do tempo, para a cadeia 5365, em $T = 1,03$, usando-se potencial estereo-hidrofóbico.

Figura 4.61: Energia configuracional, em função do tempo, para a cadeia 5365, em $T = 1,03$, usando-se o potencial hidrofóbico.

encontre diversas configurações CSA, além da nativa, enquanto o potencial estéreo-hidrofóbico (coluna cheia) conduz a cadeia a somente uma configuração CSA, que é a própria configuração nativa (com 28 contatos nativos).

Nas **Figuras 4.60 e 4.61** mostra-se a evolução da energia, em função do tempo MC de simulação, no processo de enovelamento da cadeia 5365, para a temperatura $T = 1,03$, usando-se o potencial estéreo-hidrofóbico e o potencial hidrofóbico, respectivamente; deve ser observado que a energia da configuração nativa é $E^{(n)} = -60,0$.

4.4.3 Discussão do papel das restrições estéricas

A análise comparativa dos resultados das simulações de desnaturação por calor, obtidos com o potencial hidrofóbico e com o potencial estéreo-hidrofóbico, evidencia o papel das restrições estéricas (Tarragó et al., 2003). As curvas da **capacidade térmica**, C/k , em função da temperatura, T , obtidas com estes dois potenciais, são bastante diferenciadas, como revela a **Figura 4.52**: no caso do **potencial hidrofóbico**, o pico da capacidade térmica ocorre numa temperatura $T \simeq 0,8$, e não é pronunciada; a curva C/k , em função de T (círculos vazios), revela que, à medida que a temperatura se modifica, a cadeia troca uma quantidade de energia relativamente pequena com o meio solvente. Por outro lado, com o **potencial estéreo-hidrofóbico**, a forma da curva C/k , em função de T , apresenta um pico bem pronunciado (círculos cheios), onde o valor máximo de C/k ocorre em $T \simeq 1,5$; em $T = T'$ se observam perturbações (notar o ponto cheio fora da tendência da curva) que são dependentes da história configuracional, o que não se espera que aconteça para um sistema em equilíbrio. A inexistência destas perturbações com o potencial hidrofóbico indica que, justamente na temperatura T' , existe um balanço entre os mínimos locais de energia e as restrições estéricas; T' é uma temperatura suficientemente alta para assegurar que diferentes configurações sejam acessadas, e é, ao mesmo

tempo, suficientemente pequena para aprisionar a cadeia em mínimos locais de energia, assemelhando-se a desnaturação a frio.

Nas **Figuras 4.53 e 4.54** comparam-se as atividades configuracionais em função da temperatura, T , nas situações com e sem restrições, correspondentes ao potencial estéreo-hidrofóbico (símbolos cheios) e ao potencial hidrofóbico (símbolos vazios), respectivamente. Simbolizando o **número relativo médio do total de contatos** por Ψ_S , para o **sistema sem restrições**, e por Ψ_R , para o **sistema com restrições** (correspondendo aos quadrados da **Figura 4.53**), pode-se dizer que, no intervalo entre $0,5 < T < 3,0$, Ψ_S diminui lentamente à medida que a temperatura aumenta, enquanto Ψ_R apresenta uma forma sigmoideal, rapidamente reduzindo o número de monômeros em contato para valores bem menores do que os correspondentes para Ψ_S . Este efeito das restrições estéricas indica que a porção do espaço conformacional, correspondente às conformações tipo-globulares, é peculiarmente afetada: nas regiões de temperatura em que o glóbulo é mais compacto (Ψ_R elevados), o número de configurações, que separa as configurações da cadeia que são mais distendidas das mais compactas, é severamente reduzido, o que explica o pronunciado pico da capacidade térmica. Este pico indica que a energia interna da cadeia e a entropia exibem um salto, rapidamente mudando suas respectivas quantidades para temperaturas em torno de $T = T_{max}$.

De forma semelhante, analisa-se o comportamento do **número relativo de contatos nativos** para o sistema com restrições estéricas, $\Psi_R^{(n)}$, e para o sistema sem restrições estéricas, $\Psi_S^{(n)}$. Quando as restrições não são consideradas, ou seja, para o **potencial hidrofóbico**, observa-se um número relativamente pequeno de $\Psi_S^{(n)}$ para a maioria das temperaturas, como mostrado na **Figura 4.53** (círculos vazios). Mas, para temperaturas pequenas o suficiente, quando o glóbulo é bastante compacto ($\Psi_S > 0,8$), o número relativo dos contatos nativos $\Psi_S^{(n)}$ aumenta consideravelmente, de tal forma que $\Psi_S^{(n)}$ rapidamente aproxima-se

de Ψ_S , mas continua sendo menor do que este. Observando-se pormenorizada-mente a evolução configuracional ao longo da simulação, como ilustra a **Figura 4.55**, onde pode ser visto o número de contatos total e nativos, para a temperatura $T = 0,63$, fica evidente a grande maleabilidade do glóbulo, uma vez que o total de contatos permanece acima de 80% durante toda a simulação, enquanto os contatos nativos caem “bruscamente” desde o valor máximo (28) até o valor médio de aproximadamente oito contatos nativos; ou seja, os contatos nativos foram rapidamente substituídos por contatos não-nativos, de tal maneira que o caráter globular da conformação foi preservado.

Retornando à **Figura 4.53**, para o sistema com restrições, ou seja, usando-se o **potencial estéreo-hidrofóbico**, o número relativo médio de contatos nativos $\Psi_R^{(n)}$ (círculos sólidos) segue de perto Ψ_R (quadrados sólidos). Para $T < T'$, quase todos os contatos são os nativos, isto é, a condição $\Psi_R^{(n)} = \Psi_R$ é praticamente satisfeita; mas, mesmo para temperaturas mais altas, $T > 2$, a maioria dos contatos são nativos. Este resultado é compreendido como um efeito das especificidades estéricas; na **Figura 4.54**, para temperaturas $T > 1,5$, o raio de giração para o sistema com restrições é significativamente maior do que para o sistema sem restrições. Então, em média, muitos dos contatos da cadeia são contatos locais, mas tais contatos estão restritos pelos impedimentos estéricos, que favorecem os contatos nativos como consequência do desenho da seqüência. Estes resultados também indicam que os impedimentos estéricos atuam como guias do folding, induzindo a cadeia a formar contatos nativos, mesmo em temperaturas acima de T_{max} .

Tratando-se do **potencial estéreo-hidrofóbico**, por ser um resultado marcante, salienta-se que, no pico da capacidade térmica, $T_{max} \simeq 1,5$, o **número médio de contatos nativos** aproxima-se de 50%, isto é $\Psi_R^{(n)} \simeq 1/2$ (**Figura 4.53**). Conseqüentemente, T_{max} pode ser vista como a temperatura que separa dois comportamentos distintos da atividade configuracional: abaixo de T_{max}

a atividade configuracional —limitada pelas restrições estéricas e pela flutuação térmica relativamente baixa— define uma forma globular compacta para a cadeia ($\Psi_R > 1/2$), que rapidamente se torna mais densa para temperaturas menores; enquanto que aumentando a temperatura acima de T_{max} , a forma globular da cadeia é destruída, sendo as configurações destendidas as estatisticamente mais significativas.

Finalmente, analisa-se a **frequência relativa na configuração nativa**, ou seja, a frequência relativa f com que a cadeia é encontrada na configuração nativa. Para o **sistema sem restrições**, mesmo em temperaturas bem baixas, a configuração nativa é raramente visitada ($f_S < 10^{-5}$, triângulos vazios na **Figura 4.53**), significando que o sistema não apresenta estabilidade suficiente. Entretanto, a frequência relativa na nativa f_S não ser exatamente zero tem um importante significado: isto sugere que o potencial hidrofóbico, do tipo $E_{i,j} = h_i + h_j$, é, de fato, eficiente em empacotar e em conduzir a cadeia ao estado nativo, embora falhe em sustentá-la neste último. Mas, quando as especificidades interacionais apropriadas são introduzidas —isto é, usando-se o **potencial estéreo-hidrofóbico**— elas atuam como se fossem labirintos topológicos para a configuração nativa. Esta barreira configuracional aumenta sua eficiência na medida que a temperatura diminui a partir de T_{max} , e, conseqüentemente, a frequência relativa no estado nativo, f_R (triângulos cheios, **Figura 4.53**), assume valores significativos, alcançando 10% próximo de $T = 1$, numericamente f_R é, no mínimo, cinco ordens de grandeza maior do que f_S . Note-se que, justamente em T' , o valor de f_R é menor do que a tendência da curva sugere, o que está de acordo com os comentários anteriores sobre a capacidade térmica.

4.5 Referências Bibliográficas

Plaxco, K. W. et al. (1998) Contact Order, Transition State Placement and the Folding Rates of Single Domain Proteins. *J. Mol. Biol.*, **277**: 985-994

da Silva, R. A. et al. (2001) Deterministic Folding: The Role of Entropic Forces and Steric Specificities. *J. Chem. Phys.*, **114**: 4235-4242

da Silva, R. A. (2002) **Enovelamento Determinístico de Macromoléculas**. Tese de Doutorado, FFCLRP, Departamento de Física e Matemática, Universidade de São Paulo.

Tarragó, M. E. P. et al. (2003) Steric Constraints as Folding Coadjuvant. *Phys. Rev. E*, no prelo.

Capítulo 5

CONCLUSÃO

Neste trabalho, um potencial do tipo hidrofóbico e um conjunto de especificidades estéricas, em um modelo em rede simplificado, são propostos como os dois ingredientes básicos para o processo de enovelamento proteico, constituindo o **potencial estéreo-hidrofóbico**. O presente modelo diferencia-se, em vários aspectos, dos demais modelos simplificados usados no estudo do processo de *fold-ing* de proteínas, sobretudo no que concerne à sua generalidade ao se considerar múltiplas conformações nativas caracterizadas por seus parâmetros topológicos. As simulações de desnaturação por calor de dezenas de cadeias tipo-proteínas —correspondentes a configurações nativas caracterizadas pelos mais diversos valores dos parâmetros topológicos, como ordem de contato, ordem de longo alcance, e número de estruturas *STTS*— permitiram analisar, pormenorizada-mente, o processo de enovelamento proteico, sob os aspectos termodinâmico e cinético.

A análise dos valores máximos da capacidade térmica (C_{max}/k) e a da “largura à meia altura” da curva da capacidade térmica, C/k (em função da temperatura, T), em função dos parâmetros topológicos, revela que existe uma tendência de a curva C/k apresentar-se “mais baixa” e “mais larga” na medida em que a ordem de contato, χ , da configuração nativa aumenta, ao mesmo tempo em que

o número de estruturas *STTS*, ϕ , diminui. Claramente, os resultados das simulações, para as cadeias que apresentam configurações nativas com baixo valor de χ e elevado valor de ϕ , são bastante diferenciados daqueles obtidos para as cadeias correspondentes a configurações com χ elevado e ϕ pequeno, evidenciando que o potencial estéreo-hidrofóbico adotado neste trabalho permite relacionar o comportamento termodinâmico da cadeia tipo-proteína com os atributos topológicos da sua configuração nativa.

De uma maneira geral, observa-se que existe uma temperatura, T' , em torno da qual ocorrem perturbações em todas as grandezas termodinâmicas monitoradas nas simulações de desnaturação por calor, significando que estas grandezas, em T' , são dependentes das condições iniciais. Considerando-se que as simulações de desnaturação partem da configuração nativa, em temperaturas baixas, a energia térmica, kT , apesar de pequena, é suficiente para atuar de forma diferenciada nas cadeias correspondentes a configurações nativas com diferentes parâmetros topológicos: as cadeias com estruturas nativas com maior ordem de contato, χ , e menor número de estruturas *STTS*, ϕ , tendem a apresentar maior rigidez em relação à exploração do espaço conformacional em torno da estrutura nativa; por outro lado, as cadeias com configurações nativas com χ pequeno e ϕ elevado são mais maleáveis e, induzidas pelas flutuações térmicas, exploram um espaço conformacional bem mais vasto em torno da configuração nativa, mesmo em baixas temperaturas.

A temperatura em que ocorre o máximo da capacidade térmica, T_{max} , para todas as configurações estudadas, marca a inflexão nas curvas dos seguintes parâmetros: número relativo médio do total de contatos (em T_{max} , este número aproxima-se de 50%), número relativo médio de contatos nativos, distância ponta-a-ponta e raio de giração. Esta inflexão indica mudanças na atividade configuracional. Em T_{max} , o número médio do total de contatos (Ψ_R) aproxima-se de 50%; conseqüentemente, T_{max} pode ser vista como a temperatura que

separa dois comportamentos distintos da atividade configuracional: abaixo de T_{max} a atividade configuracional —limitada pelas restrições estéricas e pela flutuação térmica relativamente baixa— define uma forma globular compacta para a cadeia ($\Psi_R > 1/2$), que rapidamente se torna ainda mais compacta para temperaturas menores; enquanto que, aumentando a temperatura acima de T_{max} , a forma globular da cadeia é destruída, sendo as configurações distendidas estatisticamente mais significativas. Além disso, observou-se que existe uma tendência em T_{max} aumentar na medida em que a ordem de contato aumenta e o número de estruturas *STTS* diminui.

As simulações de enovelamento, das mais de 40 cadeias, mostraram que as cadeias com configurações nativas caracterizadas por ordem de contato, χ , e ordem de longo alcance, χ_{LR} , pequenas, e número, ϕ , de estruturas *STTS* elevado, apresentam um intervalo relativamente amplo de temperatura (entre aproximadamente 0,9 e 2,1), em que o tempo de **folding** (tempo necessário para a cadeia encontrar pela primeira vez a configuração nativa) é bastante pequeno e independente das condições iniciais. Estas cadeias tendem a visitar apenas uma configuração maximamente compacta correspondente à CSA nativa, sendo a frequência na configuração nativa significativa desde a primeira temperatura, em que a cadeia se enovela, até a temperatura T_{max} , onde ocorre o pico da capacidade térmica. Por outro lado, as cadeias com configurações nativas com χ e χ_{LR} elevados, e ϕ pequeno, apresentam um intervalo de temperatura relativamente pequeno (entre aproximadamente 1,5 e 1,8), onde o tempo de **folding** tem um valor moderado e é fortemente dependente das condições iniciais; estas cadeias tendem a visitar mais de uma configuração maximamente compacta, além da CSA nativa, sendo a frequência com que a cadeia visita a configuração nativa dependente das condições iniciais, tendendo a zero em temperaturas T menores do que a temperatura correspondente ao pico da capacidade térmica, T_{max} .

Considerando-se, ainda, o potencial estérico-hidrofóbico, cabem alguns co-

mentários em relação às duas configurações CSA, apresentadas anteriormente na literatura como “especiais”: uma, encontrada por Govindarajan e Goldstein (1995) como a “mais otimizada” em relação a energia, e a outra, encontrada por Li e colaboradores (1996), como a “mais desenhável” —chamadas aqui de CSA de Goldstein e CSA de Li, respectivamente. A primeira apresenta ordem de contato moderada ($\chi = 0,2989$) e a segunda, ordem de contato elevada ($\chi = 0,4021$), mas ambas apresentam número de estruturas *STTS*, $\phi = 0$. Verificou-se que, de acordo com a regra de *desing* (sintaxe) adotada neste estudo, a CSA de Goldstein apresenta energia bastante elevada, $E^{(n)} = -58,8$, e não o bastante pequena para ser considerada “otimizada energeticamente”. Por outro lado, a CSA de Li apresenta energia bastante baixa, $E^{(n)} = -67,6$, e, apesar de não se ter verificado sua “desenhabilidade”, as simulações de *folding* realizadas indicaram que a “cadeia de Li” não apresenta as propriedades esperadas para uma cadeia do tipo-proteína. De uma maneira geral, o presente estudo possibilitou a indentificação de correlação linear negativa significativa entre a energia da configuração nativa, $E^{(n)}$, e ordem de contato, χ , com coeficiente de correlação igual a $-0,78$; e, também, uma correlação negativa significativa entre $E^{(n)}$ e ordem de longo alcance, χ_{LR} , com coeficiente de correlação igual a $-0,74$. Conseqüentemente, pode-se dizer que as configurações nativas, com valores baixos para χ e para χ_{LR} tendem a apresentar energias maiores, em torno de $E^{(n)} = -60$, comparadas às configurações nativas com χ e χ_{LR} mais altas, que apresentam energias em torno de $E^{(n)} = -67$.

Realizaram-se, também, simulações para diversas cadeias usando-se o potencial hidrofóbico (sem se considerar as restrições estéricas). Encontrou-se que a energia de contato, baseada apenas no potencial hidrofóbico, garante que a cadeia se empacote e que o estado nativo seja visitado, mas não provê estabilidade configuracional. Quando um conjunto de restrições estéricas apropriadas é adicionado ao potencial hidrofóbico (resultando no **potencial estéreo-hidrofóbico**),

tais especificidades estéricas interacionais atuam como coadjuvantes no processo de **folding**, selecionando caminhos para o enovelamento e aumentando de forma significativa as condições de estabilidade do glóbulo na configuração nativa. Comparando-se os resultados das simulações Monte Carlo, onde se utilizou o **potencial hidrofóbico**, com os resultados referentes às simulações com o **potencial estéreo-hidrofóbico**, verificou-se que as restrições estéricas adotadas alteram drasticamente a atividade configuracional do sistema. Mais especificamente, tais **restrições estéricas**: (i) transformam a curva original da capacidade térmica, que é bastante larga —obtida com o potencial hidrofóbico do tipo energia de contato—, em uma curva simétrica, com um pico pronunciado; e (ii) aumentam em cinco ou mais ordens de grandeza a frequência com que a cadeia permanece no estado nativo.

Esses resultados sugerem que o problema do **folding** de proteínas globulares, em muitas situações, pode ser descrito como um processo no qual os mecanismos que atuam na condução da cadeia à conformação nativa estão desacoplados daqueles que garantem estabilidade ao glóbulo. De acordo com este ponto de vista, o código estereoquímico, ou seja, o padrão de hidrofobicidade e as especificidades interacionais de cada resíduo, é responsável pelo mecanismo no qual a cadeia alcança o estado nativo, que é considerado um estado único e especial. Uma vez que a cadeia se encontre no estado nativo, e somente neste, todos os fatores energéticos e estéricos envolvidos no processo são tais que maximizam as condições de estabilidade para o glóbulo. Pode-se dizer, então, que: (1) muitas ligações de hidrogênio estão agora protegidas do meio e, como a competição com o solvente foi minimizada, elas efetivamente contribuem para a estabilidade do glóbulo; (2) as complementariedades estéricas dos resíduos aumentam a área interna em contato, reduzindo ao mesmo tempo o contato externo com o solvente, o que produz uma contribuição efetiva para a estabilidade energética do glóbulo; (3) as especificidades estéricas dos resíduos funcionam, também, como empecil-

hos no sentido de que aprisionam topologicamente a cadeia, assim como num jogo de “quebra-cabeças em 3D”, ajudando a manter a cadeia na configuração nativa.