

Aspectos de topologia e mutação no processo
de enovelamento e evolução
de proteínas.

Leandro Cristante de Oliveira

Tese de doutoramento direto
Pós-graduação em Biofísica Molecular

Oliveira, Leandro Cristante de.

Biofísica molecular : enovelamento de proteínas / Leandro Cristante de Oliveira. - São José do Rio Preto : [s.n.], 2008.

56 f. : il. ; 30 cm.

Orientador: Vitor Barbanti Pereira Leite

Orientador (exterior): José Nelson Onuchic

Co-orientador: Jorge Chahine

Tese (doutorado) – Universidade Estadual Paulista, Instituto de Biociências, Letras e Ciências Exatas

1. Biofísica. 2. Física estatística. 3. Proteínas – Estrutura. 4. Enovelamento de proteínas. I. Leite, Vitor Barbanti Pereira. II. Onuchic, José Nelson. III. Chahine, Jorge. IV. Universidade Estadual Paulista, Instituto de Biociências, Letras e Ciências Exatas. V. Aspectos de Topologia e mutação no processo de enovelamento e evolução de proteínas.

CDU – 577.3

Aspectos de topologia e mutação no processo de enovelamento e evolução de proteínas

Leandro Cristante de Oliveira

Tese apresentada ao
Instituto de Biociências, Letras e Ciências
Exatas da Universidade Estadual Paulista
“Júlio de Mesquita Filho”, Campus de
São José do Rio Preto-SP, como parte das
exigências para a obtenção do Título de
Doutor em Biofísica Molecular.

Orientador: Prof. Dr. Vitor Barbanti Pereira Leite

Orientador do estágio doutoral no exterior: Prof. Dr. José Nelson Onuchic

Co-orientador: Prof. Dr. Jorge Chahine

São José do Rio Preto - SP

Abril de 2008.

*As pessoas que acreditaram no meu trabalho
em especial aos meus pais e avós.*

“Se você mantém a calma, quando todos perderam a cabeça, é
porque você não captou o problema.”

(Axioma do Lento Tapado)

Agradecimentos

Muitas pessoas especiais passaram pela minha vida ao longo desse trabalho, algumas fundamentais, outras não. Em sua grande maioria, pessoas pela qual tenho um carinho especial. Todavia, dificilmente todas caberiam aqui, porém, sinto-me privilegiado por tê-las conhecido e por isso devo meu agradecimento.

Aos amigos Jorge Chahine e Vitor Leite, pessoas que me abriram portas e a mente para uma nova visão de vida e ciência. Considero ambos como meus segundos pais e por isso, tão queridos que os desejo sempre próximos. À Aline Bruni pela ajuda com as idéias de quimiometria testadas em uma fase do trabalho e pela amizade.

Ao Sidney, grande Moreira, pela força, pelas baladas e apoio em toda caminhada e em especial no final desse trabalho pela sua disponibilidade e motivação em ler o material, colocar suas sugestões e ajudar no que foi possível. Ele continua se achando imbatível no bilhar, no entanto continua tomando uma lavada no xadrez!

Ao Ronaldinho pelas discussões no grupo e pelos momentos de alegria, desde a piada da bolachinha até as bolas foras. Amigo querido que tenho como “parente”. E Isadora, não me esqueci de você, pelas adoráveis e turbulentas discussões, respeito e amizade.

Ao amigo José Ésio que é o maior furão que já passou pelo departamento de física, mas continua um cearense super gente boa, quando o encontramos.

À cearense Denise pelas discussões e confiança em momentos cruciais.

Por todo suporte fornecido por Daniel Schultz no exterior com abrigo, alimentação, baladas, discussões, auxílio na adaptação cultural e apoio com a língua Inglesa e a toda galerinha de San Diego: Rachael Small, Joe Hegler, Stuart Bogatko, Kelly e Mary McCormick, Troung, Alex Schug, Paul Whithford, Jeff, Sashi e Chanbong, além de José Nelson Onuchic pela ajuda e orientação. À Stéfanie Sabbag, Antony e Margareth pelas aulas de inglês e ajuda com o idioma ainda no Brasil.

Aos amigos de sala e de departamento Luciana Alonso, Ricardo Hildebrand, André Martins, Ana Helena, Diego Nolasco (em especial por adorar a minha playlist), Diogo, Priscila, Sabrina e Márcia e aos demais que sempre trombaram comigo pelos corredores.

À amiga Juliana Pirez da Silva que desde a graduação em física, sempre manteve contato e que mesmo longe me faz sentir sempre próximo à ela e amparado.

Às funcionárias Rosemar Brena e Ilva que sempre foram altamente prestativas e resolveram com eficiências todos os meus problemas de documentação, além das boas conversas e demais professores e funcionários do departamento.

Aos meus pais pela vida, sentimentos, valores, amor e guerra que me fazem hoje um homem feliz, confiante e seguro de si mesmo e à aquela que tem o nome da santa, a beleza de Helena e com um sorriso enche minha vida de paz.

Agradecimentos finais à CAPES, pelo auxílio financeiro e à FAPESP.

Resumo

A topologia do estado nativo de uma proteína desempenha um papel crucial no processo de enovelamento. Neste trabalho uma nova aproximação utilizando aspectos topológicos para investigar a evolução protéica é apresentada. O modelo utiliza uma rede cúbica $3 \times 3 \times 3$ de 27 monômeros e um mapa de conexões entre diferentes conformações em espaço de fase estrutural e de seqüência. Desenhamos a melhor seqüência não frustrada para cada uma das 103346 conformações maximamente compactas usando um algoritmo que maximiza o número de tipos de monômeros na seqüência. Isto significa que cada seqüência não pode possuir contatos desfavoráveis. O número máximo de tipos de monômeros é 5. A seqüência-conformação é considerada “protein-like” se ela tem uma única conformação de mais baixa energia, além de acessibilidade e robustez. De todas as conformações maximamente compactas, somente 4,75% geraram seqüências “protein-like”, o qual são o alvo neste estudo. Com esses dados realizamos simulações de Monte Carlo (MC) no qual examinamos as melhores seqüências-estruturas baseando-se no Z_{Score} . A simulação é iniciada com uma seqüência aleatória no qual é testada em todas as conformações, seguindo as regras estipuladas por MC. Se o Z_{Score} aumenta, assumimos que a nova conformação é mais estável que a anterior. Esse processo é repetido até que as seqüências otimamente desenhadas (com mais alto Z_{Score}) são alcançadas. Mantendo as trajetórias originadas via MC, um mapa de conectividade seqüências-estruturas é obtido. Os resultados mostram trajetórias conectadas com estruturas com baixos valores de Z_{Score} . O aumento do Z_{Score} ao longo da simulação conduz a um pequeno grupo de conformações preferenciais. O modelo sugere um funil de estruturas para a evolução de proteínas no qual as estruturas do fundo estão associadas com o

“motif” de uma proteína. Esse resultado pode ser uma possível explicação se comparado ao grande número de seqüências nos bancos de dados de proteínas (PDB). A análise do melhor cenário para seqüências frustradas também foi realizado. Um comportamento diferente para seqüências otimamente desenhadas se comparado com frustradas foi observado quando a hidrofobicidade do sistema muda. Proteínas com frustrações em um ambiente de alta hidrofobicidade tem um processo de enovelamento mais estável, proveniente do colapso hidrofóbico. Este resultado sugere que a quantidade de frustração pode determinar quando esse cenário é favorável.

Como assunto complementar modelos mais realísticos baseados na estrutura nativa e no espaço contínuo foram desenvolvidos. O objetivo é verificar a sensibilidade do sistema através de mudanças nos valores dos pré-fatores da Hamiltoniana. A proteína foi reduzida a somente esferas representando os C_α e $C_\alpha C_\beta$ e Dinâmicas Moleculares foram realizadas. Nos dois casos o carbono C_α é mantido na sua posição nativa enquanto o C_β é posicionado no centro de massa da cadeia lateral. Este modelo é suficiente para reproduzir com um bom acordo com os dados experimentais com a vantagem de baixo tempo computacional. A variação nos parâmetros de energia mostraram o mecanismo de enovelamento para uma ampla gama de valores. Todavia, propriedades energéticas podem ser sensíveis à detalhes específicos da simulação, como apresentado.

Palavras-chave: topologia, desenhabilidade, robustez, modelo de rede, modelo baseado em estrutura, enovelamento de proteínas

Abstract

The topology of a protein native state plays a crucial role in the folding process. In this work a new approach using topological aspects to investigate the protein evolutions is presented. The model uses the 27-mer in a cubic lattice of $3 \times 3 \times 3$, and a connection map between different conformations is found in the sequence and structural phase space. We designed the best unfrustrated sequence for each of the 103346 maximally compact conformation, using an algorithm that maximizes the number for monomers types in the sequence. This means that each sequence cannot have unfavorable contacts. The maximum number of types of monomer is 5. The sequence-conformation is considered protein-like if it has a unique lowest energy conformation, accessible and robust. Out of all maximally compact conformations, only 4,75% generated protein-like sequence, with are targeted in this study. With this data we performed a Monte Carlo simulations in which we probe for better sequence-structure based on Z_{score} . The simulation start with a random sequence and it is tested all conformations, finding its conformations according to the Monte Carlo rules. If the Z_{score} increases, we assume that the new conformation is more stable than the previous. This process is repeated until the optimally designed sequence (with the highest Z_{score}) is reached. Keeping track of all the Monte Carlo trajectories, a map of connectivity of sequence-structures is obtained. The results shows trajectories connected with structures of low Z_{score} values. The increase of Z_{score} along of the simulation leads to a small group of preferred conformations. The model suggest a funnel like structure for folding evolution, in which the structures at the bottom of the funnel are associated with the motif of a protein. This result can be a possible

explanation for the restricted number of conformations compared to the large number of sequences on protein data banks.

The analyses of the best folding scenario for frustrated sequences also were performed. A different behavior to optimally designed and frustrated sequences was observed when the hydrophobicity of the system changes. Proteins with frustrations in a high hydrophobic environment have a folding process more stable, provided by the hydrophobic collapse. This result suggests that the amount of frustration can determine how much the scenario is favorable.

As a complementary issue models more realistic based in the native state and off-lattice were developed. The goal is to verify the system sensibility through changes on pre-factor of the Hamiltonian. The protein was reduced to a just C_α -bead and $C_\alpha C_\beta$ -beads representation. Molecular Dynamics (MD) was performed and. On the both cases the C_α -carbon is kept in its native place while the C_β is positioned in side chain center of mass. This model is enough to reproduce with a good agreement the experimental data with the advantage of low computational time. The variation on the energetic parameter showed the folding mechanism to a broad range of values. However, energetic properties can be sensitive to specify detail of the simulation, as presented.

Keywords: Topology, designability, robustness, lattice model, based structure model, protein folding

Sumário

1	Introdução	1
2	Os modelos	5
2.1	O modelo “protein-like em rede cúbica”	5
2.2	Simulações via método de Monte Carlo	11
2.2.1	Cálculo do tempo de enovelamento	11
2.2.2	Estudo evolutivo	12
2.2.3	Termodinâmica do modelo de rede	13
2.2.4	Método do histograma simples	14
2.3	Modelos baseados em estrutura	16
2.3.1	Construção do Modelo	16
2.3.2	Simulações via Dinâmica Molecular	18
2.3.3	Método dos múltiplos histogramas	19
2.3.4	Outras avaliações	21
3	Resultados e discussões	24
4	Conclusões	35
5	Apêndices	37
5.1	Publicações	37

Lista de Figuras

2.1	<i>Representação do funcionamento do algoritmo de geração das seqüências aplicado à conformação 17467 escolhida aleatoriamente. O programa escolhe inicialmente um monômero qualquer, neste caso o indicado com a cor azul na figura B, identificando os contatos formados por este na estrutura e colorindo como apresentado em C. O processo é repetido para os novos contatos dos monômeros marcados até que todo o grupo (“cluster”) seja varrido (D). Após a busca, os itens não marcados são selecionados pela atribuição de uma nova cor e o algoritmo inicia-se novamente em busca do novo grupo (E). Em F é possível visualizar a seqüência pronta.</i>	9
2.2	<i>Tipos de movimentos para as simulações em modelo de rede cúbica. O movimento de esquina e de manivela exigem que o monômero sorteado não seja ou o primeiro ou o último da cadeia ao contrário do movimento de final, que só pode ser realizado em tais condições. Entre eles não existe nenhum privilégio ou peso ao longo da simulação.</i>	12
2.3	<i>Representação de uma torsão imprópria. Os planos formados pelos 4 átomos conectados formam um ângulo que limita a mobilidade do carbono β em torno do carbono central</i>	17
2.4	<i>O diagrama ilustra a energia dos estados desenovelado, de transição e enovelado para os casos de valores de Φ igual a 0 (à esquerda) e 1 (à direita).</i>	21

2.5	<i>Diferentes probabilidades de formação de contatos nativos ao longo da coordenada de reação Q. Cada figura mostra um intervalo de 0.10 em Q começando em 0.10. É possível acompanhar a formação das diferentes regiões da proteína. Essa figura pode ser simplificada utilizando somente a representação no intervalo do TSE e indicando a ordem dos eventos</i>	23
3.1	<i>Relação entre ordem de contato relativa e Z_{score} para as cadeias otimamente desenhadas geradas. A região “A” e “B” implicam em condições de difíceis acessibilidade e estabilidade. Em “C” o enovelamento ocorre, porém, os altos valores de OC podem ocasionar problemas em relação à estabilidade. Na região “D” encontramos as condições mais desejáveis; enovelamento rápido e estabilidade.</i>	26
3.2	<i>Perfil de enovelamento de seqüências com 5, 4 e 3 letras sem frustrações para uma mesma conformação sob os regimes de colapso e intermediário. O comportamento cinético é praticamente idêntico para as seqüências de quatro e cinco letras. Para a de três letras escrita sob a mesma estrutura, o tempo de enovelamento é relativamente maior.</i>	28
3.3	<i>Perfil de enovelamento, calor específico e energia livre para seqüências com 5 e 4 letras testadas anteriormente. O gráfico de calor específico mostra que a T_c não apresenta diferença significativa. O gráfico de energia livre mostra uma diferença na barreira de potencial onde podemos considerar a de 5 letras como praticamente zero enquanto no caso de 4 letras temos uma barreira amena . . .</i>	29

- 3.4 *Representação de um conjunto de estruturas que apresentam formato de “funil”. Somente uma conformação de mais alto Z_{score} foi visitada. Outras simulações não acessam estruturas presentes neste grupo, exceto em casos de Z_{score} extremamente baixo (presentes no topo do gráfico). Os pontos de partida da simulação estão marcados com cores e as linhas retas verticais representam a otimização da seqüência sobre a estrutura momentânea. 32*
- 3.5 *Os gráficos no topo representam duas corridas distintas e as bolas coloridas os pontos de partida. Cada uma das simulações alcançam estruturas finais diferentes: 43157 e 46646. Quando analisados, cada um dos funis, 8 coincidências foram encontradas. Neste caso, o grupo presente em um gráfico é considerado automaticamente como parte do outro e adotamos que ele assume a forma de “balde”. No gráfico presente na parte inferior em tamanho ampliado estão identificados em vermelho os pontos em comum. 34*

Lista de Tabelas

2.1	<i>Valores para as constantes da Hamiltoniana apresentada na equação 3.8.</i>	18
3.1	<i>Resultados obtidos pelo algoritmo de criação de seqüências otimamente desenhadas. Observando a tabela, é possível notar que cadeias de 5 letras são suficientemente específicas para não possuírem seqüências não degeneradas. O maior número de seqüências não degeneradas são compostas por 3 letras.</i>	24
3.2	<i>Resumo dos testes de robustez a mutações simples para estruturas pertencentes a cada uma das quatro regiões propostas. A primeira coluna representa a numeração assumida para o estudo da seqüência-conformação, na segunda o número de letras máximo permitida para a situação otimamente desenhada. A coluna seguinte apresenta o número de novas cadeias geradas através de todas as permutações simples possíveis incluindo a original. A quarta coluna mostra a quantidade de cadeias não degeneradas e em seguida, quantas delas tem a menor energia na estrutura alvo para que foram desenhadas. As três últimas colunas apresentam os valores de Z_{score}, Ordem de Contato relativa e a disposição no gráfico.</i>	27

3.3	<i>Pequeno trecho da análise das seqüências geradas à partir de permutações simples sobre a seqüência otimamente desenhada 1739. Nota-se que nenhuma das seqüências geradas apresentadas se ajustam bem à estrutura alvo e que ainda existe um aumento nos valores de Z_{score} quando uma troca de conformação ocorre.</i>	31
-----	--	----

Capítulo 1

Introdução

Proteínas são sistemas com alto grau de complexidade tanto pelas suas funções nos seres vivos quanto pela sua importância biológica, o que proporciona grande interesse quanto ao entendimento dos processos envolvidos em âmbitos físico e químico. Essas macromoléculas estão diretamente ligadas às funções de regulação, catálise e transporte entre outras e são compostas pela combinação de 20 diferentes tipos de aminoácidos.

A cadeia seqüencial de aminoácidos (seqüência) é conhecida como **estrutura primária**. A **estrutura secundária** refere-se aos ângulos formados pelas ligações peptídicas entre os aminoácidos, o que reflete em estruturas locais regulares como hélices- α e folhas- β ; a **estrutura terciária** é a disposição espacial da proteína isolada. Quando a proteína possui mais de uma cadeia, a visualização dessas é considerada **estrutura quaternária** e como exemplo temos os dímeros, trímeros, etc.

Esse tipo de classificação para essas estruturas foram introduzidas com os trabalhos de Pauling *et al.*⁴⁷ sobre estruturas secundárias para denominação de hélices- α e folhas- β e com os resultados de Watson e Crick⁵⁸ em seus estudos com o DNA e RNA.

Na década de 60, experimentos realizados por Anfinsen com a ribonu-

clease pancreática bovina em processos de desnaturação/renaturação³ *en vitro* puderam comprovar que toda a informação necessária para alcançar a estrutura funcional estava contida na seqüência de aminoácidos. Estruturas globais, e não somente locais, poderiam então ser atingidas somente com o ajuste de condições fisiológicas favoráveis em um processo controlado de maneira reversível.

Entender os mecanismos com o qual o estado biologicamente ativo de uma proteína é alcançado a partir da seqüência primária é conhecido como o *problema do enovelamento de proteínas*. Um dos grandes desafios para a elucidação deste problema está na avaliação do espaço conformacional possível. As idéias de Levinthal^{20,25,29,62} discutem a dificuldade de estabelecer uma busca aleatória através do gigantesco número de conformações desenoveladas possíveis. Para resolver o problema da busca nessa explosão de configurações, ele postulou o conceito de “pathway” que sugere uma rota única e determinada de enovelamento. Segundo sua teoria, seria impossível acessar todo o espaço conformacional em busca da menor energia livre em um curto intervalo temporal, o que ficou conhecido como “*Paradoxo de Levinthal*”. Para ilustrar o paradoxo, admita que cada um dos N aminoácidos de uma proteína possa assumir μ orientações. Isto gera μ^N conformações possíveis. Uma proteína, que segundo Creighton,²¹ tem um tamanho médio em torno de 212 a 280 aminoácidos: essa situação para o caso de somente duas orientações por aminoácido geraria um número entre 44.944 e 78.400 conformações possível, o que ilustra de maneira satisfatória o paradoxo.

A idéia de mais de um caminho começou a cogitar com os trabalhos de Harrison²³ motivado pelas novas idéias baseadas na caracterização estatística da energia de relevo (“*Energy of Landscape*”) do enovelamento de proteínas.⁶ A argumentação de Levinthal era baseada na suposição que todas as conformações eram igualmente prováveis no caminho do estado desenovelado até o enovelamento e não que conformações com menor energia são preferenciais. A argumentação foi derrubada completamente com a visão da energia de Landscape como funil.^{35,62}

Sobre esse novo ponto de vista, o entendimento do enovelamento exige

uma visão global do relevo de energia potencial. O conceito de “Funil de energia” sugere que durante a cinética a proteína passa por diversas rotas guiadas pela energia livre do sistema e pela entropia, tendo múltiplos caminhos para alcançar a estrutura nativa. A idéia fundamental do funil consiste na competição entre o estado enovelado e as armadilhas impostas pelo perfil de energia do sistema. A forma afunilada do sistema é necessária para que a proteína vença o paradoxo de Levinthal. O processo é considerada uma organização progressiva por um caminho de conformações.

A dificuldade de elucidar os mecanismos envolvidos de maneira a obter uma lei geral de formação ou padrão energético tem motivado diversos pesquisadores no estudo para obtenção de propriedades fundamentais que descrevam de maneira simples e objetiva o problema. A grande dificuldade consiste em enumerar as conformações no estado desenovelado, o que gera uma diversidade extraordinária de estados imensamente grande.^{5,20}

Nessa busca, o estado enovelado não deveria ser algo trivial levando em conta que uma proteína leva de 1 milissegundo à 1 segundo em média.⁴² Diversos grupos tem usado diferentes modelos para tentar explicar essa alta velocidade através de diferentes modelos e mecanismos; nucleação,⁵⁹ difusão-colisão,³⁰ colapso hidrofóbico,^{26-28,50} modelo do funil.^{35,53,60}

Estudos mais recentes ainda mostram que falhas no processo de enovelamento podem levar a casos letais^{10,38,51} e que os passos finais do enovelamento usualmente envolvem um ajuste das cadeias laterais,⁴² o que obriga a construção de modelos mais específicos que englobem essas características. Entre outros aspectos, é motivante verificar a existência de diferentes seqüências que enovelavam para a mesma estrutura funcional, partindo de seqüências completamente diferentes. Mèlin *et al*^{24,36,39} sugere que a estabilidade termodinâmica, robustez e ocorrência de “*motifs*” pode ser entendido através do conceito de desenhabilidade para uma proteína, que consiste na medida de quantas diferentes seqüências uma mesma conformação pode aceitar. Nesse conceito, estruturas que abrigam um

número muito grande de diferentes seqüências de aminoácidos seriam preferenciais.

As seqüências consideradas boas devem envelar rapidamente mantendo sua estabilidade a pequenas variações do ambiente. É importante manterem-se conformacionalmente inalteradas quando erros na sua construção ocorrem, o que definimos como robustez. Mudanças em 1 ou mais aminoácidos da seqüência não devem ser suficientes para destruir sua funcionalidade ou papel biológico na maioria dos casos. Em alguns casos, isto pode representar melhorias no desenvolvimento de suas tarefas, na acessibilidade ou à qualquer uma de suas características fundamentais. Para este último item consideramos que a proteína evoluiu.

Dentre os vários casos envolvidos; este trabalho é focalizando no problema do envelamento de proteínas, centralizando as discussões e idéias nos aspectos topológicos, atribuindo sugestões para atributos oriundos da evolução.

O presente trabalho forneceu nesse primeiro capítulo uma breve introdução do estudo do envelamento de proteínas e dos objetivos envolvidos no “Problema do envelamento”. O capítulo 2 apresentam as técnicas desenvolvidas, estudadas e aprimoradas. Esse conjunto é a base dos métodos e avaliações que provêm todos os resultados obtidos. Os resultados ainda não publicados são apresentados no capítulo 3 e estão focados no estudo da contribuição topológica para modelagem em rede. Conclusões encerram a primeira parte do trabalho.

Na segunda parte do trabalho, duas publicações são apresentadas com objetivos de avaliar a influência energética em modelos minimalistas em rede e avaliar a robustez do processo de envelamento para a variação dos parâmetros da Hamiltoniana do sistema para o modelo C_α e $C_\alpha C_\beta$. Essa primeira parte deve ser suficiente para o entendimento dos assuntos e técnicas discutidas nas publicações apresentadas na segunda parte do trabalho.

Capítulo 2

Os modelos

2.1 O modelo “protein-like em rede cúbica”

A idéia central para o uso deste modelo consiste no estudo de características topológicas fundamentais em proteínas ao longo da evolução. Neste contexto, um modelo minimalista de rede cúbica, ou seja, aquele que engloba o menor número de variáveis possíveis e com a mais simples descrição, foi escolhido devido à ampla referencia bibliográfica.^{11,13,16,32,35-37,39,45,52}

A proteína é modelada por uma seqüência de 27 monômeros representando aminoácidos que são unidos por ligações covalente. A estrutura biológica ativa ou biologicamente funcional é considerada uma conformação cúbica qualquer, desde que maximamente compacta, em que esta seqüência “alvo” escolhida se acomoda segundo critérios pré-estabelecidos.

A enumeração das 103346 conformações maximamente compactas possíveis num cubo $3 \times 3 \times 3$ (excluso todos os casos de rotação) foi inicialmente apresentada por Shakhnovich & Gutin.⁵² Um princípio fundamental no estudo de proteínas é o da “Mínima frustração”^{6,46,61}. A proteína deve ter o maior número de ligações não-covalentes atrativas em seu estado nativo. Outra característica que deve es-

tar presente é a robustez que consiste na propriedade de alcançar e manter-se em seu estado nativo em situações com pequenas variações no meio como pequenas mudanças de pH ou mutações não fundamentais em sua seqüência. Mutações na seqüência de aminoácidos são essenciais nos estudos evolutivo e de robustez.

A proteína deve ser termicamente estável no equilíbrio possuindo um único estado nativo e ser cineticamente acessível. Para isso as proteínas não devem ser sistemas formados por seqüências aleatórias, no entanto, todas as características desejadas podem ser ajustadas nos modelos “protein-like”^{55,56}

No “protein-like” os contatos não-covalentes entre monômeros do mesmo tipo são ditos favoráveis ou do inglês “like”, sendo os contatos não-covalentes entre monômeros de tipos diferentes desfavoráveis ou frustrados (“unlike”). Uma seqüência otimamente desenhada é aquela que em seu estado nativo possui somente contatos favoráveis, no caso do modelo de rede cúbica isso significa 28 contatos não frustrados formados na estrutura final.

A comparação de diferentes estruturas é realizada através da construção de seqüências otimamente desenhadas, sendo o alvo da modelagem cada uma das conformações. A expectativa era de escrever pelo menos 1 seqüência otimamente desenhada para cada conformação maximamente compacta possível, não considerando homopolímeros. A energia total do sistema é definida pelo número de contatos formados e tipos de energia de interação envolvida.

$$E_{total} = n_f E_f + n_d E_d \quad (2.1)$$

Onde n_f é o número de ligações favoráveis e n_d é o número de ligações menos favoráveis ou desfavoráveis. E_f e E_d são respectivamente as energias para cada tipo de contato, que podem ser definidos aleatoriamente a fim de caracterizar o papel do solvente em uma simulação.

O trabalho de Socci et al.⁵⁵⁻⁵⁷, no qual o modelo presente foi inicialmente baseado define o cenário de enovelamento através do cálculo simplificado da energia média ($E_{media} = E_f + E_d$) e através da heterogeneidade do sistema

$\Delta_{Socci} = (E_d - E_f)$ que determina a distância de energia a ser satisfeita. O sistema com $E_f = -3$ e $E_d = -1$ definem o regime de alta, o que segundo o mesmo raciocínio, consideramos $E_f = -3$ e $E_d = +3$ como de baixa hidrofobicidade. Chahine *et al.*¹² faz uma discussão aprofundada sobre diversos cenários onde o colapso específico e não específico podem ocorrer. Neste estudo, as variáveis que definem o comportamento do mecanismo de enovelamento depende da energia de ligação dos contatos favoráveis e desfavoráveis e dos tipos de monômeros presentes na seqüência. O parâmetro de colapso é obtido novamente através da relação de energia não específica e a dispersão de energia de contato. Por esse método, diferentes seqüências podem ser avaliadas tanto por uma energia fixa, como por um ajuste individual para cada seqüência garantindo um mesmo cenário de enovelamento (controle do procedimento de enovelamento - colapso específico / não específico para diferente seqüências)

A energia média não específica, ou seja, a energia média de interação entre os monômeros, é dada por:

$$\langle E_{ns} \rangle \cong \sum_{i=1}^m [f_i^2 E_f + f_i(1 - f_i) E_d] \quad (2.2)$$

onde f_i é a ocorrência de um tipo de monômero i na cadeia, e m é a quantidade de tipos de monômeros diferentes. A rugosidade do sistema pode ser medida através de

$$\Delta \cong \sum_{i=1}^m [f_i^2 (E_f - \langle E_{ns} \rangle)^2 + f_i(1 - f_i) (E_d - \langle E_{ns} \rangle)^2]. \quad (2.3)$$

A razão entre essas duas grandezas (equações 2.2 e 2.3) fornece o parâmetro de hidrofobicidade empregado

$$\kappa = -\frac{\langle E_{ns} \rangle}{\Delta}. \quad (2.4)$$

O colapso ocorre para valores de κ maiores que 0.5. Seqüências com valores de κ iguais, obrigatoriamente, estão sob o mesmo regime de hidrofobicidade. Mantendo E_f fixo podemos manipular E_d a fim de garantir o ajuste desejado. Para todos os casos, assumimos $E_f = -3$.

O cálculo da energia total do sistema permite verificar o intervalo de energia entre o estado nativo e o próximo estado de mais baixa energia (GAP). Alguns autores sugerem que quanto maior o GAP de energia, mais estável e acessível será o estado enovelado. O teste é realizado inserindo cada seqüência obtida em cada uma das conformações existentes (103346). Quando a mais baixa energia é obtida por duas conformações diferentes (para a mesma seqüência) dizemos que o GAP de energia é igual a zero e que esta possui um estado nativo degenerado, e portanto é descartada. As seqüências selecionadas possuem somente uma configuração de mais baixa energia, sendo não-degeneradas.

O algoritmo para a geração de seqüências otimamente desenhadas é escrito de maneira à sempre maximizar o número de letras possíveis. Isto é obtido uma vez que a seqüência é “montada” utilizando o mapa de contato da estrutura “alvo” que deseja-se estudar. O número máximo de monômeros possíveis para desenhar seqüências sem frustrações energéticas é 5 e foi comprovado pela execução do algoritmo. Isto sugere que, para modelos em rede cúbica, a conformação limita o número de aminoácidos.

Dentre os parâmetros conhecidos para modelos teóricos em rede, consideramos o mais satisfatório para o cálculo de estabilidade termodinâmica e acessibilidade cinética o Z_{score} introduzido por Dima *et al.*¹⁶ como:

$$Z_{score} = \frac{\langle E \rangle - E_{nativo}}{\sqrt{\langle E^2 \rangle - \langle E \rangle^2}} \quad (2.5)$$

que mede a diferença de energia entre o estado nativo (E_{nativo}) e a energia média de todas as outras conformações para uma mesma seqüência dividido pelo valor do desvio padrão.

A ordem de contado relativa (OC) é a medida da distância media entre pares de monômeros vizinhos não covalentes, normalizado pelo comprimento da cadeia

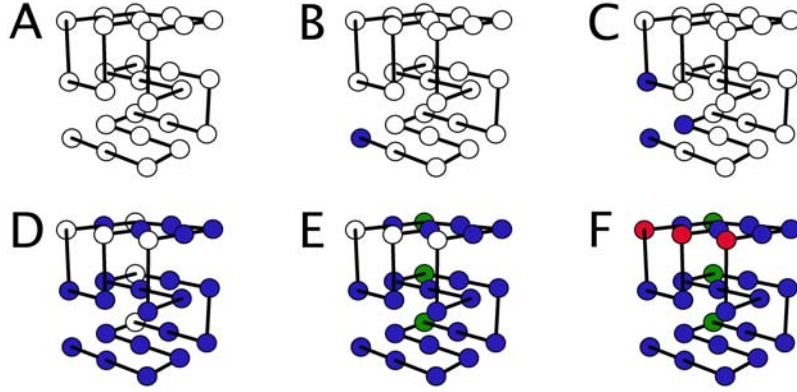


Figura 2.1: *Representação do funcionamento do algoritmo de geração das seqüências aplicado à conformação 17467 escolhida aleatoriamente. O programa escolhe inicialmente um monômero qualquer, neste caso o indicado com a cor azul na figura B, identificando os contatos formados por este na estrutura e colorindo como apresentado em C. O processo é repetido para os novos contatos dos monômeros marcados até que todo o grupo (“cluster”) seja varrido (D). Após a busca, os itens não marcados são selecionados pela atribuição de uma nova cor e o algoritmo inicia-se novamente em busca do novo grupo (E). Em F é possível visualizar a seqüência pronta.*

$$OC = \frac{1}{NL} \sum_{i,j}^N \Delta S_{i,j} \quad (2.6)$$

onde N é o número total de contatos (28), $S_{i,j}$ é o número de monômeros entre os vizinhos não covalentes i e j , e L o número total de monômeros na seqüência.

Os parâmetros estatísticos utilizados (OC e Z_{score}) concordam com os valores experimentais no modelo em questão. Kaya *et al.*³¹ demonstram a correlação entra a taxa de enovelamento/reenovelamento em relação a ordem de contato. Os resultados de Z_{score} foram calculados e verificados como igualmente bons, através de simulações de Monte Carlo e não serão apresentados ou discutidos aqui.

A partir do momento que o algoritmo foi aplicado para cada uma das con-

formações (como ilustrado na figura 2.1), podemos dizer que existe uma seqüência associada para cada uma das conformações. A robustez do sistema é avaliada através de permutações simples e duplas. Permutações simples consistem na troca de posição de 2 monômeros de tipos diferentes ao longo da cadeia, mutações duplas na troca de 2 pares de monômeros de tipos diferentes. Para uma dada seqüência, todas as permutações simples e duplas são realizadas e testadas em todas as conformações. À partir das seqüências não-degeneradas obtidas é obtida a porcentagem de cadeias que mantêm a estrutura alvo (e o Z_{score}). A porcentagem obtida é a maneira direta de quantizar a robustez nesse tipo de modelo e quanto maior o seu valor, mais desejável.

2.2 Simulações via método de Monte Carlo

2.2.1 Cálculo do tempo de enovelamento

O método de Monte Carlo tem sido empregado com sucesso nos mais diversos estudos simulacionais afim de calcular propriedades de sistemas de vários corpos. O método consiste em varrer o espaço conformacional utilizando números aleatórios ou pseudo-aleatórios. Dentre as várias maneiras possíveis, o trabalho faz uso do Critério de Metropolis⁴⁰. A simulação inicia-se pelo crescimento de uma conformação aleatória numa rede cúbica à partir de uma seqüência pré-definida. Os movimentos realizados são baseados na física de polímeros (Figura 2.2) e aceitos segundo a probabilidade de Boltzmann definida por P_{Btz} . A cada movimento aceito ou não, este é contabilizado como um passo de Monte Carlo (MCs).

$$P_{Btz} = e^{\frac{-(E_{nova} - E_{anterior})}{K_b T}} \quad (2.7)$$

Após um movimento realizado, esse é aceito se a nova energia for mais baixa que a anterior, caso contrário, a diferença de energia do sistema é calculada e comparada com um número aleatório entre 0 e 1. K_b é definido como a constante de Boltzmann. Se o valor da probabilidade de Boltzmann for maior que o número gerado, a nova conformação é aceita.

O critério de Metrópolis possibilita a cadeia de escapar de mínimos locais, além de, em simulações no estado de equilíbrio manter a vibração natural de uma proteína para uma dada temperatura.

O tempo de enovelamento é obtido através das médias aritméticas dos passos de Monte Carlo a partir de 100 tomadas da estrutura inicial. As corridas são realizadas para temperaturas suficientes afim de construir o perfil dos tempos de enovelamento. Considera-se a temperatura de enovelamento T_f a que os estados enovelado e desenovelado são igualmente populados.

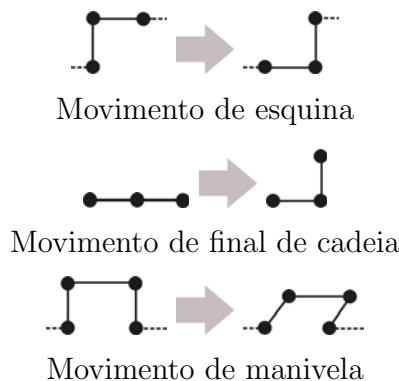


Figura 2.2: *Tipos de movimentos para as simulações em modelo de rede cúbica. O movimento de esquina e de manivela exigem que o monômero sorteado não seja ou o primeiro ou o último da cadeia ao contrário do movimento de final, que só pode ser realizado em tais condições. Entre eles não existe nenhum privilégio ou peso ao longo da simulação.*

2.2.2 Estudo evolutivo

O processo evolutivo é realizado através de simulações de Monte Carlo com uma pequena modificação no critério de Metrópolis. O ponto inicial é uma seqüência de 2 letras não-degenerada, assim, cadeias aleatoriamente são geradas até que uma satisfaça essa condição. Permutações simples ou mutações são realizadas. Mutações devem ser entendidas como a troca de 1 monômero outro de tipo diferente que vai desde 1 a 5 (posteriormente representado por cores ou letras). Essas mutações são a essência evolutiva do algoritmo. A nova seqüência é testada em todas as conformações possíveis. Caso encontre uma estrutura de mais baixa energia única, a seqüência é redesenhada de maneira a ficar sem frustrações (utilizando o algoritmo de construção das seqüências apresentado anteriormente). A aceitação é baseada no critério de Metrópolis, onde a energia é substituída por $-Z_{score}$ de maneira que a simulação de não equilíbrio caminhe para conformações que aumentam a estabilidade. A simulação é finalizada quando um determinado Z_{score} é alcançado por um número definido de vezes. Em todas as simulações

a seqüência/conformação final deve ter Z_{score} igual ou superior a 8.5. Diversas corridas são realizadas de maneira à obter de diversos pontos de partida, diversos estados finais que caracterizam esse conjunto de estruturas como um funil (com 1 única estrutura final) ou balde (diversas).

Com várias simulações, é possível verificar se existe uma mesma estrutura em duas ou mais simulações com finais diferentes, definindo assim um conjunto ou grupo, o que vai caracterizar mais tarde como um balde ou funil.

Ainda é possível, à partir de uma única corrida, medir o número médio de conexões que diferentes estruturas realizam ao longo do caminho.

2.2.3 Termodinâmica do modelo de rede

As simulações para obtenção de propriedades termodinâmicas seguem os mesmos moldes da para o cálculo do tempo de enovelamento, exceto por pequenas diferenças. A corrida é realizada somente em uma temperatura próxima à de enovelamento ou na temperatura crítica que são próximas, em geral. O valor dessa temperatura é obtido através do pico de calor específico e representa o valor onde a proteína está sujeita ao maior número de transições do estado enovelado para o desenovelado e vice-versa. Após ser gerada uma conformação espacial aleatória inicial, o sistema passa por um período de termalização muito superior ao tempo de enovelamento médio da proteína. Todos os resultados obtidos durante a termalização são descartados. A simulação é realizada até que o número de passos de Monte Carlo seja pelo menos 4 ordens de grandeza maior que o tempo médio de enovelamento, ou que o número de transições sejam suficientes para uma estatística coerente. Os valores obtidos são aplicados ao método do histograma simples.

2.2.4 Método do histograma simples

O método do histograma simples¹⁷ é uma técnica utilizada para calcular propriedades termodinâmicas para pequenos sistemas. Neste trabalho ela é aplicada para estudos na rede cúbica, uma vez que todo o espaço conformacional pode ser varrido, a partir de uma única simulação, na temperatura crítica do sistema. Nesta técnica, a probabilidade do sistema é aproximada ao histograma normalizado do sistema.

O número total de contatos nativos formados (ou fração de contatos nativos formados em alguns casos) Q foi escolhido como coordenada de reação devido à inúmeros trabalhos mostrando que para proteínas pequenas com dois estados definidos, a transição tem probabilidade igual na análise de retornar a forma desenovelada ou seguir para o estado nativo^{9,15,43}. Para os estudos em rede, o número de contatos formados quaisquer Z foi incluído afim de verificar colapsos hidrofóbicos.

Dada a função probabilidade para uma dada temperatura T e T' , respectivamente temos:

$$P_{\beta}(E, Z, Q) = \frac{\Omega(E, Z, Q)e^{-\beta E}}{\sum_E \Omega(E, Z, Q)e^{-\beta E}} \quad (2.8)$$

$$P_{\beta'}(E, Z, Q) = \frac{\Omega(E, Z, Q)e^{-\beta' E}}{\sum_E \Omega(E, Z, Q)e^{-\beta' E}} \quad (2.9)$$

onde $P_X(E, Z, Q)$ definem a probabilidade para os parâmetros de ordem E, Z e Q , β é o inverso da temperatura e x representa a condição de temperatura desejada. Nessas equações, a degenerescência $\Omega(E, Z, Q)$ do sistema é um fator geométrico, ou seja, não depende da temperatura (a ocorrência é a mesma com uma elevação nos níveis de energia). Considerando que a função de partição é dada por $Z_{\beta} = \sum \Omega(E, Z, Q)e^{-\beta E}$ genérico, podemos dividir a equação 2.8 pela 2.9 obtendo uma expressão que estima uma probabilidade desconhecida à partir

de uma já conhecida.

$$P_{\beta'} = \frac{P_{\beta}(E e^{-(\beta'-\beta)E})}{\frac{Z_{\beta'}}{Z_{\beta}}}$$

$$P_{\beta'(E,Z,Q)} = \frac{P_{\beta}(E, Z, Q) e^{-(\beta'-\beta)E}}{\sum P_{\beta}(E, Z, Q) e^{-(\beta'-\beta)E}} \quad (2.10)$$

Com as probabilidades definidas, grandezas termodinâmicas são facilmente obtidas:

$$C_v = k\beta^2(\langle E^2 \rangle - \langle E \rangle^2)$$

$$\langle E^2 \rangle_{\beta'} = \sum_E P_{\beta'}(E) E^2$$

$$\langle E \rangle_{\beta'} = \sum_E P_{\beta'}(E) E$$

Como a proteína deve ter somente um estado nativo único e de mais baixa energia, ajustamos o valor de $\Omega(-84, 28, 28) = 1$. A probabilidade de encontrar a proteína no estado nativo é dada por $P_{nativo} = \frac{e^{-E_{nat}/t}}{Z}$.

2.3 Modelos baseados em estrutura

2.3.1 Construção do Modelo

Apesar da possibilidade de um estudo na rede para esses modelos, foi adotado o espaço contínuo de maneira a aumentar a liberdade conformacional e se aproximar o máximo possível de uma proteína real. Modelos baseados em estrutura^{1, 7, 8, 22, 41, 48, 49} apresentam bom acordo com resultados experimentais em relação ao estado de transição. Em sua confecção mais simples, a proteína é representada por esferas localizadas nas posições dos carbonos alfa (C_α) da estrutura definida experimentalmente via cristalografia ou ressonância magnética (NMR). Esse modelo é também conhecido como $C_\alpha - G\bar{o}$. Um aperfeiçoamento desse modelo consiste em adicionar um carbono beta (C_β) na posição do centro de massa da cadeia lateral. A representação mais realística consiste no uso de todos os átomos, porém esse caso não será abordado neste trabalho.

A energia total do sistema é definida tomando como referência as conformações geradas ao longo da simulação Γ em relação ao estado nativo Γ_0 dado por:

$$\begin{aligned}
 U(\Gamma, \Gamma_0) = & \sum_{\text{covalentes}} \frac{1}{2} \epsilon_r (r - r_0)^2 + \sum_{\text{angulares}} \frac{1}{2} \epsilon_\theta (\theta - \theta_0)^2 \\
 & + \sum_{\text{diedral}} \{ \epsilon_\phi [1 - \cos(\phi - \phi_0)] + [\frac{1}{2} \epsilon_\phi [1 - \cos(3(\phi - \phi_0))]] \} \\
 & + \sum_{i < j - 3} \{ \epsilon_{LJ}(i, j) [5 (\frac{\sigma_{ij}}{r_{ij}})^{12} - 6 (\frac{\sigma_{ij}}{r_{ij}})^{10}] \} + \sum_{ij} \epsilon_{LJrep}(i, j) (\frac{\sigma_{ij}}{r_{ij}})^{12} \quad (2.11)
 \end{aligned}$$

As variáveis com índice $_0$ indicam os valores na estrutura nativa ou de referencia. As constantes $\epsilon_r, \epsilon_\theta, \epsilon_\phi, \epsilon_{LJ}$ e ϵ_{LJrep} são respectivamente os pré-fatores energéticos de ligação covalente, angulares, diedrais e de interações de Lennard-Jones e repulsão e podem ser ajustados com certa flexibilidade⁴⁴. Os valores

utilizados são apresentados na tabela 2.1. Ainda na equação, r é a distância entre dois átomos ligados covalentemente, θ é o ângulo entre duas ligações covalentes adjacentes, ϕ é o ângulo formado entre 2 planos formados por 4 átomos conectados.

Diferenciamos diferentes diedros que dependem do tipo de átomos envolvidos e isso influencia nos valores das constantes, como pode ser verificado na tabela 2.1. As torsões impróprias inseridas e representadas na figura 2.3 são necessárias para que o C_β não gire em torno do C_α diretamente ligado a ele. Elas possuem valor máximo na conformação *cis* e ajudam a manter a quiralidade da molécula. No termo de Leonard-Jones, $r_{i,j}$ é a distância entre dois átomos quaisquer e $\sigma_{i,j}$ a distância entre os contatos ligados nativamente, caso não sejam, é assumido como 4 Å.

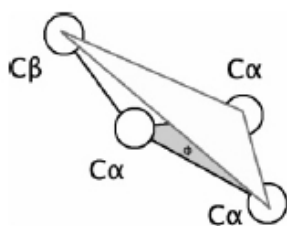


Figura 2.3: *Representação de uma torsão imprópria. Os planos formados pelos 4 átomos conectados formam um ângulo que limita a mobilidade do carbono β em torno do carbono central*

Os contatos não ligados covalentemente são definidos pelo CSU^{54} (*Contacts of Structural Units*). Consideramos somente os contatos entre carbonos do mesmo tipo, sendo sugerido ao leitor o tratamento para demais casos⁴⁴, no qual discute os limites para que não haja a quebra do mecanismo correto de enovelamento. Fazer com que o sistema faça interações entre todos os pares de átomos possível não acrescenta vantagens no modelo e exige uma modificação na maioria dos parâmetros energéticos.

Ainda quanto as constantes, é importante notar que para uma com-

ϵ_r		$100 K_B T$
ϵ_θ		$20 K_B T$
ϵ_ϕ	$C_\alpha C_\alpha C_\alpha C_\alpha$	$0.8 K_B T$
	$C_\alpha C_\alpha C_\alpha C_\beta$ ou $C_\beta C_\alpha C_\alpha C_\alpha$	$0.2 K_B T$
	$C_\beta C_\alpha C_\alpha C_\beta$	$0.1 K_B T$
	Torsões impróprias	$1.0 K_B T$
ϵ_{LJ}		$1.0 K_B T$
$\epsilon_{LJ,ep}$		$1.0 K_B T$

Tabela 2.1: Valores para as constantes da Hamiltoniana apresentada na equação 3.8.

paração entre barreiras de potencial para diferentes proteínas é necessário forçar que a energia no estado nativo seja o mesmo em todos os casos. Ao não tomar esse cuidado, somente os mecanismos de enovelamento envolvidos podem ser comparados.

2.3.2 Simulações via Dinâmica Molecular

Simulações com Dinâmica Molecular (MD) tem grande proximidade com a situação real uma vez que descrevem a evolução temporal tão bem quanto possível. A configuração inicial é comumente obtida através de uma estrutura determinada experimentalmente. Em MD as equações clássicas de movimento são resolvidas para todos os átomos presentes no sistema, sendo estes tratados explicitamente.

A equação do movimento 2.12 é resolvida afim de obter posições atômicas e velocidades como uma função do tempo.

$$m_i \frac{d^2 \vec{r}}{dt^2} = -\nabla_i [U(\vec{r}_1, \vec{r}_2, \dots, \vec{r}_i)], i = 1, N \quad (2.12)$$

Na equação apresentada m_i representa a massa de cada uma das partículas i enquanto \vec{r}_i a sua respectiva posição espacial e U a energia potencial, que de-

pende da posição das N partículas do sistema. A energia potencial inclui termos de ligações covalentes, angulares, torsionais (diedrais), impróprios (para manterem tetraedros ou geometrias planares) e termos não ligados covalentemente, como Lennard-Jones e possíveis contribuições eletrostáticas.

O sistema é equilibrado pela integração das equações do movimento enquanto valores de temperatura são ajustados. Esse controle é dirigido ao intervalo de interesse e reajustados pela distribuição de Maxwell no aumento da temperatura ou pelo escalonamento de todas as velocidades. A temperatura do sistema $T(t)$ é determinada em termos da energia cinética média

$$T(t) = \frac{1}{(3N - n)K_b} \sum_{i=1}^N m_i |\vec{v}_i|^2 \quad (2.13)$$

onde, $(3N - n)$ é o número total de graus de liberdade do sistema, v_i é a velocidade da partícula i no tempo t . O escalonamento de velocidades é dado por um fator de $\sqrt{T'/T(t)}$ que resultará na energia cinética correspondente à temperatura T' .

A cada integração numérica a força ($F = -\nabla U$) é calculada e os átomos atualizados. Sistemas grandes tornam esse tipo de simulação inviável devido ao alto tempo computacional envolvido no cálculo de todos os pares de interações para o cálculo da Força requeridos pela MD.

2.3.3 Método dos múltiplos histogramas

O método dos múltiplos histogramas^{2,34} ou WHAM (*The Weighted Histogram Analysis Method*) é um aperfeiçoamento do método do histograma simples apresentado anteriormente. Para sistemas maiores, infelizmente 1 simulação longa na temperatura crítica não é suficiente para representar todo *Ensemble* de estados energéticos/conformacionais e desta maneira é necessário o uso de vários histogramas atribuindo então diferentes pesos para cada um deles. Partindo da

equação 2.10

$$P_{\beta'}(E, Q) = \frac{P_{\beta}(E, Q)e^{-(\beta'-\beta)E}}{\sum_E P_{\beta}(E, Q)e^{(\beta'-\beta)E}}$$

$$P_{\beta}(E, Q) = \frac{\Omega(E, Q)e^{-\beta E}}{\sum_E \Omega(E, Q)e^{-\beta E}} \quad (2.14)$$

onde, novamente, P_{β} e $P_{\beta'}$ são as probabilidades e $\Omega(E, Q)$ a degenerescência. Aqui não podemos aproximar o histograma à própria probabilidade como no método do histograma simples, portanto, ele é definido por $N_{\eta}(E, Q)$ ou $n_{\eta}(E, Q)$ quando normalizado. A partir da geração vários histogramas a média é obtida através de:

$$\Omega(E, Q) = \bar{N}_{\eta}(E, Q)e^{-\beta_{\eta}(f\beta_{\eta}-E)} \quad (2.15)$$

A energia livre $f = -\frac{1}{\beta}\ln Z$, sendo Z a função de partição. Tendo agora vários histogramas deve ser atribuído pesos para cada R corridas, assim

$$\Omega(E, Q) = \sum_{n=1}^R p_{\eta}(E, Q)N_{\eta}(E, Q)e^{-\beta_{\eta}(f\beta_{\eta}-E)} \quad (2.16)$$

$$\sum_{n=1}^R p_{\eta}(E, Q) = 1$$

O conjunto de pesos $p_{\eta}(E, Q)$ que minimiza os erros no cálculo de $\Omega(E, Q)$ foi apresentado por Bennett⁴ e é dado por,

$$p_{\eta}(E, Q) = \frac{n_{\eta}(E, Q)e^{\beta_{\eta}(f\beta_{\eta}-E)}}{\sum_n^R n_{\eta}(E, Q)e^{\beta_{\eta}(f\beta_{\eta}-E)}} \quad (2.17)$$

e substituindo 2.16 em 2.17, temos:

$$P_{\beta}(E, Q) = \frac{\sum_{n=1}^R N_{\eta}(E, Q)e^{-\beta_{\eta}E}}{\sum_{n=1}^R n_{\eta}(E, Q)p_{\eta}(E, Q)^{\beta_n(f\beta_n-E)}} \quad (2.18)$$

$$e^{-f_n} = \sum_E P_{\beta}(E, Q) \quad (2.19)$$

O método é auto-consistente fazendo-se necessário atribuir valores iniciais para a energia livre (f) com o intuito de encontrar p , fazendo então o caminho inverso até a convergência.

2.3.4 Outras avaliações

As análises complementares são obtidas através do *Ensemble* do estado de transição (*TSE*). Essa região é definida como o intervalo delimitado pelos valores de Q (fração de contatos nativos) que cortam o gráfico de energia livre por uma reta à $1K_B T$ abaixo do ponto máximo. A avaliação dos valores de Φ ⁵³ é essencial na identificação de aminoácidos essenciais no processo de enovelamento. Experimentalmente são testados todos os pontos de mutação e avaliados as contribuições de cada resíduo mutado.¹⁸

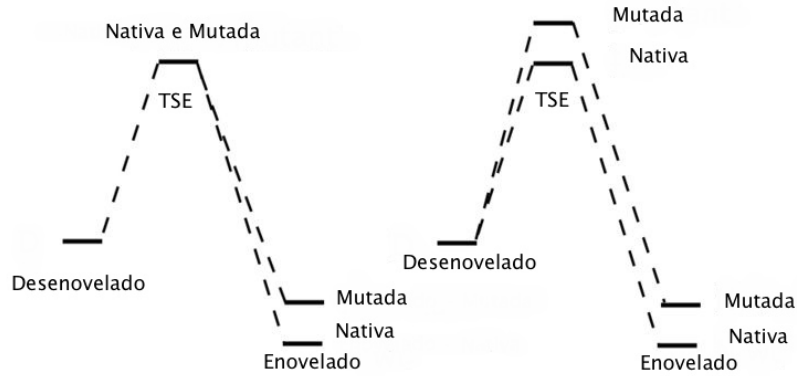


Figura 2.4: O diagrama ilustra a energia dos estados desenovelado, de transição e enovelado para os casos de valores de Φ igual a 0 (à esquerda) e 1 (à direita).

$$\Phi = \frac{(\Delta G_{nativa}^{TSE-Desenovelado} - \Delta G_{mutada}^{TSE-Desenovelado})}{(\Delta G_{nativa}^{Nativo-Desenovelado} - \Delta G_{mutada}^{Nativo-Desenovelado})} \quad (2.20)$$

onde $\Delta G \simeq \ln \langle e^{\Delta E/k_b T} \rangle$,

$$\Phi = \frac{\Delta\Delta G^{TSE-Desenovelado}}{\Delta\Delta G^{Enovelado-Desenovelado}} \quad (2.21)$$

Em muitos casos, as mutações em regiões específicas na proteína impedem completamente o alcance ao estado funcional. Uma outra forma interessante de calcular os valores de Φ em simulações consiste na comparação das contribuições de cada resíduo mutado e seu estado referencial nativo através da probabilidade de formação de contatos

$$\Phi_i = \frac{P_{TSE}^i - P_{desenovelado}^i}{P_{enovelado}^i - P_{desenovelado}^i} \quad (2.22)$$

sendo P_{estado}^i a probabilidade de formar um contato i no estado proposto. Valores de $\Phi = 0$ indicam que o contato nunca é formado no *TSE* e portanto não são importantes para a proteína vencer a barreira de energia livre. Ao contrário, $\Phi = 1$ indicam que esse contato é fundamental para que a proteína alcance o estado nativo.

De maneira similar ao limite que define o intervalo do *TSE*, informações sobre o mecanismo de enovelamento e a ordem dos eventos pode ser verificada. A análise das rotas são obtidas através de janelas de intervalo de Q^{14} onde é possível observar a ordem de formação dos contatos através da probabilidade de contatos formados ao longo da coordenada Q . Um exemplo de como o procedimento funciona está apresentado na figura 2.5, que mostra a evolução temporal em relação à coordenada de reação para a proteína CI2 (*Chymotrypsin Inhibitor 2*), uma proteína com rota única e bem definida.

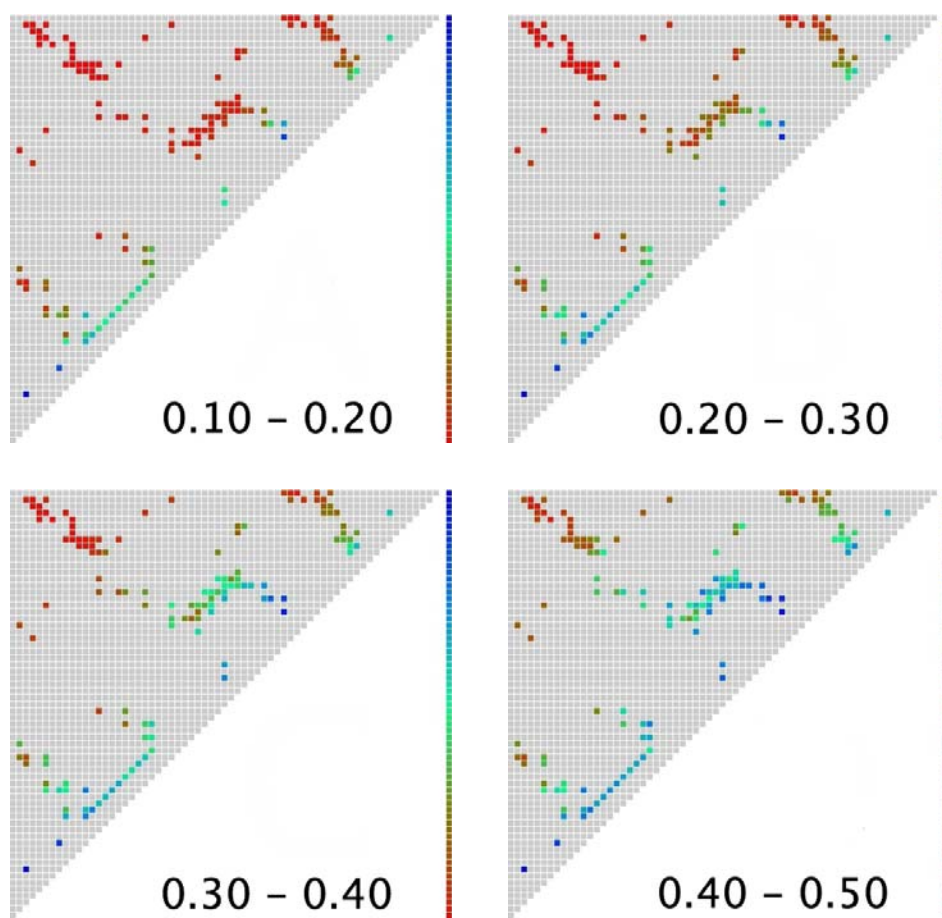


Figura 2.5: *Diferentes probabilidades de formação de contatos nativos ao longo da coordenada de reação Q . Cada figura mostra um intervalo de 0.10 em Q começando em 0.10. É possível acompanhar a formação das diferentes regiões da proteína. Essa figura pode ser simplificada utilizando somente a representação no intervalo do TSE e indicando a ordem dos eventos*

Capítulo 3

Resultados e discussões

A fim de estudar as características topológicas presentes em proteínas, foram geradas 103346 conformações maximamente compactas possíveis em uma rede cúbica $3 \times 3 \times 3$ onde podemos abrigar 27 monômeros que correspondem ao tamanho de nossa cadeia. Iniciamos a busca de cada seqüência otimamente desenhada para cada uma das estruturas através do algoritmo proposto no Capítulo 2. Foi escrita uma seqüência não-frustrada para cada uma das estruturas possíveis. Nesta abordagem, tanto a seqüência quanto a sua estrutura equivalente recebem a mesma numeração, uma vez que cada seqüência esta associada a estrutura na qual esta foi desenhada.

seq	letras	% degenerada	n degeneradas
5	5	0	5
702	4	66,94	482
12833	3	23,22	2980
50654	2	2,83	1434
39152	1	100	39152

Tabela 3.1: *Resultados obtidos pelo algoritmo de criação de seqüências otimamente desenhadas. Observando a tabela, é possível notar que cadeias de 5 letras são suficientemente específicas para não possuírem seqüências não degeneradas. O maior número de seqüências não degeneradas são compostas por 3 letras.*

O valor total de seqüências (4901) 4,75% que podem ser utilizadas no estudo concorda com o trabalho anterior de Helling *et. al.*³⁷ (e concorda também com o conceito de desenhabilidade^{24,37,39} proposto pelo mesmo). Nosso trabalho apresenta um modelo aprimorado em relação ao modelo HP^{11,19} proposto por Helling, mas que pode facilmente ser reduzido ao mesmo. No HP os aminoácidos são definidos como hidrofóbicos ou polares, o que seria similar a uma modelagem com 2 letras, bastando um conjunto de simples alterações. Estruturas com um maior número de letras podem ter suas seqüências “alvo” reescritas com um número menor de letras e ainda manterem-se otimamente desenhadas, concordando com os resultados de Li *et al.*³⁶ Temos então que para seqüências de cinco letras podemos reescrever outras novas de quatro letras sem frustrações; para 4 letras novas de três letras e assim por diante. No entanto, é importante notar que essa redução no número de letras podem resultar em cadeias fisicamente iguais. Por exemplo, a seqüência hipotética 12334 é fisicamente idêntica à 13224.

Restringimos ao estudo, de agora em diante, somente aos casos não degenerados. As seqüências selecionadas possuem o primeiro passo para serem consideradas no estudo, possuindo indícios de características desejáveis para proteínas, porém, nem todas ainda possuem. Esse dado confirma a baixa probabilidade de seqüências aleatórias apresentarem características de proteínas.

As relações entre o Z_{Score} e a Ordem de Contato relativa são apresentadas na figura 3.1, onde podemos identificar quatro regiões com comportamentos distintos, sendo “D” a região onde, segundo a literatura estão as seqüências/estruturas que enovelam com tempo hábil e são estáveis. As regiões “A” e “B” apresentam características indesejáveis e dependem de um tempo muito alto para alcançarem o estado nativo. Quando alcançam possuem pouca estabilidade, especialmente em “A”, onde por causa da alta ordem de contato relativa, a maioria das interações são de longo alcance, o que diminui a estabilidade. A região “C” deve ser acessível, porém, sofre dos mesmos problemas comentados anteriormente impostos pela ordem de contato alta.

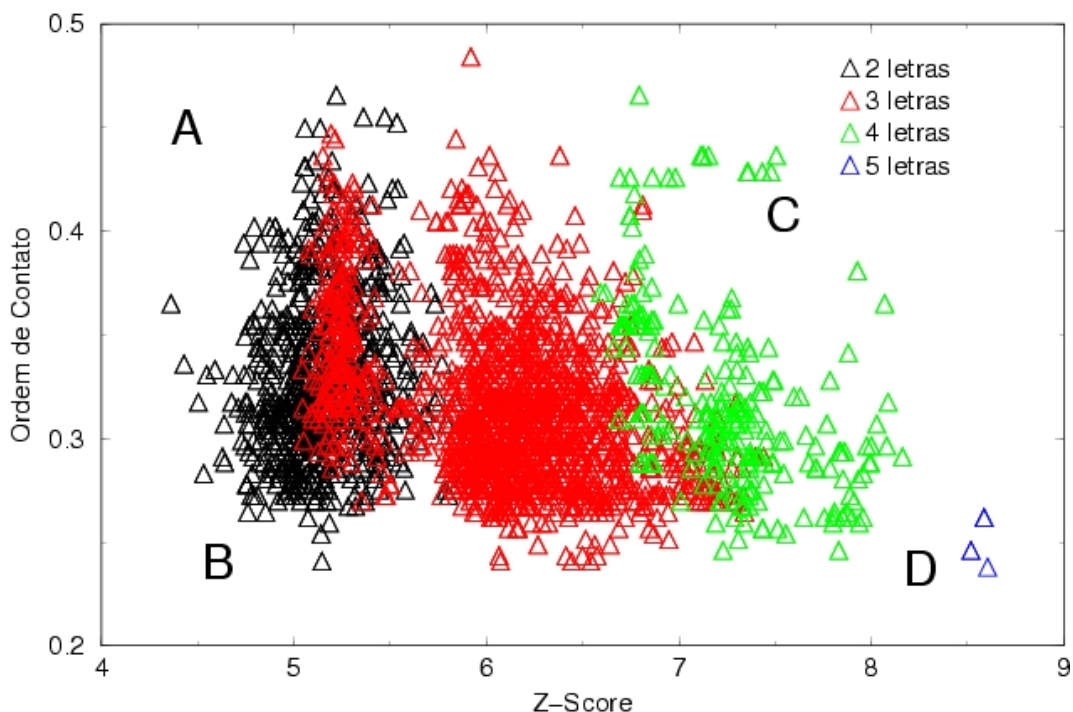


Figura 3.1: *Relação entre ordem de contato relativa e Z_{score} para as cadeias otimamente desenhadas geradas. A região “A” e “B” implicam em condições de difíceis acessibilidade e estabilidade. Em “C” o enovelamento ocorre, porém, os altos valores de OC podem ocasionar problemas em relação à estabilidade. Na região “D” encontramos as condições mais desejáveis; enovelamento rápido e estabilidade.*

Para os testes robustez foram feitas mutações simples gerando novas seqüências, de acordo com o processo descrito anteriormente. Provindas de cada uma das quatro regiões apresentadas no gráfico, avaliamos a capacidade das seqüências manterem-se com mais baixa energia na sua estrutura alvo de desenho.

Os resultados comprovam claramente que as regiões “A” e “B” possuem baixa robustez. A tabela 3.2 apresenta a geração de seqüências por permutações simples. Essa robustez aumenta com o número de letras se encaminhando para o caso mais desejável, onde todas elas são não degeneradas e permanecem na es-

seqüência	letras	novas geradas	não degene	mesma estrutura	%	Z_{score}	OC	Região
70731	2	163	17	2	11,76	5,05	0,41	A
12128	2	172	37	22	59,45	5,06	0,45	A
1739	2	127	12	1	8,33	4,77	0,26	B
56328	2	177	39	21	53,85	5,38	0,29	B
66055	4	216	147	140	95,23	7,48	0,43	C
41818	4	216	135	123	91,11	7,50	0,44	C
46646	5	280	257	255	99,22	8,52	0,25	D
60265	5	280	265	265	100	8,60	0,24	D

Tabela 3.2: *Resumo dos testes de robustez a mutações simples para estruturas pertencentes a cada uma das quatro regiões propostas. A primeira coluna representa a numeração assumida para o estudo da seqüência-conformação, na segunda o número de letras máximo permitida para a situação otimamente desenhada. A coluna seguinte apresenta o número de novas cadeias geradas através de todas as permutações simples possíveis incluindo a original. A quarta coluna mostra a quantidade de cadeias não degeneradas e em seguida, quantas delas tem a menor energia na estrutura alvo para que foram desenhadas. As três últimas colunas apresentam os valores de Z_{score} , Ordem de Contato relativa e a disposição no gráfico.*

trutura alvo. Na última linha da tabela a robustez apresentada pela conformação 60265 é alta a ponto de não aceitar nenhum caso não degenerado em outra estrutura, algo desejável biologicamente, uma vez que sugere o cumprimento de funções vitais se associarmos a atividade como ligada à estrutura. Ao tomar o grupo total não degenerado é nítido que o número de seqüências disponíveis para simulações de propriedades de proteínas é ainda menor que os 4.75% citado anteriormente.

A avaliação entre o número de letras, acessibilidade, estabilidade e topologia foi confrontado tomando a melhor conformação (60265) e comparando as seqüências otimamente desenhadas de mais alto Z_{score} para 5,4 e 3 letras em dois cenários diferentes; colapso e intermediário ao colapso. Tomando com 5 letras A

B C D B A E B D D B E A B D C B A A B C D B A E B D , com 4 A B C D B A C B D D D B C A B D C B A A B C D B A C B D e com 3 A B C B B A A B B B B A A B B C B A A B C B B A A B B foi comparado o perfil de envelhecimento. As modificações foram realçadas em negrito.

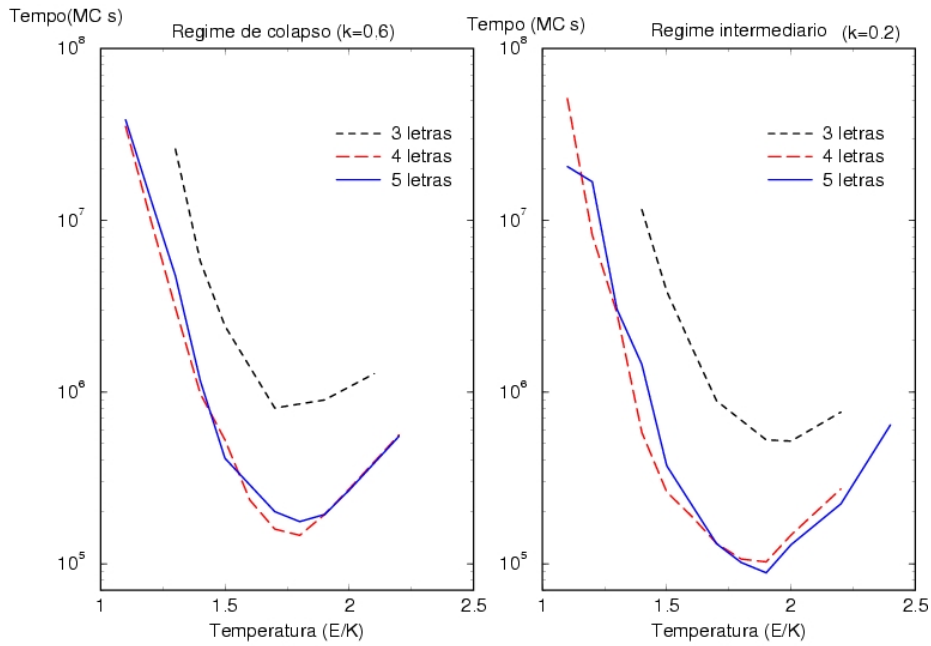


Figura 3.2: Perfil de envelhecimento de seqüências com 5, 4 e 3 letras sem frustrações para uma mesma conformação sob os regimes de colapso e intermediário. O comportamento cinético é praticamente idêntico para as seqüências de quatro e cinco letras. Para a de três letras escrita sob a mesma estrutura, o tempo de envelhecimento é relativamente maior.

Os resultados apresentados na figura 3.2 sugerem a existência de um número ótimo de aminoácidos para uma dada conformação, uma vez que as seqüências de quatro letras possuem praticamente o mesmo comportamento cinético.

Como caso adicional, apenas as cadeias de 5 e 4 letras foram comparadas no regime de $E_f = -3$ e $E_d = -1$ e são apresentados na figura 3.3. Este número “ótimo” estaria ligado à limitações do espaço conformacional imposto na modelagem do sistema. Provavelmente, em redes cúbicas maiores esse número

continuará sendo 4 ou 5 limitado pela necessidade de escrever seqüências sem frustrações. Como exercício mental, considere uma conformação qualquer onde o primeiro ou último monômero está situado no centro do cubo. Devido à ligação covalente, esse monômero pode fazer no máximo 5 contatos, o que limita-se todas as faces, exceto a em contato pelo monômero seguinte (ou anterior). Entretanto, o modelo é muito simples para qualquer outra especulação.

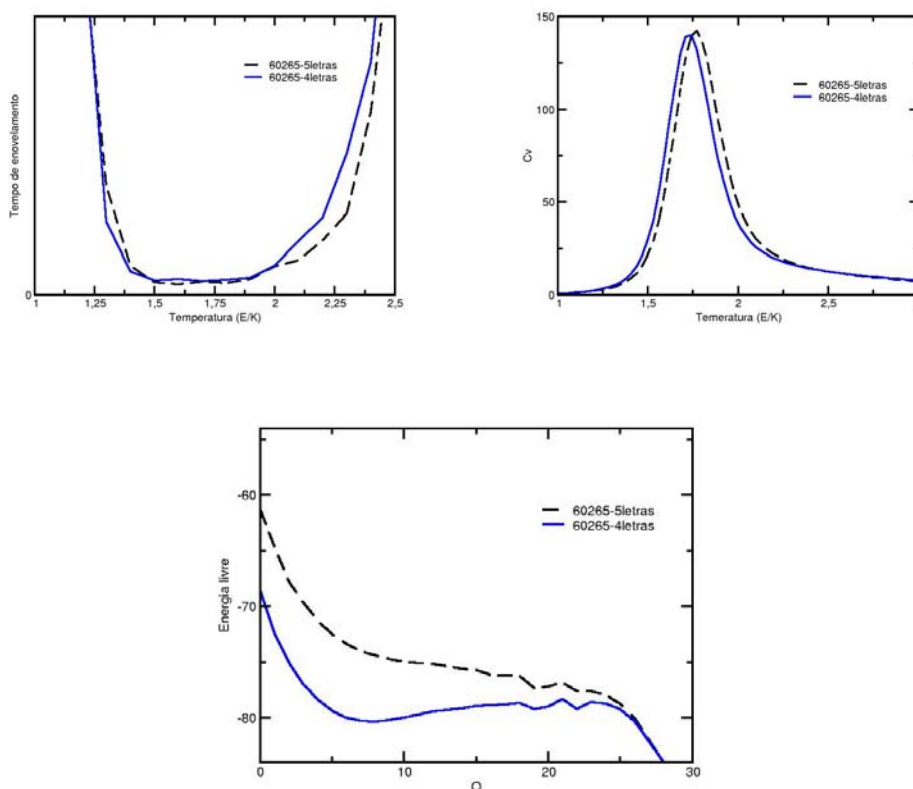


Figura 3.3: Perfil de enovelamento, calor específico e energia livre para seqüências com 5 e 4 letras testadas anteriormente. O gráfico de calor específico mostra que a T_c não apresenta diferença significativa. O gráfico de energia livre mostra uma diferença na barreira de potencial onde podemos considerar a de 5 letras como praticamente zero enquanto no caso de 4 letras temos uma barreira amena

O gráfico de energia livre apresentado na figura 3.3 merece uma com-

paração com os trabalhos de Kubelka *et al*³³ sobre proteínas com enovelamento ultra-rápido. Em seu trabalho, a maioria das proteínas apresentadas possuem barreira de potencial baixa ou próximas de zero. Por se tratar de pequenas proteínas, provavelmente estas alcançaram um alto grau de otimização, o que em nossa teoria, equivalente a seqüência 60265 na situação de 5 letras.

Na abordagem, classificamos a região “D”, apresentada na figura 3.1, como o melhor comportamento para o estudo de proteínas. A melhor seqüência (desenhada sobre a melhor estrutura: 60265), quando sujeita a uma permutação simples, gera novas 280 seqüências, onde 256 apresentam estado nativo não degenerado. Todas permanecem na conformação 60265, mostrando que quando testadas em cada uma das 103346 conformações possíveis, não existe nenhuma outra que ela se acomode tão bem como a sua estrutura alvo (60265). Essa robustez é reduzida em direção das regiões “C”, “B” e “A”, aumentando o número de degeneradas e diminuindo o número que seqüências que se alocam melhor na estrutura em que foi desenhada.

Partindo do caminho inverso percebemos que, para uma seqüência “ruim” (região “A”), pequenas perturbações levam a seqüência a outras conformações, fora de sua estrutura alvo. Tomamos como exemplo a seqüência 1739 apresentada na tabela 3.3, onde as seqüências mutadas não se “acomodam” na estrutura.

Propomos investigar a existência de uma espécie de *Funil de Estruturas*. O funil não apresentaria uma visão do processo de enovelamento de uma proteína; e sim um conceito para o estudo de evolução de proteínas, justificando a robustez, estabilidade e enovelabilidade das proteínas. Neste processo, pequenas mutações ocorridas ao longo da evolução “forçaram” trocas de estruturas, ou melhor, uma alocação em conformações topologicamente melhor adaptadas. Estruturas pouco robustas por não poderem abrigar suas seqüências alvo levemente modificadas seriam naturalmente extintas em uma espécie de seleção natural. Tal processo seria uma possível explicação para a ocorrência de grupos muito bem definidos de proteínas, como, por exemplo, as hemoglobinas e alguns “motifs”.

seqüência	Z_{score}	Gap	Conformação
nativa	4.76059	8	1739
mutada	4.76745	8	91324
mutada	4.77393	8	1760
mutada	4.89017	8	95046
mutada	4.91654	8	1026
mutada	4.92565	8	1126
mutada	4.94938	8	1143
mutada	4.95764	8	95054

Tabela 3.3: *Pequeno trecho da análise das seqüências geradas à partir de permutações simples sobre a seqüência otimamente desenhada 1739. Nota-se que nenhuma das seqüências geradas apresentadas se ajustam bem à estrutura alvo e que ainda existe um aumento nos valores de Z_{score} quando uma troca de conformação ocorre.*

No caso apresentado, somente quando otimamente desenhada, a cadeia 1739 se ajusta bem a sua estrutura respectiva. Quando mutada, além de assumir sua mais baixa energia em outra conformação, ainda existe um aumento nos valores de Z_{score} , sugerindo que, em qualquer das outras conformações, seu tempo de envelamento e estabilidade aumentaram. Consideramos que a troca de estrutura representa uma evolução no âmbito de conformação. Fazendo um paralelo com as teorias propostas no conceito de “Funil de energia”, no caminho evolutivo protéico deve haver um caminho de estruturas até a condição mais estável. Assim, o “funil de estruturas” se diferencia por não avaliar uma única seqüência na rota de envelamento. A busca acontece no espaço de seqüência-estrutura através de sucessivos melhoramentos e reotimização.

As simulações de evolução são realizadas segundo a descrição apresentado no capítulo 2. A temperatura foi ajustada em 2 unidade de E/K_B de maneira a manter a taxa de aceitação em torno de 0.5. Simulações em outras temperaturas foram realizadas, mas não apresentaram diferenças no resultado final, em geral, basicamente o conjunto interligado é acessado, porém o tempo computacional é

muito maior.

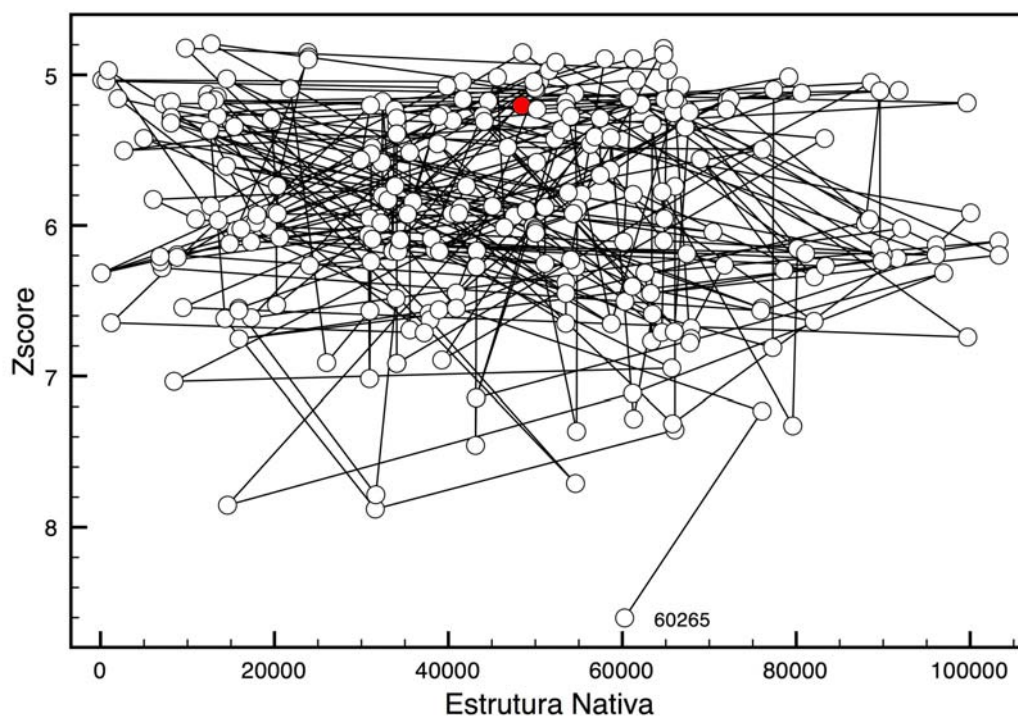


Figura 3.4: Representação de um conjunto de estruturas que apresentam formato de “funil”. Somente uma conformação de mais alto Z_{score} foi visitada. Outras simulações não acessam estruturas presentes neste grupo, exceto em casos de Z_{score} extremamente baixo (presentes no topo do gráfico). Os pontos de partida da simulação estão marcados com cores e as linhas retas verticais representam a otimização da seqüência sobre a estrutura momentânea.

Todo conjunto de resultados da simulação, quando esta alcança uma estrutura final com valor de $Z_{score} \geq 8.5$, é considerado a princípio um funil. Os funis são comparados entre si em busca de estruturas comuns. Quando uma é encontrada, dizemos que os funis estão interligados. Se o Z_{score} das estruturas comuns é muito baixo (abaixo de 5.5) possivelmente eles estão ligados por uma região não muito importante e que nem devem ter características de proteínas. Quando ligados por estruturas com valores mais altos de Z_{score} dizemos que existe

um “*Balde de estruturas*”. O balde interliga diretamente estruturas de iguais estabilidades.

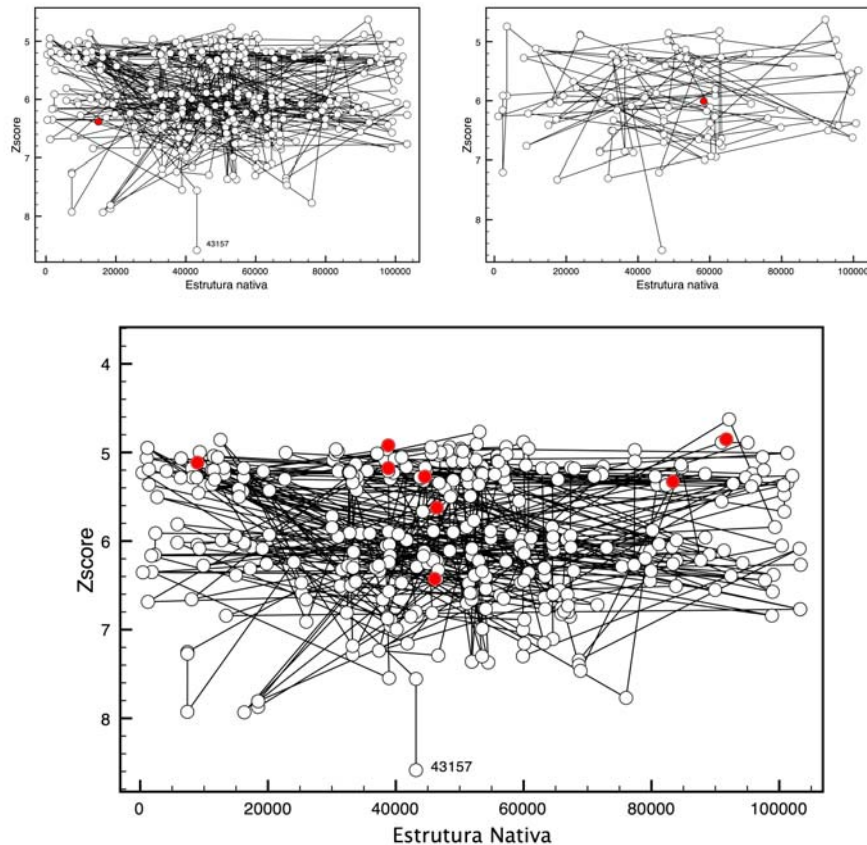


Figura 3.5: Os gráficos no topo representam duas corridas distintas e as bolas coloridas os pontos de partida. Cada uma das simulações alcançam estruturas finais diferentes: 43157 e 46646. Quando analisados, cada um dos funis, 8 coincidências foram encontradas. Neste caso, o grupo presente em um gráfico é considerado automaticamente como parte do outro e adotamos que ele assume a forma de “balde”. No gráfico presente na parte inferior em tamanho ampliado estão identificados em vermelho os pontos em comum.

Capítulo 4

Conclusões

Neste trabalho investigamos a capacidade de armazenar o maior número de letras por uma estrutura. Essa característica nos permite identificar proteínas mais robustas a mutações. Quanto maior o número de letras possíveis em uma conformação, mais nos aproximamos das regiões resistentes a mutações.

O número de tipos de monômeros em uma seqüência também apresenta mudanças no contexto do enovelamento. Um número reduzido de letras pode gerar armadilhas durante a cinética de enovelamento, deixando-a lenta. O acréscimo de letras, quando suportado pela estrutura, pode causar uma melhora na enovelabilidade, aumentando a especificidade da seqüência. No estudo, foi verificado a existência de um número limite, sugerindo um número ótimo de monômeros para cada estrutura. Assim, o número restrito de tipos de aminoácidos estaria vinculado a liberdade conformacional da estrutura (assim como dos graus de liberdade espaciais).

O número máximo de tipos de monômeros de que podem se alocar em uma estrutura nativa não frustrada está associado a topologia de sua estrutura nativa. Quanto maior o número de letras possíveis, maior a robustez da seqüência/estrutura frente a mutações. O número restrito de estruturas que permitem alta desenhabilidade está associado ao número restrito de estruturas que

são bem otimizáveis. Esta afirmação é apoiada pelo fato de existir um conjunto restrito de “*motifs*” estruturais em proteínas reais comparado ao número de seqüências.

Um conjunto limitado das seqüências geradas apresentam características de heteropolímeros, o que mostra a não viabilidade de comparações entre seqüências aleatórias, principalmente quando estas são feitas a casos otimamente desenhados.

O conceito de “*Funil*” ou “*Balde*” de estruturas parece ser suficiente para explicar a inter-relação evolutiva entre diferentes conformações e a grande disparidade na relação número de seqüências/estruturas depositadas nos bancos de dados de proteínas. Nele existe uma seleção natural na organização tridimensional até sua forma estável e daí funcional. Situações desfavoráveis na seqüência obrigam esta a procurar uma nova conformação, mais estável que a anterior, mesmo não sendo otimamente desenhada para esta. Este ajuste segue até que variações de seqüências otimamente desenhadas não se apresentem como um problema na conformação suficientemente evoluída.

Resultados com somente 1 estrutura de alto Z_{score} final e outras com mais de 1 sugerem funis e baldes com diferentes rotas evolutivas para conformações onde no topo existe um alto grau de estruturas conectadas. Com o aumento do Z_{score} , o número de conexões diminui dirigindo a um restrito grupo de conformações.

O “*Funil de estruturas*” compila em si conceitos de desenhabilidade e de escolha de estruturas preferenciais propostas por outros autores por uma enumeração completa associada ao tempo.

Capítulo 5

Apêndices

5.1 Publicações

As páginas seguintes apresentam os trabalhos “*Frustration and Hydrophobicity interplay in protein folding and protein evolution*” publicado no *The Journal of Chemical Physics* em 2006 e “*Geometrical Features of the Protein Folding Mechanism are a robust property of the Energy Landscape: A detailed investigation of several reduced models*” no *The Journal of Physical Chemistry B* em 2008.

No primeiro, uma análise dos diferentes critérios para temperatura de enovelamento proposto por diversos autores é avaliado e discutido para o modelo de rede cúbica proposto aqui. Uma extensiva enumeração para o caso de mutações simples e duplas é realizada e estudos entre seqüências otimamente desenhadas e frustradas concluem o trabalho, correlacionando diferentes cenários de enovelamento sob a perspectiva evolutiva.

No segundo trabalho, diferentes parâmetros e configurações foram avaliados para modelos baseados em estrutura. Essas variações fornecem uma visão dos limites onde resultados experimentais são reproduzidos pelo modelo. Em um caso específico foi avaliado a influência da cadeia lateral para uma proteína com rota de enovelamento bem definida.

Frustration and hydrophobicity interplay in protein folding and protein evolution

Leandro C. Oliveira, Ricardo T. H. Silva, Vitor B. P. Leite,^{a),b)} and Jorge Chahine^{a),c)}
 Departamento de Física, IBILCE, Universidade Estadual Paulista, São José do Rio Preto,
 São Paulo 15054-000, Brazil

(Received 6 April 2006; accepted 17 July 2006; published online 25 August 2006)

A lattice model is used to study mutations and compacting effects on protein folding rates and folding temperature. In the context of protein evolution, we address the question regarding the best scenario for a polypeptide chain to fold: either a fast nonspecific collapse followed by a slow rearrangement to form the native structure or a specific collapse from the unfolded state with the simultaneous formation of the native state. This question is investigated for optimized sequences, whose native state has no frustrated contacts between monomers, and also for mutated sequences, whose native state has some degree of frustration. It is found that the best scenario for folding may depend on the amount of frustration of the native structure. The implication of this result on protein evolution is discussed. © 2006 American Institute of Physics. [DOI: 10.1063/1.2335638]

I. INTRODUCTION

Proteins are not random heteropolymers, but have, rather, been selected through evolution.¹ The effect of mutation on the stability of proteins is a crucial issue in protein evolution. Theoretical studies generally emphasize, following Gö's principle of minimal frustration, that good sequences must be optimized.² It is accepted that biological proteins have been selected through the natural evolution of the species. It is implied that there must be some tolerance in this optimization while still allowing a protein to exhibit satisfactory thermodynamic stability and efficient folding dynamics. The tolerance to amino acid substitution has been observed experimentally.³⁻⁵ The motivation for the present study is to understand (i) how mutations affect optimally designed protein sequences and (ii) the role of hydrophobicity in the mutation processes.

Much work on hydrophobicity has been done in an attempt to answer the following questions: Do compact conformations due to hydrophobic collapse help protein folding?⁶⁻¹¹ Which scenario would proteins choose in order to fold faster: a fast nonspecific collapse followed by a slow rearrangement to reach the native state or a specific collapse with simultaneous formation of the native state? While studies have shown that some proteins undergo a burst hydrophobic collapse followed by their folding,¹²⁻¹⁴ there is experimental evidence that some proteins collapse concomitantly with the formation of their native structure.¹⁵ Some studies⁶⁻⁸ correlate folding kinetics with four parameters defined as follows: (i) The folding temperature T_f is defined as that where half of the chains are folded, or, in protein models used in simulations, as that at which the probability for finding half of the native contacts is 0.5; (ii) T_g is the glass temperature at which the kinetics is dominated by traps due

to many local minima in the energy landscape and at which the kinetics deviates from exponential behavior; (iii) T_θ is a temperature associated with a nonspecific collapse of the chain, which represents the burst hydrophobic collapse of some proteins; (iv) the stability gap is the difference between the energy of the native structure and all other states. These quantities have been used to define dimensionless parameters, which have revealed a correlation with the folding rates. One of these parameters, defined as $\sigma = (T_\theta - T_f)/T_\theta$, was introduced by Klimov and co-worker,^{10,11} who demonstrated that fast folding sequences have small values of σ , which means that sequences that display a specific collapse with simultaneous formation of the native structure fold faster. This behavior precludes the scenario where the chain collapses to nonspecific structures, after which it rearranges itself slowly to find the native structure. Experimental data from small-angle x-ray scattering and circular dichroism⁹ have corroborated the theoretical studies of Klimov and co-worker by showing that the proteins which fold the fastest are those associated with small values of σ . On the other hand, Chiu and Goldstein⁸ have shown, through the use of the diffusion equation, that marginally stable proteins fold faster in the presence of a nonspecific interaction that favors compact states. Whether collapse occurs before folding was also a matter of analysis in a theoretical study by Gutin *et al.*¹⁶ Their results suggested that, if an overall attraction among residues dominates, then collapse precedes folding. As regards to the requisites for fast folding, the ratio T_f/T_g , proposed by Onuchic *et al.*⁷ and Gillespie and Plaxo¹⁷ on simplified models, was shown to be correlated with the folding rates. A different criterion proposed by Sali *et al.*⁶ correlates fast folding kinetics with a large stability gap.

The main purpose of the present study is to show that the way that the folding rates change with temperature and also with the degree of frustration of the native state due to mutation depends significantly on the regime of hydrophobicity. This paper is organized as follows. In Sec. II the lattice

^{a)}Authors to whom correspondence should be addressed.

^{b)}Electronic mail: vleite@ibilce.unesp.br

^{c)}Electronic mail: chahine@ibilce.unesp.br

model, mutation procedure, and simulation methods are discussed. In Sec. III, starting from a native nonfrustrated sequence, all possible sequences obtained by the mutation procedure are analyzed. Two selected mutated sequences, along with the native one, are studied in detail. In Sec. IV the criteria for folding transition temperature are discussed. In Sec. V the thermodynamic and kinetic behaviors of the selected sequences are described. In Sec. VI, the implications of the results are discussed.

II. MODEL AND METHODS

The model used for the kinetic simulations has been extensively employed in previous studies.^{18–20} The protein is modeled by a 27-length polymer chain (27-mer) on a three-dimensional cubic lattice. The energy of a given three-dimensional protein configuration is associated with the interaction between nonbonded monomers, and it is given by

$$E = N_l E_l + N_u E_u, \quad (1)$$

where N_l is the number of contacts between monomers of the same type (like contacts) and N_u the number of contacts between monomers of different types (unlike contacts). We used as a native sequence a three-letter code sequence ABABBBBCBACBABABACACBACAACAB (to which we gave the name native sequence), which forms a well-known nonfrustrated (optimized) native structure.^{20,21} We studied the thermodynamics and kinetics of this lattice model at the low- and high-hydrophobicity limits. The hydrophobicity of this model was discussed in detail by Chahine *et al.*²⁴ The average nonbonded contact energy is proportional to E_u and E_l and the relative frequency of the u and l contacts. The hydrophobicity is normalized by the dispersion in the contact energies, which is associated with the roughness of the energy landscape. In our simulations, $E_l = -1$, $E_u = -3$ regime yields a favorable energy for contact formation, characterizing the occurrence of collapse; this is the high-hydrophobicity (HH) limit. The low-hydrophobicity (LH) limit is defined by $E_l = 3$, $E_u = -3$, which shows on average no attraction between monomers. In a $3 \times 3 \times 3$ cube conformation, the maximum value of the number of contacts is 28, and so the lowest possible energy is -84 . This is the energy of the unfrustrated (native) state, which means a state where there are no contacts between monomers of different types. For the 27-mer it is possible to generate all the cube conformation,²¹ which makes it possible to verify that the mentioned sequence has a nondegenerate ground state.

The mutated sequences were obtained by the permutation of pairs of different monomers. In this way, the proportion of monomers A , B , and C is maintained constant. The native structure for the mutated sequences was found among the 103 346 maximally compact conformations. Mutations were classified based on its frustration, gap, and Z_{score} . Frustration f in a sequence is quantified by the number of unfavorable contacts in the native state, since the number of frustrated contacts raises the energy of the native state; frustration has a direct influence on the native state's stability and kinetics,

$$\text{gap} = E_1 - E_0, \quad (2)$$

where E_0 is the native state energy and E_1 the energy of the first excited states, and

$$Z_{\text{score}} = \frac{\langle E \rangle - E_0}{\sigma}, \quad (3)$$

where $\langle E \rangle$ is the average energy and σ the energy standard deviation. These parameters for each sequence were calculated taking into account only the maximally compact structures. Good mutations in principle maximize both gap and Z_{score} . These parameters display good correlation with the stability of the ground state, even when only maximally compact conformations are considered.²²

In the folding simulations, we used the Monte Carlo algorithm with the standard polymer local lattice moves, which are end, corner flip, and 90° crankshaft moves.²¹ In order to find the density of states $\Omega(E, Q, Z)$ we used the single histogram Monte Carlo method.²³ Once the density is known the averages of the quantities of interest such as the mean energy $\langle E \rangle$, the average of the number of all contacts $\langle Z \rangle$, and the average of the number of native contacts $\langle Q \rangle$ can be calculated. Normalized histograms are used as an approximation for the probability and, once the degeneracy of the ground state is known [$\Omega(-84, 28, 28) = 1$], the free energy is readily obtained, which allows for the determination of the density of states.

III. MUTATIONS

Starting from the initial native sequence N , mutations were obtained by permuting the position of two monomers of different types in the sequence. This procedure follows some criteria. Among many mutated sequences generated by these permutations, it is selected sequences that have the same nondegenerate ground state as that of the native sequence. The feasibility of this procedure is guaranteed by the known maximally compact cube conformations. The mutated sequence is threaded in all the cube conformations, making it possible to verify the degeneracy of the native state. For one permutation there are 236 different sequences. Of these sequences, 206 present a single native state, and 198 (96%) present the same native structure as N . For two permutations there are 19 815 different sequences. Of these sequences 8933 present a single native state, and 3991 (45%) present the same native structure as N . The mutated sequence score distributions of gap versus Z_{score} for one and two mutations are shown in Fig. 1.

“Good” and “bad” mutations were classified based on f , gap, and Z_{score} . In the process of evolution of a protein, it is expected that good mutations fold satisfactorily, with comparable folding times and folding temperatures to the native nonfrustrated sequence. Following this reasoning, mutations with considerably longer folding time and a less stable native state are unlikely to survive the natural selection process, since they do not satisfy this folding criterion. The distributions of gap versus Z_{score} and f vs Z_{score} are shown in Fig. 1. Two good mutations, in which f equals 3 and 6 and with the same native structure as the native one, were selected and

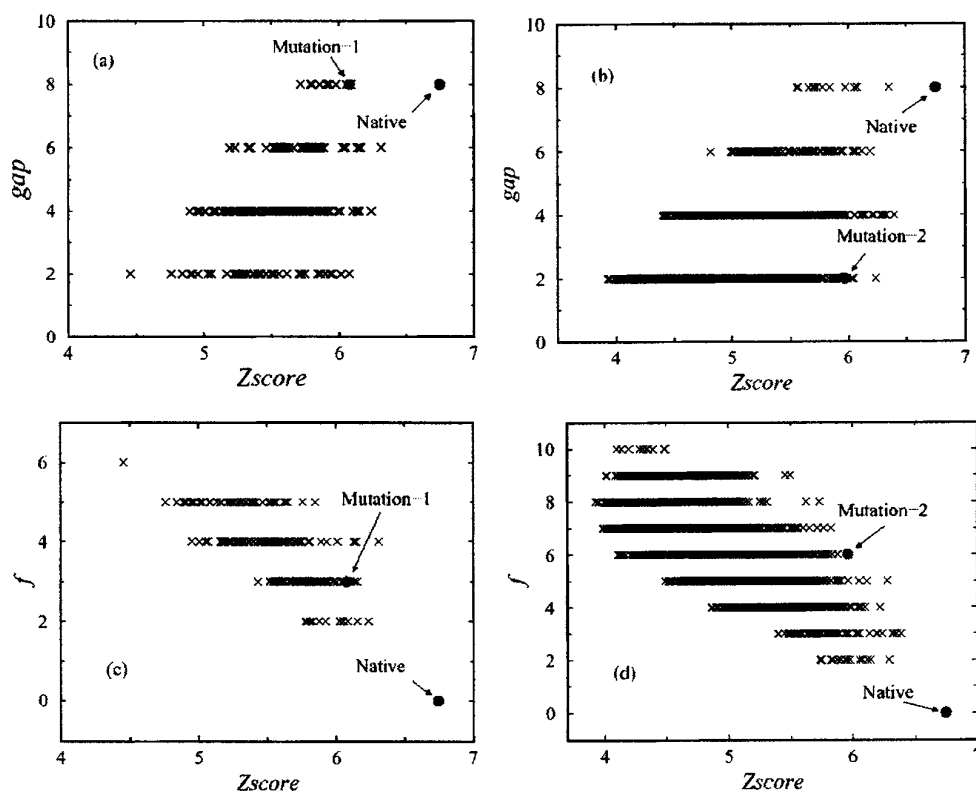


FIG. 1. Gap vs Z_{score} is shown for mutation 1(a) and mutation 2(b); the number of frustrated contacts f vs Z_{score} is shown for mutation 1(c) and mutation 2(d). The native and selected good mutations are marked. Mutation 1 and mutation 2 have, respectively, three and six frustrated contacts in the native conformation.

their thermodynamic and kinetic behaviors were compared with those of the native one. From the single mutation distribution [Figs. 1(a) and 1(c)], a good sequence with three frustrated contacts was ABABBBBCBACBACACACBABAACAB (mutation 1). From the double mutation distribution [Figs. 1(b) and 1(d)], a good sequence with six frustrated contacts was CBABBBBCBCCBABAACACBAAAAAAB (mutation 2). If the argument of robust good mutations is reasonable and, from the distributions of Fig. 1, one would expect that the selected sequences (mutation 1 and mutation 2) should provide reasonable folding features. For constant values of gap and f , both mutations have high Z_{score} compared to the average Z_{score} .

IV. CRITERIA FOR FOLDING TEMPERATURE

Three criteria have been used in simulation studies of protein models in order to define the folding temperature: (i) P_{nat} is the probability to find the native structure calculated from histogram techniques,²³ and the folding temperature is that which makes the probability equal to 0.5; (ii) $\langle Q \rangle$ is the average number of native contacts between the monomers that form the chain normalized by the total number of native contacts in the native conformation, in which case the folding temperature is that which turns the value of $\langle Q \rangle$ to 0.5; (iii) $\langle Q^*Q \rangle - \langle Q \rangle^2$ is the fluctuation of $\langle Q \rangle$, where the folding temperature is that for which the fluctuation has a peak. Since Q is related to the energy, this criterion is similar to the peak of the heat capacity (as a function of the temperature) to

find the folding temperature. We performed simulations for two different sequences: the native one, whose folded structure has no frustration, and the mutation 2 sequence, whose folded structure has six frustrated contacts. The reason for conducting such simulations was due to the possible dependence of the difference in folding temperatures of the two sequences ΔT , according to the three criteria. Once this dependence is calculated, for computational reasons, we chose the criterion that produced the lowest ΔT . The mutated sequence may show a very low folding temperature, which may lead to a high computational time to fold the chain. Figure 2 shows the results for the two sequences, the native (nat) and mutation 2 (mut) in the two regimes of hydrophobicity (LH and HH). The lower portion of the figure refers to the native sequence. Curve (c) is the native sequence in the HH regime and (d) is the native sequence in the LH regime. The higher portion formed by curves (a) and (b) shows the corresponding results for the mutated sequence. When frustration is absent the three criteria give nearly similar values for the folding temperature as shown by curves (c) and (d). The folding temperatures differ by at most 6% for each case. These differences increase by a small amount for the mutated sequence in the HH regime as shown by curve (a). As for the LH regime, the differences are significant when mutation is introduced, as shown by curve (b), which also shows that criterion (2) (represented by the dotted line) produces the smallest ΔT , when comparing the folding temperatures between plots (b) and (d). This criterion makes use of $\langle Q \rangle$, the

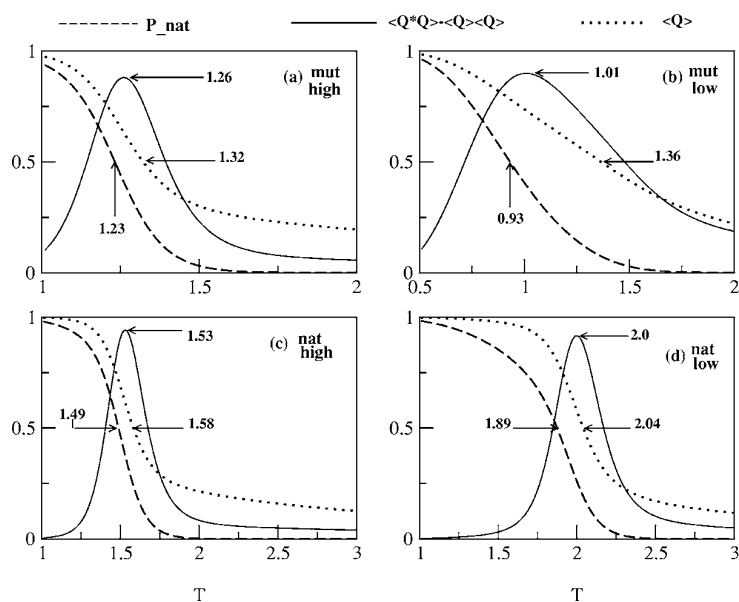


FIG. 2. The figures show the three different criteria for defining the folding temperature for the two sequences, the native (nat) and mutation 2 (mut) in the two regimes of hydrophobicity (LH and HH). Criterion P_{nat} (with dashed line) is the probability of finding the native state. Criterion $\langle Q^*Q \rangle - \langle Q \rangle^*$ (continuous line) is the fluctuation of the number of native contacts. Finally criterion $\langle Q \rangle$ (dotted line) is the average of the native contact, which is the average number of native contacts divided by the total number of native contacts in the native structure (28). Curve (c) is the native sequence in the HH regime and (d) is the native sequence in the LH one. Curves (a) and (b) show the corresponding results for the mutated sequence. In the absence of frustration the three criteria give nearly similar values for the folding temperatures as shown by curves (c) and (d). In the presence of frustration the differences between those temperatures increase by a small amount for the mutated sequence in the HH regime as shown by curve (a), but increase significantly for the LH limit.

normalized average number of native contacts, which is 0.5 at the folding temperature. This is also the order parameter used by Gutin *et al.*¹⁶

Monte Carlo simulations and histogram techniques were used to study the behavior of the native sequence in the two regimes of hydrophobicity. The occurrence of a collapse transition which is not related to folding, i.e., a nonspecific collapse, is studied through the parameter $\langle Z \rangle$, which is the average number of any contacts, native or not. Figure 3(a) shows $\langle Z \rangle$ and its fluctuation as a function of the temperature, for the LH regime. The midtransition for $\langle Z \rangle$ and the peak in the fluctuation occur at temperatures close to $T = 2.0$. In this regime the chain collapses directly to the native structure. Figure 3(b) shows the corresponding quantities for the HH regime. Now the midtransition for $\langle Z \rangle$ and the peak

in its fluctuation occur for temperature $T = 3.0$, which is significantly different from the temperatures shown in Fig. 2(c) (the dotted and continuous lines) for the native quantities $\langle Q \rangle$ and $\langle Q^*Q \rangle - \langle Q \rangle^*$. The increase of $\langle Z \rangle$ and a peak in the fluctuation at a temperature much higher than the folding temperature are indicative of a nonspecific collapse of the chain. This conclusion correlates with the results of Ref. 16: when the overall interaction between monomers is attractive (HH), folding is preceded by a collapse, which does not occur for the LH regime where the overall interaction is nearly zero or slightly repulsive. Collapse was studied in detail for the mutated sequences at HH, and it was observed that the collapse transition always occurs at the same temperature around 2.8.

V. THERMODYNAMICS AND KINETICS

This section sets out the main results of this work, which are related to the folding thermodynamics and kinetics of native and mutated sequences in the two regimes of hydrophobicity. The transition temperatures T_f and T_g and the folding parameter T_f/T_g were calculated for native, mutation 1 ($f=3$), and mutation 2 ($f=6$) sequences at HH and LH regimes. For both hydrophobicity regimes T_g does not depend on the degree of frustration, which is expected, since T_g is associated with the roughness of the landscape and does not depend on the sequence details. T_f decreases significantly with f at LH, and remains approximately constant at HH. The summary of these results is shown through the ratio T_f/T_g in Fig. 4. From this figure it is evident that frustration strongly affects the stability and foldability of a sequence at LH, which suggests that in this regime an optimized sequence is not robust with respect to mutations. That is not the case for the HH regime, which shows no dependence on degree of frustration. The kinetic results reinforce this evidence.

The kinetics of three sequences was studied by measuring their folding time (τ), which was calculated by perform-

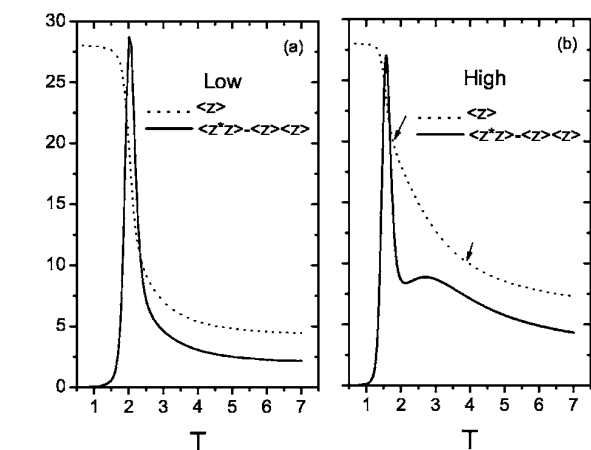


FIG. 3. The curves show the average number of contacts $\langle Z \rangle$ (right axis) and its fluctuation (left axis) as a function of the temperature for LH and HH. For the HH limit, the arrows show a fast increase in $\langle Z \rangle$ which causes the emergence of the smaller peak. Below $T = 2$, the value of $\langle Z \rangle$ experiences an even faster increase due to the folding transition, which is related to the higher peak of the specific heat. The curves for the LH limit show a single transition related to the folding of the chain.

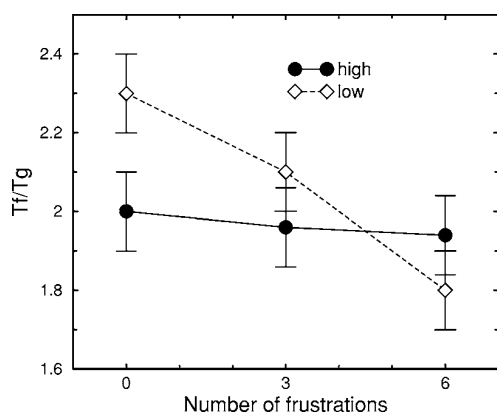


FIG. 4. Folding parameter T_f/T_g as a function of the number of frustrated contacts f in the native conformation and hydrophobicity. $f=0, 3$, and 6 correspond to native, mutation 1, and mutation 2 sequences, respectively.

ing at least 100 independent runs to reach the native state. Twelve values for τ were calculated in the range of temperature $1.2 < T < 3.0$. The results are shown in Fig. 5, where the continuous lines are only a visual guide. Since there is a scaling factor between the temperatures and the energy parameters, the temperature for each sequence and regime of hydrophobicity is normalized by its respective folding transition temperature, T/T_f . By doing this, it is possible to compare their kinetics in the same range of temperature, which means that their folding temperature occurs at the same relative value $T/T_f=1$. Consistent with previous studies,^{25,26} folding mean first passage time, or simply folding time, as a function of temperature has a U-shaped dependence for all sequences and regime of hydrophobicity. τ increases as frustration increases in a very obvious way at LH regime. At HH regime, τ remains approximately constant. For the native sequence, the kinetics at LH is much faster than for the HH. For a mutation 1 sequence ($f=3$), the kinetics in HH and LH regimes have similar rates. For mutation 2 ($f=6$) the kinetics at the LH limit is slower than in the HH limit, and an inversion is seen from Fig. 5(a)–5(c). The results are summarized in Fig. 6, where the minimum folding time τ_{\min} in Fig. 5 is

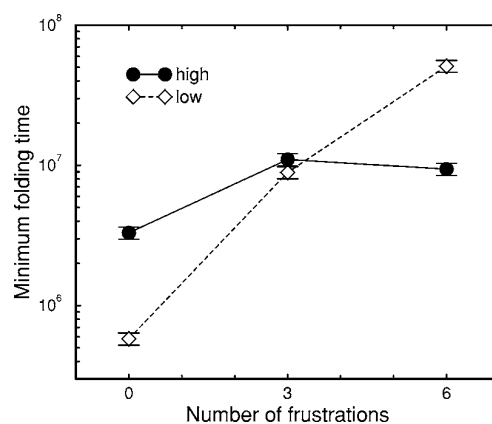


FIG. 6. Minimum folding time of each curve of Fig. 5 as a function of frustration f and hydrophobicity. $f=0, 3$, and 6 correspond to native, mutation 1, and mutation 2 sequences, respectively.

shown as a function of frustration and hydrophobicity. Minimum folding time is a good overall parameter for the kinetics. While τ_{\min} is approximately constant at HH, at LH τ_{\min} varies by about two orders of magnitude.

VI. DISCUSSION

With regard to the folding criteria, the value of the folding temperature is nearly independent of the criteria when the sequence is optimally designed or a small amount of frustration is present in the native state. In the present model, this means that less than 10% of all native contacts are unfavorable. When this amount is increased, the low hydrophobic limit indicates a strong dependence on the chosen folding criteria, whereas the high hydrophobic limit is less sensitive to the choice of the criterion. In the HH limit, characterized by an overall attraction between monomers, folding is preceded by a nonspecific collapse, which is absent at the other limit. This result is in agreement with previous studies of Gutin *et al.*¹⁶ As for the kinetic aspect, it is crucial to study two regimes of stability: the higher stability regime for the native sequence and the lower stability regime for the mu-

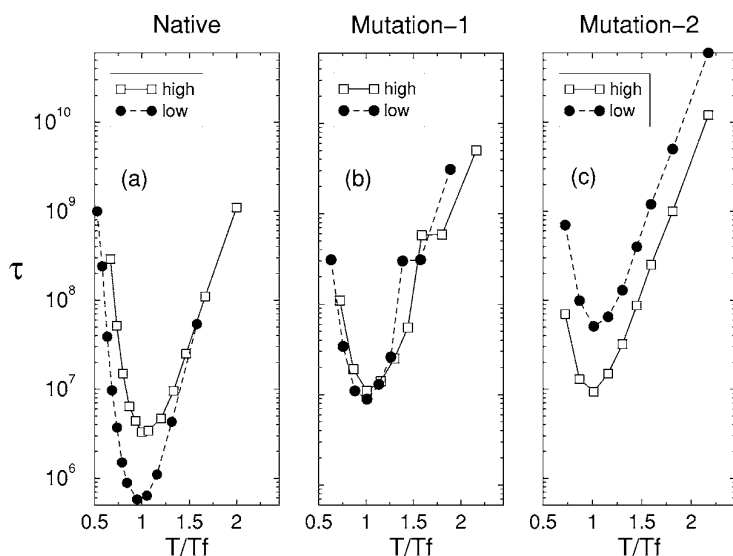


FIG. 5. Folding times for the native (a), mutation 1 (b), and mutation 2 (c) sequences as a function of temperature at the two limits of hydrophobicity. For a better comparison, the temperature for each sequence and regime of hydrophobicity is normalized by its respective folding transition temperature, T/T_f .

tated sequences. The criterion of Thirumalai and Klimov¹¹ states that the chain folds faster when the parameter $\sigma=(T_\theta - T_f)/T_\theta$ is relatively small which is consistent with $T_\theta \cong T_f$ or the absence of a nonspecific collapse. In the absence of frustration in the native state, i.e., strongly optimized sequence and high stability, the present results are also in agreement with the above criterion. On the other hand, for a lower stability regime, due to a certain amount of frustration in the native state of the protein, which in this model means that 20% of all native contacts are unfavorable, the simulations in Fig. 5 indicate faster folding rates for an overall attraction between monomers (HH limit). For this lower stability regime, our results are in agreement with the theoretical results of Chiu and Goldstein⁸ who studied the kinetic behavior of marginally stable proteins. They have demonstrated that, for marginal protein stability, compaction, induced by nonspecific interactions, leads to the increase of the folding rates. The results of Figs. 5 and 6 also suggest that, when stability is reduced, the stronger attractive interaction responsible for the chain collapse produces faster folding rates.

Our results suggest that optimally designed sequences, characterized by high stability, will fold faster at the LH limit characterized by the absence of a nonspecific collapse. For less optimized sequences and lower stability, the HH limit, characterized by the presence of a nonspecific collapse preceding folding, provides faster folding rates. The first case correlates with previous results of Thirumalai and Klimov,¹¹ and the second correlates with the results of Chiu and Goldstein.⁸

The nonspecific collapse of the chain may be a way of overcoming frustration existing in the native state and leads us to speculate on issues regarding the evolution of proteins. An evolutionary process that would make sequences optimally designed would remove the nonspecific collapse that precedes folding. On the other hand, if evolutionary pressures did not result in strongly optimized sequences, Figs. 4 and 5 suggest a scenario where nonspecific collapse precedes folding in order to make the folding process faster. Collapse would help folding when some degree of frustration is present in the native state. If the time for optimizing sequences in the huge space of sequences (20^N for N amino acids) would be prohibitively large, especially for large proteins, the resulting sequences, not strongly optimized, would first collapse and then fold in a biologically relevant time; thus, the necessity to further optimize the sequence would no longer exist. These sequences could be the result of the least effort in the evolutionary process of choosing polypeptide chains that could fold in some functional structures. Particularly, for large proteins, the space of sequences would be so vast that it would make strongly optimized sequences improbable. This would imply that large proteins are likely to collapse before folding into their native state. Among all the large proteins whose sequences are not strongly optimized, those that collapse before folding would predominate as suggested by Fig. 5(c). Also, at the HH limit, the folding process is less sensitive to mutations that worsen the sequence design. If we look at the continuous line in Figs. 5 (HH) we see that the increase in folding time is moderate from Fig. 5(a)

and 5(b) and is nearly absent from Fig. 5(b) and 5(c). At the opposite limit, the LH one, represented by the dotted lines, a continuous increase in the folding times is observed, suggesting a strong dependence on mutations. In addition, sequences that experience a nonspecific collapse before folding would prevail after mutations, not only due to the faster kinetics but also because the folding temperature T_f is less sensitive to mutation.

In short, sequences that collapse concomitantly with the formation of their native structure were strongly optimized during evolution and have the fastest rates. Among the sequences which were not strongly optimized by evolution (probably large proteins), those which collapse before folding have faster rates and would be predominant. Thus, the chain collapse would be the result of three features: (i) Sequences not strongly optimized imply an amount of frustration in the native structure, which, in the terms of the present study, means that 20% (or more) of all the native contacts are unfavorable; (ii) for this type of sequence, the HH limit (collapse) provides faster folding rates; (iii) protein evolution seeks optimal biological relevant time rather than the maximal folding rates. By this last statement we mean that collapse could be the result of an evolutionary process that was sufficient for the protein to achieve its biological functions and necessary stability. In this scenario, as far as foldability and stability are concerned, there is no need for further sequence optimization that would remove collapse from the chain folding.

The present results contribute to the debate on the intriguing subject of hydrophobic collapse. Of course, the statements regarding protein evolution should be considered more as plausible explanations than conclusions, which cannot be drawn from this simple study.

ACKNOWLEDGMENTS

The authors are grateful to Dr. José Nelson Onuchic for helpful discussions. Two of the authors (L.C.O. and R.T.H.S.) were supported by CAPES, Brazil. Two of the authors (V.B.P.L. and J.C.) were partially supported by the Brazilian agency CNPq. One of the authors (V.B.P.L.) was supported by FAPESP, Brazil.

¹ A. Frauenfelder, J. Deisenhofer, and P. Wolynes, *Simplicity and Complexity in Proteins and Nucleic Acids* (Dahlem University Press, Berlin, 1999).

² J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes, *Proteins: Struct., Funct., Genet.* **21**, 167 (1995).

³ D. Rennell, S. E. Bouvier, L. W. Hardy, and A. R. Poteete, *J. Mol. Biol.* **222**, 67 (1991).

⁴ J. Sodek and D. Shortle, *Proteins* **13**, 132 (1992).

⁵ L. H. Weaver, M. G. Grütter, S. J. Remington, T. M. Gray, N. W. Isaacs, and B. W. Matthews, *J. Mol. Evol.* **21**, 97 (1985).

⁶ A. Sali, E. Shakhnovich, and M. Karplus, *J. Mol. Biol.* **235**, 1614 (1994).

⁷ J. N. Onuchic, Z. Luthey-Schulten, and P. G. Wolynes, *Annu. Rev. Phys. Chem.* **48**, 545 (1997).

⁸ T. L. Chiu and R. A. Goldstein, *J. Chem. Phys.* **107**, 4408 (1997).

⁹ I. S. Millet, L. E. Townsley, F. Chiti, S. Doniach, and K. W. Plaxo, *Biochemistry* **41**, 321 (2002).

¹⁰ D. K. Klimov and D. Thirumalai, *Phys. Rev. Lett.* **76**, 4070 (1996).

¹¹ D. Thirumalai and D. K. Klimov, *Curr. Opin. Struct. Biol.* **9**, 197 (1999); C. J. Camacho and D. Thirumalai, *Proc. Natl. Acad. Sci. U.S.A.* **90**, 6369 (1993).

¹² T. Kiefhaber, *Proc. Natl. Acad. Sci. U.S.A.* **92**, 9029 (1995).

- ¹³M. Mucke and F. X. Schmid, *J. Mol. Biol.* **239**, 713 (1994).
- ¹⁴S. Khorasanizadeh, I. D. Peters, T. R. Butt, and H. Roder, *Biochemistry* **32**, 7054 (1993).
- ¹⁵K. W. Plaxo, I. S. Millet, D. J. Segel, S. Doniach, and D. Baker, *Nat. Struct. Biol.* **6**, 554 (1999).
- ¹⁶A. M. Gutin, V. I. Abkevich, and E. I. Shakhnovich, *Biochemistry* **34**, 3066 (1995).
- ¹⁷B. Gillespie and K. W. Plaxo, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 12014 (2000).
- ¹⁸P. E. Leopold, M. Nontal, and J. N. Onuchic, *Proc. Natl. Acad. Sci. U.S.A.* **89**, 8721 (1992).
- ¹⁹A. Li, R. Helling, C. Tang, and N. Wingreen, *Science* **273**, 666 (1996).
- ²⁰N. D. Socci and J. N. Onuchic, *J. Chem. Phys.* **101**, 1519 (1994).
- ²¹E. Shakhnovich and A. Gutin, *J. Chem. Phys.* **93**, 5967 (1990).
- ²²R. I. Dima, J. R. Banavar, M. Cieplak, and A. Maritan, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 4904 (1999).
- ²³A. M. Ferrenberg and R. H. Swendsen, *Phys. Rev. Lett.* **61**, 2635 (1988).
- ²⁴J. Chahine, H. Nymeyer, V. Leite, N. Socci, and J. N. Onuchic, *Phys. Rev. Lett.* **88**, 168101 (2002).
- ²⁵V. B. P. Leite, J. N. Onuchic, G. Stell, and J. Wang, *Biophys. J.* **87**, 3633 (2004).
- ²⁶N. D. Socci and J. N. Onuchic, *J. Chem. Phys.* **103**, 4732 (1995).

Geometrical Features of the Protein Folding Mechanism Are a Robust Property of the Energy Landscape: A Detailed Investigation of Several Reduced Models[†]

Leandro C. Oliveira, Alexander Schug, and José N. Onuchic*

University of California San Diego, Center for Theoretical Biological Physics, La Jolla, California 92093-0374

Received: August 30, 2007; In Final Form: December 6, 2007

The concept of a funneled energy landscape and the principle of minimal frustration are the theoretical foundation justifying the applicability of structure-based models. In simulations, a protein is commonly reduced to a C_α -bead representation. These simulations are sufficient to predict the geometrical features of the folding mechanism observed experimentally utilizing a concise formulation of the Hamiltonian with low computational costs. Toward a better understanding of the interplay between energetic and geometrical features in folding, the side chain is now explicitly included in the simulations. The simplest choice is the addition of C_β -beads at the center-of-mass position of the side chains. While one varies the energetic parameters of the model, the geometric aspects of the folding mechanism remain robust for a broad range of parameters. Energetic properties like folding barriers and protein stability are sensitive to the details of simulations. This robustness to geometry and sensitivity to energetic properties provide flexibility in choosing different parameters to represent changes in sequences, environments, stability or folding rate effects. Therefore, minimal frustration and the funnel concept guarantee that the geometrical features are robust properties of the folding landscape, while mutations and/or changes in the environment easily influence energy-dependent properties like folding rates or stability.

Introduction

Protein folding is a diffusive process where multiple routes lead to transitions from the unfolded (U) to the folded (F) ensembles. A protein's native configuration and native interactions are sufficient to determine its funneled energy landscape.^{1–5} During a long evolutionary process, the energy landscape has been sufficiently smoothed out, minimizing frustration and local roughness. This principle of minimal frustration and the funneled energy landscape provide a sufficiently large bias toward the native state relative to any roughness arising from local minima.

Structure-based models, inspired by the work of Go,⁶ take advantage of these principles. Molecular dynamics (MD) simulations using unfrustrated structure-based models can reproduce Φ -values⁷ in good agreement with experiments.^{8,9} The folding barriers for two-state folders give good qualitative agreement with experimental data.^{10,11} Going beyond folding, recent work investigates multiple folding basins to describe conformational transitions underlying molecular processes in biological systems.^{12–16} Single molecule techniques provide valuable insight into free-energy profiles and kinetic data of conformational transitions.^{17–21} In the case of the ROP-dimer, conformational transitions between two competing states can explain the mutational behavior of this protein by a trapdoor mechanism.^{16,22}

Commonly, in these simple models, the representation of the protein is reduced to single beads localized at the positions of its C_α -atoms. This coarse-grained yet concise minimal model has the additional advantage of low computational demands. This paper, however, focuses on the minimal addition of complexity to the C_α -level, simply by adding an explicit C_β -

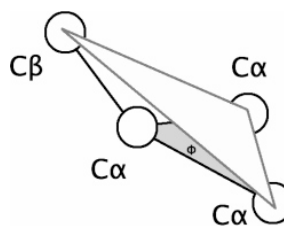


Figure 1. Illustration of improper dihedrals. In molecular dynamics simulations, the angle between two planes defined by four atoms is called its dihedral. The four atoms constituting the two planes of improper dihedrals are branched and not sequentially connected like for proper dihedrals. In $C_\alpha C_\beta$ -structure-based simulations, three C_α -atoms and one C_β -atom form one improper dihedral. The three C_α -atoms form the one plane, the two outer C_α -atoms and the C_β -atom the other plane.

bead for each amino acid at the position of the side chain center-of-mass. This minimal addition is sufficient to start to investigate the interplay between energetics and geometry during folding. For example, it makes it possible to explore the effect of chain packing or to model mutations. This new approach is tested for three different proteins: the chymotrypsin inhibitor 2 (CI2²³), the N-terminal domain of ribosomal protein L9 (L9²⁴) and the colicin E9 immunity protein IM9 (IM9²⁵). Although there is extensive previous work on $C_\alpha C_\beta$ -models,^{26–31} the current goal is to perform a full sensitivity analysis on energetic parameters to demonstrate that geometric properties of the folding mechanism are robust to a broad range of these parameters.

Methods

Diverse approaches can investigate protein folding *in silico*. Among the commonly used methods are knowledge-based homologue modeling,³² molecular dynamics with empirical biomolecular forcefields,^{33–40} global optimization of free-energy

[†] Part of the "Attila Szabo Festschrift".

* To whom correspondence should be addressed. E-mail: Jonuchic@ucsd.edu.

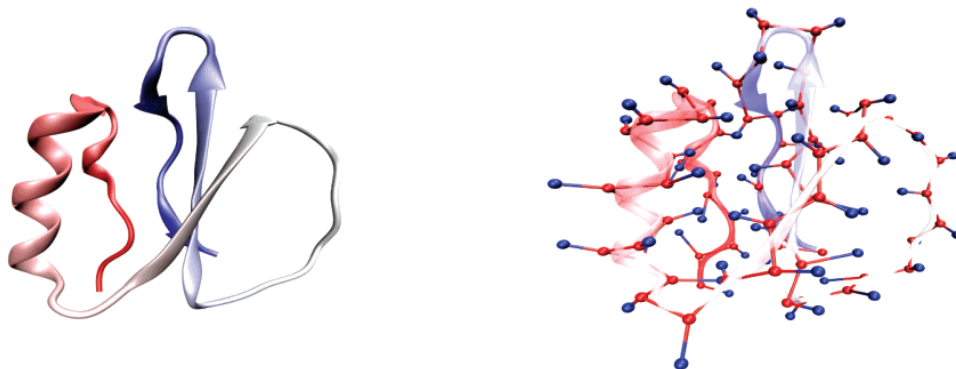


Figure 2. The $C_\alpha C_\beta$ -structure-based model for CI2. Cartoon representation of CI2 (left, PDB code 1ypa) and the coarse-grained model on a $C_\alpha C_\beta$ -level (right). The C_α -atoms are placed at their respective positions, while the C_β -atoms are placed at the center-of-mass position of the individual amino-acid side chains.

functions,^{41–44} lattice-based approaches^{31,45–47} and native structure-based models.^{2,3,5,8,16,22}

In many structure-based models, a protein is reduced to a C_α -bead representation. Molecular dynamics simulations based on C_α -beads have correctly described the folding mechanism of multiple proteins.¹⁰ As long as one does not enter the regime where diffusion becomes the rate-limiting step (for barriers < 3 kT),⁴⁸ the calculated free-energy barriers separating the F and U states and the associated folding rates as obtained from structure-based simulations are in qualitative agreement with experimental data, particularly for geometrical properties.^{10,11,49} It is also possible to investigate solvent/dehydration effects.⁵⁰

Going beyond the simplistic C_α -representation, one can add additional layers of complexity. Ideally, one would like to span the whole range of coarse-grained models from C_α to the all-atom level with the option of adding additional terms such as directional hydrogen bonding.^{26–31,51,52} To achieve this goal, the minimal next step after the C_α -representation is placing explicit C_β -beads at the center-of-mass position of the side chains for all but glycine residues (see Figure 2). The parameters are chosen similar to earlier work.²⁸ The total energy of the system is then given by

$$E_{\text{total}} = E_{\text{bond}} + E_{\text{angle}} + E_{\text{dihedrals}} + E_{\text{LJ}}$$

The bonded energies

$$E_{\text{bond}} = \sum_{\text{bonds}} \frac{1}{2} \epsilon_r (r - r_0)^2$$

are summed over the energy of all covalent bonds. $\epsilon_r = 80$ kT is the bond constant, r is the distance between the two bonded atoms and r_0 is the reference distance of these atoms in the native structure. The angular energy

$$E_{\text{angle}} = \sum_{\text{angles}} \frac{1}{2} \epsilon_0 (\theta - \theta_0)^2$$

has the angle constant $\epsilon_0 = 20$ kT. θ is the angle between two adjacent bonds, and θ_0 is the reference angle in the native structure. The dihedral energy

$$E_{\text{dihedral}} = \sum_{\text{dihedrals}} \left[\epsilon_\phi [1 - \cos(\phi - \phi_0)] + \left[\frac{1}{2} \epsilon_\phi [1 - \cos[3(\phi - \phi_0)]] \right] \right]$$

possesses the dihedral constant ϵ_ϕ and the angle Φ between the two planes formed by four connected atoms. We distinguish between different dihedrals, dependent on the type of the involved atoms. If all four atoms are C_α , $\epsilon_\phi = 0.8$. For the cases $C_\alpha C_\alpha C_\alpha C_\beta$ and $C_\beta C_\alpha C_\alpha C_\alpha$, $\epsilon_\phi = 0.2$. For the case $C_\beta C_\alpha C_\alpha C_\beta$, $\epsilon_\phi = 0.1$. Improper torsions (or improper dihedrals) are implemented using the same equation. For the improper dihedrals, $\epsilon_\phi = 1.0$ (see Figure 1). Improper dihedrals have a maximum value at the cis conformation and help to maintain the system's chirality.

The nonbonded interactions are defined by a protein's contact map obtained with CSU.⁵³ Only contacts at least three amino acids apart in sequence are considered. Each contact is regarded as an attractive interaction between the two involved amino acids. In our $C_\alpha C_\beta$ -model, these interactions are divided into different contributions. The possible combinations are contacts of the type $C_\alpha C_\alpha$, $C_\alpha C_\beta$ and $C_\beta C_\beta$.

$$E_{\text{LJ}} = \sum_{|i-j| \geq 3} \epsilon_{\text{LJ}} \left[5 \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - 6 \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{10} \right] + \sum_{i,j} \epsilon'_{\text{LJ}} \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12}$$

E_{LJ} is the Leonard-Jones potential energy plus a repulsive term. The parameters are $\epsilon_{\text{LJ}} = 0.8$ kT and $\epsilon'_{\text{LJ}} = 0$ kT for formed contacts. If two amino acids are not in contact, $\epsilon_{\text{LJ}} = 0$ kT and $\epsilon'_{\text{LJ}} = 1$ kT prevent atomic clashes. r_{ij} is the distance between any two atoms. σ_{ij} are the native distances for contacts and 4 Å otherwise. In our standard model for $C_\alpha C_\beta$ -simulations, we only include $C_\alpha C_\alpha$ - and $C_\beta C_\beta$ -type contacts. Mixed contacts of the type $C_\alpha C_\beta$ are only considered when indicated, although $C_\alpha C_\beta$ -atoms still interact by the repulsive term given by $\sigma_{ij} = 4$ Å.

Sensitivity Analysis on the Parametric Range

Evaluation of the Simulations. MD simulations at different temperatures and with different parameter sets are used to probe the sensitivity of our model on energetic parameters. Multiple transitions between the F and U states indicate sufficient sampling. The free-energy profiles are calculated with the WHAM algorithm⁵⁴ over the reaction coordinate Q (the fraction of formed tertiary native contacts⁵⁵). Utilizing the simulation data, the folding barrier ΔG is determined as the free-energy barrier between the F and U states at the folding temperature T_F . We define the transition state ensemble (TSE) as the Q -range of the free-energy profile that has an energetic difference of up to 1 kT from the maximum barrier between the F and U states. It is in general around $Q = 0.45$.

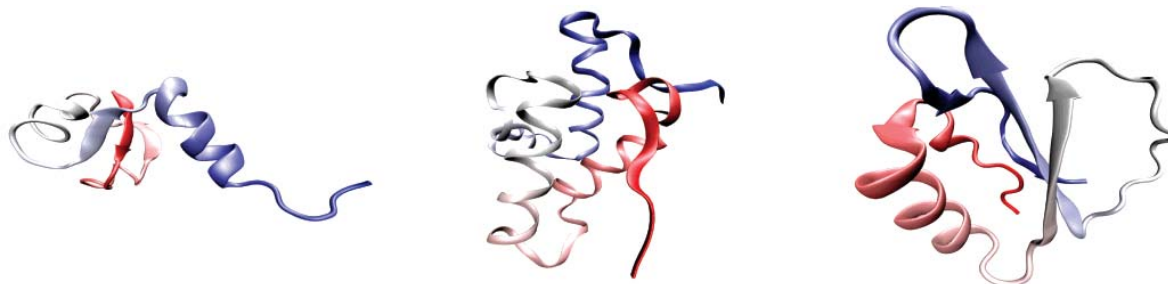


Figure 3. The simulated proteins. The simulated proteins are L9 (left, PDB code 1cqu), IM9 (middle, PDB code 1imq) and CI2 (right, PDB codes 2CI2 and 1ypa). They possess different folds and vary in their secondary structure.

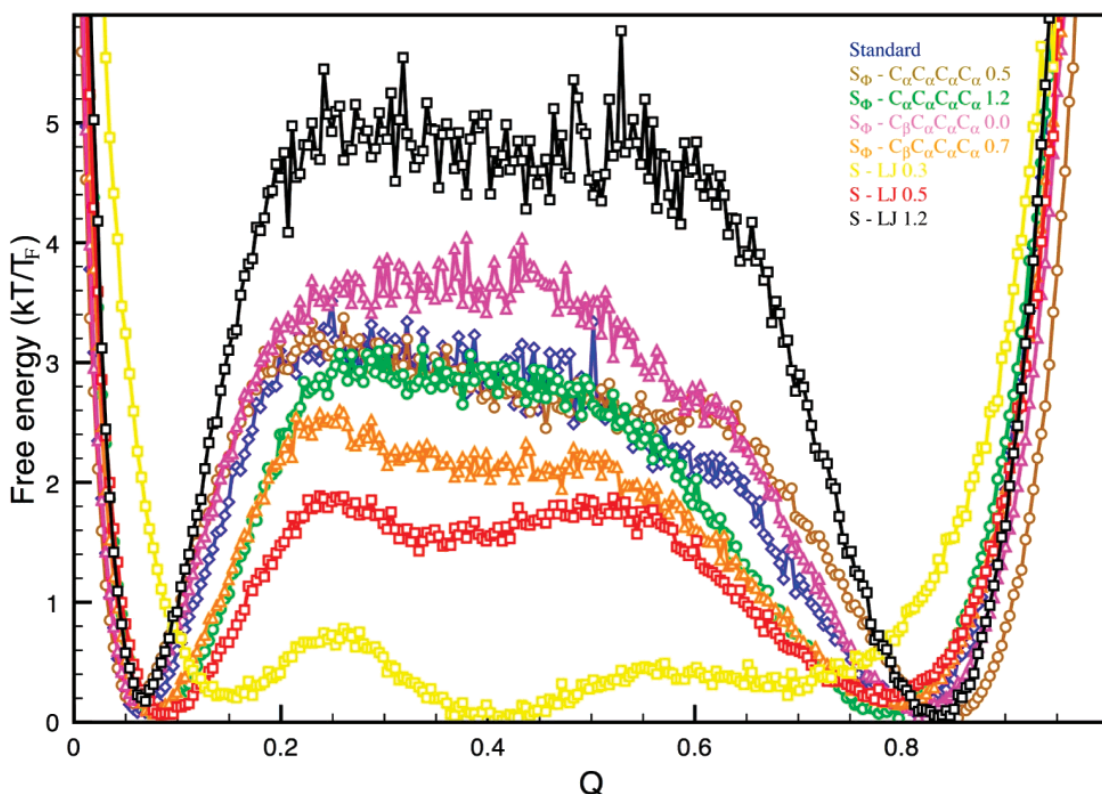


Figure 4. Free-energy profiles of CI2. The colored lines display the free-energy profiles for different parameter sets at their respective T_F 's. The standard model is $C_\alpha C_\beta$ with only $C_\alpha C_\alpha$ - and $C_\beta C_\beta$ -type contacts. The other models use modified interaction strengths. Each $S-X$ is the linear prefactor, which scales the indicated interaction ($S-\Phi$ scales the dihedral, $S-LJ$ the Lennard-Jones interaction). When varying the parameters, the heights of the folding barriers change. The TSE regions are very broad and roughly between $Q = 0.2$ and $Q = 0.6$. The unfolded basins are around $Q = 0.1$, the folded one around $Q = 0.8$. When the Lennard-Jones interaction is decreased, intermediates in the TSE region appear.

Φ -value analysis is an experimental method to probe the TSE by point mutations⁷. Φ -values measure the contribution from each mutated residue to the TSE relative to the native state. In structure-based simulations, contact Φ -values give the probability of contact formation in the TSE.^{56–58} Φ for residue i is calculated as $\Phi_i = (P_{TSE}^i - P_U^i)/(P_F^i - P_U^i)$, with P_X^i being the probability of contact formation for residue i in state X (with X being F, TSE or U). These values provide insight into the folding mechanism. A Φ -value of 0 shows that this contact is never formed in the TSE, while a Φ -value of 1 indicates that the region around this contact is highly nativelike in the TSE and might be part of a folding nucleus.

Similar to Φ -values, the analysis of the probability of contact formation versus the fraction of formed native contacts Q gives

insight into the folding routes and the geometric properties of the energy landscape.¹⁰ At different intervals of Q , one observes the ongoing stabilization of interfaces between the different regions of the protein. The order in which contact formation in different regions of the contact map occurs vs Q defines the folding mechanism, as exemplified in Figure 5.

In some simulations, one or more interactions get *scaled*; i.e., the according interaction is multiplied with a linear prefactor. This allows probing the effect changes of the investigated interaction have on the folding mechanism or also on energetic properties like folding barriers.

The Simulated Proteins. To explore the parametric sensitivity of our model, we modified the parameter strengths for four different protein structures. These different proteins express a

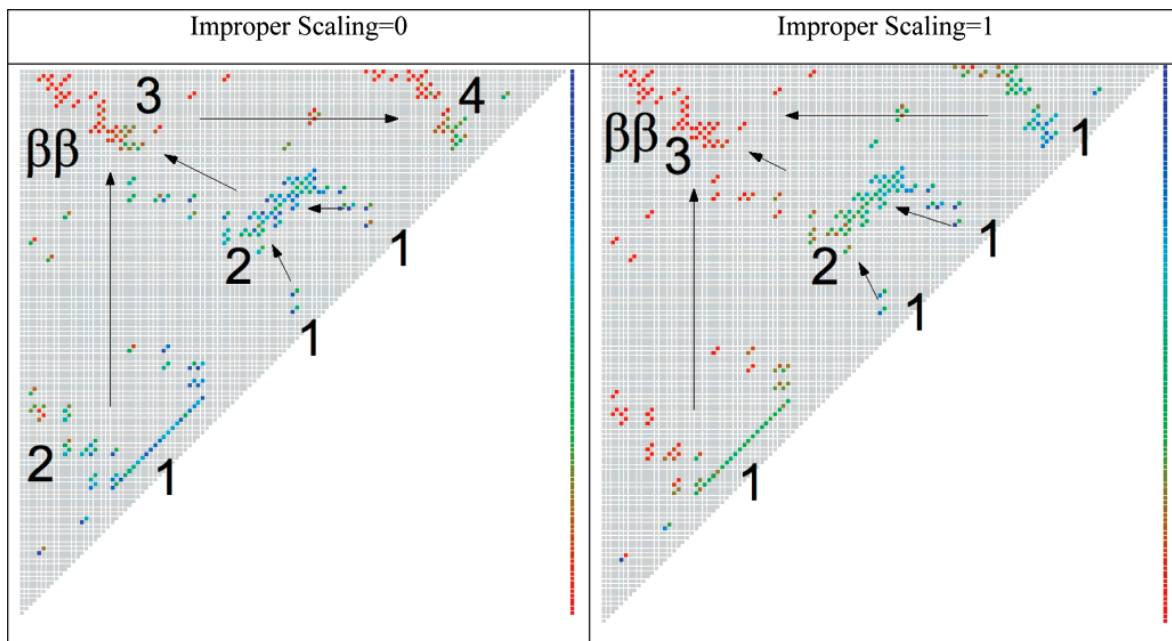


Figure 5. Variation in the folding mechanism for different improper dihedral scaling strengths on CI2. Each image displays contact Φ -values by color from red (“hot”, $P = 0$) to blue (“frozen”, $P = 1$). Each dot represents the probability of the contact formation between amino acids i and j in the TSE. The numbers and arrows indicate the order of events during the structural formation process. When not using improper dihedrals, the N-terminal region starts forming early, while the C-terminal region forms late. We see the opposite order of events when using the improper dihedrals. The folding mechanism with included improper dihedrals agrees with the one observed in experiments and C_{α} -structure-based simulations.

TABLE 1: Varying the Strength of Improper Dihedrals on CI2, L9 and IM9^a

scaling factor improper	T_F			barrier (kT/ T_F)			folding mechanism compared to C_{α} -structure-based simulations		
	CI2	L9	IM9	CI2	L9	IM9	CI2	L9	IM9
0	1.05	1.014	1.092	7.2	3.3	4.1	differ	same	same
0.5	1.098	1.06	1.134	4.6	3.5	4.3	same	same	same
1	1.14	1.1	1.158	3.3	3.2	4.7	same	same	same
1.5	1.174	1.124	1.172	2.4	3.3	3.2	same	same	same

^a When varying the strength of improper dihedrals, both barriers and folding temperatures change. When no improper dihedrals are used, the folding mechanism is altered for CI2. The general behavior upon modification of the dihedral strength can be easily understood as one recalls that strong improper dihedrals restrict the U ensemble the most, the TSE less and have little influence on the already strongly constrained F ensemble. Therefore, the free energies of the U and TSE ensembles rise at a given temperature, resulting in a higher T_F -value. As both U and TSE become concurrently more nativelike, the folding barrier might be lowered.

wide variety of folds from all-alpha to mixed alpha-beta (see Figure 3). Simulations on two different PDB structures for CI2 (PDB codes 2ci2⁵⁹ and 1ypa,⁶⁰ root-mean-square deviation between them = 0.5 Å) investigate how sensitive the simulations are with respect to slight changes in the structure. Additionally, we compared the folding mechanism from C_{α} -type simulations to the $C_{\alpha}C_{\beta}$ folding mechanism under parametrical changes for L9 (1cqu²⁴) and IM9 (1imq²⁵). CI2 has 64 residues that form multiple β -strands and one α -helix. L9 has 56 residues that form two α -helices and three β -strands. IM9 is an all-helical protein with 86 residues that form four α -helices.

Variation of the Improper Dihedral Strength. In contrast to C_{α} -based simulations, $C_{\alpha}C_{\beta}$ -simulations include improper dihedrals. They are named “improper” because the involved atoms are branched instead of being bonded sequentially (see Figure 1). Their purpose is to maintain chirality in the structure. Without these dihedrals, simulations on CI2 express higher barriers and a different folding mechanism (see Figure 5). Strengthening these dihedrals results in increased folding temperatures and decreased folding barriers. This result can be rationalized. Stronger improper dihedrals impose constraints on the U and TSE ensembles. While the highly structured F

TABLE 2: Varying the Included Contacts on CI2^a

contacts included	T_F	folding barrier (kT/ T_F)	number of contacts	folding mechanism compared to C_{α} -structure-based simulations
$C_{\alpha}C_{\alpha}$	0.85	1.1	136	same
$C_{\beta}C_{\beta}$	0.71	0.2	127	same
$C_{\alpha}C_{\alpha}, C_{\beta}C_{\beta}$	1.14	3.3	263	same
$C_{\alpha}C_{\alpha}, C_{\beta}C_{\beta}, C_{\alpha}C_{\beta}$	1.14	3.3	523*	same
$C_{\alpha}C_{\alpha}, C_{\beta}C_{\beta}, C_{\alpha}C_{\beta}$	1.78	3.1	523	same**

^a These data underline the importance of the native contacts for the folding process. The inclusion of $C_{\alpha}C_{\alpha}$ - or $C_{\beta}C_{\beta}$ -contacts alone results in very low barriers, though one already obtains the same folding mechanism as expressed by C_{α} -simulations. When adding both $C_{\alpha}C_{\alpha}$ - and $C_{\beta}C_{\beta}$ -contacts, the folding barriers become comparable to C_{α} -simulations. When the mixed $C_{\alpha}C_{\beta}$ -contacts are added, one sees changes of the barrier shape in the TSE. By scaling the interaction strengths of all contacts by a factor of 0.5 (indicated by one asterisk) to get a similar native energy as in the $C_{\alpha}C_{\alpha}$ - and $C_{\beta}C_{\beta}$ -simulations, one can obtain a highly similar shape of the barrier. When these mixed contacts are added without adjusting the interaction strength of all contacts, the shape of the folding barrier is changed (indicated by two asterisk). In all cases, the folding mechanism is preserved.

ensemble remains unaffected, the conformational space available to the TSE ensemble decreases slightly and that available to

TABLE 3: Varying the Strength of Dihedral and Lennard-Jones Interactions on CI2, L9 and IM9^a

parameter	scaling factor	T_F			barrier (kT/ T_F)			folding mechanism compared to C_α -structure-based simulations		
		CI2	L9	IM9	CI2	L9	IM9	CI2	L9	IM9
standard		1.14	1.10	1.15	3.3	3.2	3.7	same	same	same
$C_\alpha C_\alpha C_\alpha C_\alpha$	0.30	1.05	0.95	0.99	4.9	5.0	5.1	same	same	same
	1.50	1.17	1.27	1.33	3.6	1.9	2.0	same	same	same
$C_\alpha C_\alpha C_\alpha C_\beta$ and	0.00	1.04	1.06	1.13	4.4	3.0	4.5	same	same	same
$C_\beta C_\alpha C_\alpha C_\alpha$	0.70	1.35	1.17	1.23	3.3	3.5	2.2	same	same	same
$C_\beta C_\alpha C_\alpha C_\beta$	0.00	1.27	1.06	1.11	4.3	3.4	5.1	same	same	same
	0.70	1.12	1.32	1.37	2.9	2.0	1.5	same	same	same
Lennard-Jones	0.30	0.64	0.57	0.60	1.8	1.3	0.9	same	same	same
	0.50	0.87	0.80	0.86	2.6	2.1	2.1	same	same	same
	1.50	1.79	1.71	1.79	5.0	4.4	6.1	same	same	same

^a The standard model is $C_\alpha C_\beta$ with only $C_\alpha C_\alpha$ - and $C_\beta C_\beta$ -type contacts. $C_\alpha C_\alpha C_\alpha C_\alpha$, $C_\alpha C_\alpha C_\alpha C_\beta$, $C_\beta C_\alpha C_\alpha C_\alpha$ and $C_\beta C_\alpha C_\alpha C_\beta$ represent different sets of dihedrals. In spite of the strong parametric variations, these are not sufficient to significantly change the folding mechanism. However, the energetic properties, like folding temperatures and barriers, are affected strongly. Increasing the dihedral interaction strengths tends to decrease the folding barrier. In contrast, increasing the Lennard-Jones interaction has an opposite effect. In both cases, the folding temperature is increased.

the U ensemble decreases strongly, as natively structures are preferred for those ensembles. As both U and TSE become more nativelylike, the folding barrier might slightly decrease (Table 1). Concurrently, the destabilization of U compared to F results in a higher T_F .

Simulations on IM9 and L9 also show an increase in T_F when increasing the improper dihedral strength (see Figure 5). The impact on the folding barriers is weaker for IM9 than for CI2, and the barriers remain mostly the same for L9. In contrast to CI2, the folding mechanism stays stable upon removal of the improper dihedrals. For the following simulations, we maintain a scaling of 1 for the improper dihedrals.

Variation of the Included Contacts. The C_β -beads add an additional layer of complexity. A crucial task is the proper inclusion of the contact map. Should each contact in the contact map be expressed by all possible atomic combinations of the involved two residues? We simulated CI2, IM9 and L9 with different contact-type combinations (Table 2). The folding mechanism is similar, even when only $C_\alpha C_\alpha$ - or $C_\beta C_\beta$ -contacts are used. However, this residual number of contacts results in very low folding temperatures and barriers. Results comparable to the C_α -model are obtained by including both $C_\alpha C_\alpha$ - and $C_\beta C_\beta$ -contacts. Including mixed contacts of the type $C_\alpha C_\beta$ demands special care. The sheer number of additional contacts of this type results in strong energetic changes and a change in the folding barrier shape. One can maintain the original barrier shape by scaling these contacts with a linear prefactor, so that the stabilizing energy for the native state remains similar to the other simulations. Apart from these changes in the folding barriers, the folding mechanisms remain unchanged for all investigated cases. These results, together with the fact that these mixed contacts are in general between the same amino acids as the $C_\alpha C_\alpha$ - and $C_\beta C_\beta$ -contacts, one can therefore simplify the model by ignoring those contacts without risking changed folding behavior.

Modifying the Parameter Strength of the Dihedral and Lennard-Jones Interaction. The contact map is a crucial part of structure-based simulations, as it defines the Lennard-Jones interaction strength. We scale this interaction for simulations of four different protein structures to investigate the geometric and energetic sensitivity of folding. Additionally, we scaled the strength for different subsets of dihedral interaction strength.

Simulations of the two different PDB structures for CI2 should ensure that our simulations are consistent for slightly dissimilar native structures. The contact maps have 136 contacts for 2ci2 and 142 contacts for 1ypa. In simulation, they show small differences in the folding barrier heights but the same folding mechanism and ϕ -values (data not shown).

Similarly, for all four investigated proteins, scaling the Lennard-Jones and dihedral interactions strongly affects the folding barrier and T_F (see Table 3 and Figure 4). In all of these cases, this appears to be a result from the changed energetics in the system. In contrast, the scaling does not significantly change the folding mechanism.

Conclusions and Perspectives

In minimal C_α -structure-based simulations, all degrees of freedom real proteins possess are condensed into a few interactions which capture the essential properties. However, a pure C_α -bead description cannot capture effects like the sterics of side chain packing. As a minimal next step, we added an explicit representation of C_β -beads at the center-of-mass position of the side chains to structure-based simulations. In this work, we show that the geometric properties of the energy landscape and therefore the folding mechanism are robust as we move from the C_α - to the $C_\alpha C_\beta$ -level and is maintained on the $C_\alpha C_\beta$ -level over a broad range of parameters. Since the geometric properties are insensitive to the parametric details, we now have the flexibility to adjust the energetic parameters to represent different sequences, environments, stability and folding rate effects. Though outside the scope of the current investigation, this is an important finding, as $C_\alpha C_\beta$ -structure-based simulations are the simplest models with a reasonable realistic side chain representation. These models pave the way to future work addressing mutations and the steric effects of side chain packing on folding.

Acknowledgment. This work was funded by the National Science Foundation-sponsored Center for Theoretical Biological Physics (grants Phy-0216576 and 0225630) and also by grant MCB-0543906. L.C.O. thanks the Brazilian Agency CAPES for financial support. The authors thank Jorge Chahine and Paul C. Whitford for insightful discussions on modeling. We would also like to thank the anonymous referees for their helpful comments.

References and Notes

- (1) Frauenfelder, H.; Sligar, S. G.; Wolynes, P. G. *Science* **1991**, *254*, 1598.
- (2) Bryngelson, J. D.; Onuchic, J. N.; Socci, N. D.; Wolynes, P. G. *Proteins: Struct., Funct., Genet.* **1995**, *21*, 167.
- (3) Clementi, C.; Nymeyer, H.; Onuchic, J. N. *J. Mol. Biol.* **2000**, *298*, 937.
- (4) Lyubovitsky, J. G.; Gray, H. B.; Winkler, J. R. *J. Am. Chem. Soc.* **2002**, *124*, 5481.
- (5) Onuchic, J. N.; Wolynes, P. G. *Curr. Opin. Struct. Biol.* **2004**, *14*, 70.

- (6) Ueda, Y.; Taketomi, H.; Go, N. *Biopolymers* **1978**, *17*, 1531.
- (7) Fersht, A. R. *Curr. Opin. Struct. Biol.* **1995**, *5*, 79.
- (8) Clementi, C.; Jennings, P. A.; Onuchic, J. N. *J. Mol. Biol.* **2001**, *311*, 879.
- (9) Lindberg, M.; Tangrot, J.; Oliveberg, M. *Nat. Struct. Biol.* **2002**, *9*, 818.
- (10) Chavez, L. L.; Onuchic, J. N.; Clementi, C. *J. Am. Chem. Soc.* **2004**, *126*, 8426.
- (11) Clementi, C.; Plotkin, S. S. *Protein Sci.* **2004**, *13*, 1750.
- (12) Best, R. B.; Chen, Y. G.; Hummer, G. *Structure* **2005**, *13*, 1755.
- (13) Hyeon, C.; Lorimer, G. H.; Thirumalai, D. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 18939.
- (14) Okazaki, K.; Koga, N.; Takada, S.; Onuchic, J. N.; Wolynes, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 11844.
- (15) Whitford, P. C.; Miyashita, O.; Levy, Y.; Onuchic, J. N. *J. Mol. Biol.* **2007**, *366*, 1661.
- (16) Schug, A.; Whitford, P. C.; Levy, Y.; Onuchic, J. N. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 17674.
- (17) Hummer, G.; Szabo, A. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 3658.
- (18) Rief, M.; Grubmüller, H. *ChemPhysChem* **2002**, *3*, 255.
- (19) Hummer, G.; Szabo, A. *Biophys. J.* **2003**, *85*, 5.
- (20) Dudko, O. K.; Mathe, J.; Szabo, A.; Meller, A.; Hummer, G. *Biophys. J.* **2007**, *92*, 4188.
- (21) Dudko, O. K.; Hummer, G.; Szabo, A. *Phys. Rev. Lett.* **2006**, *96*.
- (22) Levy, Y.; Cho, S. S.; Shen, T.; Onuchic, J. N.; Wolynes, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 2373.
- (23) McPhalen, C. A.; Svendsen, I.; Jonassen, I.; James, M. N. G. *Proc. Natl. Acad. Sci. U.S.A.* **1985**, *82*, 7242.
- (24) Luisi, D. L.; Kuhlman, B.; Sideras, K.; Evans, P. A.; Raleigh, D. P. *J. Mol. Biol.* **1999**, *289*, 167.
- (25) Osborne, M. J.; Breeze, A. L.; Lian, L. Y.; Reilly, A.; James, R.; Kleanthous, C.; Moore, G. R. *Biochemistry* **1996**, *35*, 9505.
- (26) Cheung, M. S.; Finke, J. M.; Callahan, B.; Onuchic, J. N. *J. Phys. Chem. B* **2003**, *107*, 11193.
- (27) Takada, S.; Luthey-Schulten, Z.; Wolynes, P. G. *J. Chem. Phys.* **1999**, *110*, 11616.
- (28) Finke, J. M.; Cheung, M. S.; Onuchic, J. N. *Biophys. J.* **2004**, *87*, 1900.
- (29) Ding, F.; Dokholyan, N. V.; Buldyrev, S. V.; Stanley, H. E.; Shakhnovich, E. I. *Biophys. J.* **2002**, *83*, 3525.
- (30) Irback, A.; Sjunnesson, F.; Wallin, S. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 13614.
- (31) Klimov, D. K.; Thirumalai, D. *Folding Des.* **1998**, *3*, 127.
- (32) Ginalska, K. *Curr. Opin. Struct. Biol.* **2006**, *16*, 172.
- (33) Ponder, J. W.; Case, D. A. Force fields for protein simulations. *Adv. Protein Chem.* **2003**, *66*, 27–85.
- (34) Brooks, B. R.; Brucoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187.
- (35) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K. *J. Comput. Chem.* **2005**, *26*, 1781.
- (36) Van der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. *J. Comput. Chem.* **2005**, *26*, 1701.
- (37) Adcock, S. A.; McCammon, J. A. *Chem. Rev.* **2006**, *106*, 1589.
- (38) Gnanakaran, S.; Nussinov, R.; Garcia, A. E. *J. Am. Chem. Soc.* **2006**, *128*, 2158.
- (39) Jayachandran, G.; Vishal, V.; Pande, V. S. *J. Chem. Phys.* **2006**, *124*.
- (40) Scheraga, H. A.; Khalili, M.; Liwo, A. *Annu. Rev. Phys. Chem.* **2007**, *58*, 57.
- (41) Schug, A.; Wenzel, W. *Biophys. J.* **2006**, *90*, 4273.
- (42) Schug, A.; Herges, T.; Verma, A.; Lee, K. H.; Wenzel, W. *ChemPhysChem* **2005**, *6*, 2640.
- (43) Schug, A.; Wenzel, W.; Hansmann, U. H. E. *J. Chem. Phys.* **2005**, *122*.
- (44) Schug, A.; Herges, T.; Wenzel, W. *Phys. Rev. Lett.* **2003**, *91*.
- (45) Thirumalai, D.; Klimov, D. K.; Dima, R. I. Insights into specific problems in protein folding using simple concepts. *Computational Methods for Protein Folding Advances in Chemical Physics*; John Wiley and Sons: New York, 2002; Vol. 120, pp 35–76.
- (46) Oliveira, L. C.; Silva, R. T. H.; Leite, V. B. P.; Chahine, J. *J. Chem. Phys.* **2006**, *125*.
- (47) Vieth, M.; Kolinski, A.; Brooks, C. L.; Skolnick, J. *J. Mol. Biol.* **1995**, *251*, 448.
- (48) Ma, H. R.; Gruebele, M. *J. Comput. Chem.* **2006**, *27*, 125.
- (49) Kubelka, J.; Hofrichter, J.; Eaton, W. A. *Curr. Opin. Struct. Biol.* **2004**, *14*, 76.
- (50) Cheung, M. S.; Garcia, A. E.; Onuchic, J. N. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 685.
- (51) Shimada, J.; Kussell, E. L.; Shakhnovich, E. I. *J. Mol. Biol.* **2001**, *308*, 79.
- (52) Linhananta, A.; Zhou, Y. Q. *J. Chem. Phys.* **2002**, *117*, 8983.
- (53) Sobolev, V.; Sorokine, A.; Prilusky, J.; Abola, E. E.; Edelman, M. *Bioinformatics* **1999**, *15*, 327.
- (54) Ferrenberg, A. M.; Swendsen, R. H. *Phys. Rev. Lett.* **1989**, *63*, 1195.
- (55) Cho, S. S.; Levy, Y.; Wolynes, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 586.
- (56) Levy, Y.; Cho, S. S.; Onuchic, J. N.; Wolynes, P. G. *J. Mol. Biol.* **2005**, *346*, 1121.
- (57) Onuchic, J. N.; Socci, N. D.; Luthey-Schulten, Z.; Wolynes, P. G. *Folding Des.* **1996**, *1*, 441.
- (58) Shoemaker, B. A.; Wang, J.; Wolynes, P. G. *J. Mol. Biol.* **1999**, *287*, 675.
- (59) McPhalen, C. A.; James, M. N. G. *Biochemistry* **1987**, *26*, 261.
- (60) Harpaz, Y.; Elmasry, N.; Fersht, A. R.; Henrick, K. *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *91*, 311.

Referências Bibliográficas

- [1] E. Alm and D. Baker, *Prediction of protein-folding mechanisms from free-energy landscapes derived from native structures.*, Proc Natl Acad Sci U S A **96** (1999), no. 20, 11305–11310.
- [2] A.M. Ferrenberg and R.H. Swendsen, *Optimized monte carlo data analysis.*, Phys Rev Lett **63** (1989), no. 12, 1195–1198.
- [3] C.B. Anfinsen, *Principles that govern the folding of protein chains.*, Science **181** (1973), no. 96, 223–230.
- [4] C.H. Bennett, *Efficient estimation of free-energy differences from monte-carlo data*, J Comp Phys **22** (1976), no. 2, 245–268.
- [5] J.D. Bryngelson, J.N. Onuchic, N.D. Socci, and P.G. Wolynes, *Funnels, pathways, and the energy landscape of protein folding: a synthesis.*, Proteins **21** (1995), no. 3, 167–195.
- [6] J.D. Bryngelson and P.G. Wolynes, *Spin glasses and the statistical mechanics of protein folding.*, Proc Natl Acad Sci U S A **84** (1987), no. 21, 7524–7528.
- [7] C. Clementi, H. Nymeyer and J.N. Onuchic, *Topological and energetic factors: what determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? an investigation for small globular proteins.*, J Mol Biol **298** (2000), no. 5, 937–953.

- [8] C. Clementi, P.A. Jennings and J.N. Onuchic, *How native-state topology affects the folding of dihydrofolate reductase and interleukin-1beta.*, Proc Natl Acad Sci U S A **97** (2000), no. 11, 5871–5876.
- [9] C. Clementi, P.A. Jennings and J.N. Onuchic, *Prediction of folding mechanism for circular-permuted proteins.*, J Mol Biol **311** (2001), no. 4, 879–890.
- [10] R.W. Carrell and B. Gooptu, *Conformational changes and disease—serpins, prions and alzheimer’s.*, Curr Opin Struct Biol **8** (1998), no. 6, 799–809.
- [11] H. Cejtin, J. Edler, A. Gottlieb, R. Helling, H. Li, J. Philbin, N. Wingreen, and C. Tang, *Fast tree search for enumeration of a lattice model of protein folding*, J Chem Phys **116** (2002), no. 1, 352–359.
- [12] J. Chahine, H. Nymeyer, V.B. Leite, N.D. Socci, and J.N. Onuchic, *Specific and nonspecific collapse in protein folding funnels.*, Phys Rev Lett **88** (2002), no. 16, 168101.
- [13] J. Chahine, R.J. Oliveira, V.B. Leite, and J. Wang, *Configuration-dependent diffusion can shift the kinetic transition state and barrier height of protein folding.*, Proc Natl Acad Sci U S A **104** (2007), no. 37, 14646–14651.
- [14] L.L. Chavez, S. Gosavi, P.A. Jennings, and J.N. Onuchic, *Multiple routes lead to the native state in the energy landscape of the beta-trefoil family.*, Proc Natl Acad Sci U S A **103** (2006), no. 27, 10254–10258.
- [15] L.L. Chavez, J.N. Onuchic, and C. Clementi, *Quantifying the roughness on the free energy landscape: entropic bottlenecks and protein folding rates.*, J Am Chem Soc **126** (2004), no. 27, 8426–8432.
- [16] R.I. Dima, J.R. Banavar, M. Cieplak, and A. Maritan, *Statistical mechanics of protein-like heteropolymers.*, Proc Natl Acad Sci U S A **96** (1999), no. 9, 4904–4907.

- [17] A.M. Ferrenberg and R.H. Swendsen, *New monte carlo technique for studying phase transitions.*, Phys Rev Lett **61** (1988), no. 23, 2635–2638.
- [18] A.R. Fersht, *Characterizing transition states in protein folding: an essential step in the puzzle.*, Curr Opin Struct Biol **5** (1995), no. 1, 79–84.
- [19] K. M. Fiebig and K. A. Dill, *Protein core assembly process*, J Chem Phys **98** (1993), no. 4, 3475–3487.
- [20] A.V. Finkelstein, *Rate of protein folding near the point of thermodynamic equilibrium between the coil and the most stable chain fold.*, Fold Des **2** (1997), no. 2, 115–121.
- [21] T.E. Greigton, *Protein folding*, W. H. Freeman and Company, 1992.
- [22] R. Guerois and L. Serrano, *The sh3-fold family: experimental evidence and prediction of variations in the folding pathways.*, J Mol Biol **304** (2000), no. 5, 967–982.
- [23] S.C. Harrison and R. Durbin, *Is there a single pathway for the folding of a polypeptide chain?*, Proc Natl Acad Sci U S A **82** (1985), no. 12, 4028–4030.
- [24] R. Helling, H. Li, R. Melin, J. Miller, N. Wingreen, C. Zeng, and C. Tang, *The designability of protein structures.*, J Mol Graph Model **19** (2001), no. 1, 157–167.
- [25] B. Honig, *Protein folding: from the levinthal paradox to structure prediction.*, J Mol Biol **293** (1999), no. 2, 283–293.
- [26] K.A. Dill, *Theory for the folding and stability of globular proteins.*, Biochemistry **24** (1985), no. 6, 1501–1509.
- [27] K.A. Dill, *Dominant forces in protein folding.*, Biochemistry **29** (1990), no. 31, 7133–7155.

- [28] K.A. Dill, *The meaning of hydrophobicity.*, Science **250** (1990), no. 4978, 297–298.
- [29] K.A. Dill, *Polymer principles and protein folding.*, Protein Sci **8** (1999), no. 6, 1166–1180.
- [30] M. Karplus and D.L. Weaver, *Protein folding dynamics: the diffusion-collision model and experimental data.*, Protein Sci **3** (1994), no. 4, 650–668.
- [31] H. Kaya and H.S. Chan, *Contact order dependent protein folding rates: Kinetic consequences of a cooperative interplay between favorable nonlocal interactions and local conformational preferences*, Proteins Struct Func and Gen **52** (2003), no. 4, 524–533.
- [32] D.K. Klimov and D. Thirumalai, *Cooperativity in protein folding: from lattice models with sidechains to real proteins*, Fol Des **3** (1998), no. 2, 127–139.
- [33] J. Kubelka, J. Hofrichter, and W.A. Eaton, *The protein folding 'speed limit'.*, Curr Opin Struct Biol **14** (2004), no. 1, 76–88.
- [34] S. Kumar, D. Bouzida, R.H. Swendsen, P.A. Kollman, and J.M. Rosenberg, *The weighted histogram analysis method for free-energy calculations on biomolecules .1. the method*, J Comp Chem **13** (1992), no. 8, 1011–1021.
- [35] P.E. Leopold, M. Montal, and J.N. Onuchic, *Protein folding funnels: a kinetic approach to the sequence-structure relationship.*, Proc Natl Acad Sci U S A **89** (1992), no. 18, 8721–8725.
- [36] H. Li, R. Helling, C. Tang, and N. Wingreen, *Emergence of preferred structures in a simple model of protein folding.*, Science **273** (1996), no. 5275, 666–669.

- [37] H. Li, C. Tang, and N.S. Wingreen, *Designability of protein structures: a lattice-model study using the miyazawa-jernigan matrix.*, Proteins **49** (2002), no. 3, 403–412.
- [38] C.L. Masters and K. Beyreuther, *Spongiform encephalopathies. tracking turncoat prion proteins.*, Nature **388** (1997), no. 6639, 228–229.
- [39] R. Melin, H. Li, N. S. Wingreen, and C. Tang, *Designability, thermodynamic stability, and dynamics in protein folding: A lattice model study*, J Chem Phys **110** (1999), no. 2, 1252–1262.
- [40] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller, *Equation of state calculations by fast computing machines*, J Chem Phys **21** (1953), no. 6, 1087–1092.
- [41] V. Munoz and W.A. Eaton, *A simple model for calculating the kinetics of protein folding from three-dimensional structures.*, Proc Natl Acad Sci U S A **96** (1999), no. 20, 11311–11316.
- [42] B. Nolting and K. Andert, *Mechanism of protein folding.*, Proteins **41** (2000), no. 3, 288–298.
- [43] H. Nymeyer, N.D. Socci, and J.N. Onuchic, *Landscape approaches for determining the ensemble of folding transition states: success and failure hinge on the degree of frustration.*, Proc Natl Acad Sci U S A **97** (2000), no. 2, 634–639.
- [44] L.C. Oliveira, A. Schug, and J.N. Onuchic, *Geometrical features of the protein folding mechanism are a robust property of the energy landscape: A detailed investigation of several reduced models.*, J Phys Chem B (2008), 10.1021.

- [45] L.C. Oliveira, R.T. Silva, V.B. Leite, and J. Chahine, *Frustration and hydrophobicity interplay in protein folding and protein evolution.*, J Chem Phys **125** (2006), no. 8, 084904.
- [46] J.N. Onuchic and P.G. Wolynes, *Theory of protein folding.*, Curr Opin Struct Biol **14** (2004), no. 1, 70–75.
- [47] L Pauling and R B Corey, *Atomic coordinates and structure factors for two helical configurations of polypeptide chains.*, Proc Natl Acad Sci U S A **37** (1951), no. 5, 235–240.
- [48] S.S. Plotkin, *Speeding protein folding beyond the $g(o)$ model: how a little frustration sometimes helps.*, Proteins **45** (2001), no. 4, 337–345.
- [49] S.S. Plotkin and J.N. Onuchic, *Investigation of routes and funnels in protein folding by free energy functional methods.*, Proc Natl Acad Sci U S A **97** (2000), no. 12, 6509–6514.
- [50] S. Rackovsky and H.A. Scheraga, *Hydrophobicity, hydrophilicity, and the radial and orientational distributions of residues in native proteins.*, Proc Natl Acad Sci U S A **74** (1977), no. 12, 5248–5251.
- [51] S.E. Radford and C.M. Dobson, *From computer simulations to human disease: emerging themes in protein folding.*, Cell **97** (1999), no. 3, 291–298.
- [52] E.I. Shakhnovich and A.M. Gutin, *Engineering of stable and fast-folding sequences of model proteins.*, Proc Natl Acad Sci U S A **90** (1993), no. 15, 7195–7199.
- [53] B.A. Shoemaker, J. Wang, and P.G. Wolynes, *Exploring structures in protein folding funnels with free energy functionals: the transition state ensemble.*, J Mol Biol **287** (1999), no. 3, 675–694.

- [54] V. Sobolev, A. Sorokine, J. Prilusky, E.E. Abola, and M. Edelman, *Automated analysis of interatomic contacts in proteins.*, Bioinformatics **15** (1999), no. 4, 327–332.
- [55] N. D. Socci and J. N. Onuchic, *Folding kinetics of proteinlike heteropolymers*, J Chemical Phys **101** (1994), no. 2, 1519–1528.
- [56] N D Socci and J N Onuchic, *Kinetic and thermodynamic analysis of proteinlike heteropolymers - monte-carlo histogram technique*, J Chem Phys **103** (1995), no. 11, 4732–4744.
- [57] N.D. Socci, H. Nymeyer, and J.N. Onuchic, *Exploring the protein folding funnel landscape*, Physica D **107** (1997), no. 2-4, 366–382.
- [58] J.D. Watson and F.H. Crick, *Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid.*, Nature **171** (1953), no. 4356, 737–738.
- [59] D.B. Wetlaufer, *Nucleation, rapid folding, and globular intrachain regions in proteins.*, Proc Natl Acad Sci U S A **70** (1973), no. 3, 697–701.
- [60] P.G. Wolynes, *Symmetry and the energy landscapes of biomolecules.*, Proc Natl Acad Sci U S A **93** (1996), no. 25, 14249–14255.
- [61] P.G. Wolynes, J.N. Onuchic, and D. Thirumalai, *Navigating the folding routes*, Science **267** (1995), no. 5204, 1619–1620.
- [62] R. Zwanzig, A. Szabo, and B. Bagchi, *Levinthal's paradox.*, Proc Natl Acad Sci U S A **89** (1992), no. 1, 20–22.