UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Campus de São José do Rio Preto

Ronaldo Junio de Oliveira

Estudo do Coeficiente de Difusão no Enovelamento de Proteína

São José do Rio Preto, 1 de Agosto de 2011

Oliveira, Ronaldo Junio de.
Estudo do coeficiente de difusão no enovelamento de proteína / Ronaldo Junio de Oliveira. - São José do Rio Preto : [s.n.], 2011.
115 f. : il. ; 30 cm.

Orientador: Vitor Barbanti Pereira Leite
Co-orientador: Jorge Chahine
Tese (doutorado) – Universidade Estadual Paulista, Instituto de Biociências, Letras e Ciências Exatas

1. Biofísica molecular. 2. Proteínas - Enovelamento. 3. Coeficiente de difusão. 4. Simulação computacional. 5. Dinâmica molecular. I. Leite, Vitor Barbanti Pereira. II. Chahine, Jorge. III. Universidade Estadual Paulista, Instituto de Biociências, Letras e Ciências Exatas. IV. Título.

CDU – 577.112

Ficha catalográfica elaborada pela Biblioteca do IBILCE
Campus de São José do Rio Preto - UNESP

Ronaldo Junio de Oliveira

Estudo do Coeficiente de Difusão no Enovelamento de Proteína

Tese apresentada como parte das exigências para obtenção do título de Doutor em Biofísica Molecular, área de concentração Biofísica Molecular do Departamento de Física do Instituto de Biociências, Letras e Ciências Exatas da Universidade Estadual Paulista "Júlio de Mesquita Filho" – UNESP, Campus de São José do Rio Preto, São Paulo, Brasil.

BANCA EXAMINADORA

Prof. Dr. Vitor Barbanti Pereira Leite
Professor Doutor
UNESP – São José do Rio Preto
Orientador

Prof. Dr. Antônio Francisco Pereira de Araújo
Professor Doutor
UnB – Brasília

Prof. Dr. Nelson Augusto Alves
Professor Doutor
USP – Ribeirão Preto

Prof. Dr. Luis Paulo Barbour Scott
Professor Doutor
UFABC – Santo André

Prof. Dr. José Roberto Ruggiero
Professor Doutor
UNESP – São José do Rio Preto

São José do Rio Preto, 1 de Agosto de 2011

À minha família...

"This structure has novel features which are of considerable biological interest."

Watson and Crick

April, 1953 – Nature.

Agradecimentos

Foram muitas as pessoas que contribuíram direta ou indiretamente para a realização desse trabalho de doutorado e a elas devo esse agradecimento.

Primeiramente, agradeço o Prof. Dr. Vitor Barbanti Pereira Leite pela excelente orientação e o Prof. Dr. Jorge Chahine pela co-orientação durante os seis anos de minha vida acadêmica. Também, agradeço o Prof. Dr. Jin Wang pela orientação durante meu estágio no exterior.

Ao Dr. Paul Whitford pela colaboração intensa e por ser meu "old brother" em ciência.

Aos professores do Departamento de Física pelo ensino de qualidade.

Aos colegas da pós-graduação em Biofísica Molecular do Departamento de Física, em especial aos alunos do grupo: Antônio, André, Débora, Tiago e Vinícius.

Aos funcionários do Departamento de Física e da seção de Pós-Graduação, em especial a Rosemar Rosa de Carvalho Brena.

À minha companheira Isadora Pfeifer Dalla Picola pelo carinho, amor, compreensão e encorajamento em todos os momentos.

Aos meus pais José de Oliveira e Vânia Mara de Souza Oliveira, minha irmã Aline Mara de Oliveira e seu companheiro Gustavo Vassoler pela paciência e dedicação por mim.

Às instituições de fomento Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) pelo auxílio financeiro.

Sumário

Lista de Abreviaturas

2D	Duas dimensões
3D	Três dimensões
AMBER	Assisted Model Building with Energy Refinement
CSU	Contact of Structural Units
*Tm*CSP	*Thermotoga maritima* Cold-Shock Protein
CTBP	Center for Theoretical Biological Physics
DNA	Deoxyribose Nucleic Acid
FRET	Förster Resonance Energy Transfer
GROMACS	GROningen MAchine for Chemical Simulations
LD	Lambda Descriptor
MFPT	Mean First Passage Time
NMR	Nuclear Magnetic Resonance
PDB	Protein Data Bank
RG	Radius of Gyration
RMSD	Root Mean Square Deviation
RMN	Ressonância Magnética Nuclear
SMOG	Structure-Based Models in GROMACS
VMD	Visual Molecular Dynamics
WHAM	Weighted Histogram Analysis Method

Lista de Símbolos

$C(Q_0, \Delta)$	Função correlação de Q
C_α	Carbono alfa
$D(Q, T)$	Coeficiente de difusão dependente da coordenada de reação e da temperatura
D_0	Coeficiente de difusão equivalente a uma superfície de energia lisa
$F(Q, T)$	Energia livre dependente da coordenada de reação e da temperatura
$F(\phi)$	Função de diedro
k_B	Constante de *Boltzmann*
K_Q	Constante do potencial harmônico de restrição
$n(E)$	Densidade de estados na energia E
$P(Q, t)$	Probabilidade dependente da coordenada de reação e do tempo
Q	Coordenada de reação, número de contatos corretos nativos
r	Distância entre dois átomos
R	Razão entre os termos de interação diedral
R_g	Raio de giração
S_0	Entropia numa superfície lisa
$S(Q)$	Entropia configuracional em Q
T	Temperatura do sistema
T_f	Temperatura de transição para o estado enovelado
T_g	Temperatura de transição para estado de vidro
V	Potencial de interação

β	Inverso da temperatura multiplicado pela constante de *Boltzmann*
$\Delta E^2(Q)$	Flutuação quadrática média da energia
δE	gap, diferença entre a energia do estado fundamental e a energia média
ϵ	Constante de energia de interação
Γ	Configuração de uma determinada estrutura
σ	Raio de volume de exclusão
ω_{unf}	Curvatura em Q_{unf}
ω_{fold}	Curvatura do topo da barreira de energia
Ω	Número de estados do sistema
τ_f	Tempo de enovelamento
τ_{corr}	Tempo de relaxação associado com o decaimento da função correlação

Lista de Figuras

Resumo

A difusão desempenha um papel importante na cinética de enovelamento de proteínas. Nessa tese, desenvolvemos métodos analíticos e computacionais para o estudo do coeficiente de difusão dependente da posição e do tempo. Para estes estudos, utilizou-se sobretudo o modelo baseado na estrutura (modelo Gō) via simulação computacional da representação em carbonos alfa. Investigou-se o efeito da difusão no enovelamento da proteína *cold-shock* (*Tm*CSP). Encontrou-se que o efeito temporal da difusão leva a cinéticas não-exponenciais e a estatística não-poissônica da distribuição de tempos de enovelamento. Com relação a dependência com a posição, o coeficiente de difusão revelou ter um comportamento não-monotônico que foi compreendido pela análise dos valores-ϕ e da entropia residual no estado nativo. Para uma versão frustrada do modelo, encontrou-se que um baixo nível de frustração energética aumenta a difusão no estado nativo e torna o estado de transição mais homogêneo. Esses resultados corroboram com experimentos recentes de fluorescência de uma única molécula. Esse trabalho também propõe um método para a determinação da superfície de energia de enovelamento de proteína. A partir da caracterização da superfície de energia, definimos a quantidade Λ (LD – *Landscape Descriptor*) que mostrou uma forte correlação entre a cinética e a termodinâmica de uma dezena de proteínas globulares, tornando-se um método útil para classificar proteínas.

Palavras-chave: enovelamento de proteína, coeficiente de difusão, superfície de energia, modelo baseado em estrutura, dinâmica molecular.

Abstract

Diffusion plays an important role in protein folding kinetics. In this thesis we developed analytical and computational methods in order to study the diffusion coefficient dependent on position and time. For these studies we used mainly the structure-based model (Gō model) via computer simulation of the alpha-carbon representation. We investigated the effect of diffusion in the folding of the cold-shock protein (TmCSP). We found that the time dependence on diffusion leads to non-exponential kinetics and non-Poisson statistics of folding time distribution. With respect to the position dependence, the diffusion coefficient reveled a non-monotonic behavior that was understood by analyzing the ϕ-values and the residual entropy in the native state. For a frustrated version of the model, we found that a low level of frustration energy stabilizes and increases the diffusion in the native state and the transition state becomes more homogeneous. These results are supported by recent single-molecule fluorescence experiments. This work also proposes a method to determine the protein folding energy landscape. With the energy landscape characterized, we defined the quantity Λ (LD – Landscape Descriptor) which showed a strong correlation between kinetics and thermodynamics of a dozen globular proteins making it a useful method to classify proteins.

Keywords: protein folding, diffusion coefficient, energy landscape, structure-based model, molecular dynamics.

Organização da Tese

Essa tese começa por discutir, no Capítulo 1, o problema do enovelamento de proteína desde os experimentos bioquímicos de Anfinsen, que motivaram os estudos nessa área, até a mais recente teoria de superfície de energia. Mostra-se que a reação de enovelamento de proteína pode ser tratada a partir de uma abordagem difusiva, para a qual se emprega uma equação de difusão do tipo Fokker-Planck para descrever a dinâmica sobre a superfície de energia e suas quantidades comumente estudadas. Ao final do capítulo, apresentam-se as motivações e os objetivos desse trabalho. O Capítulo 2 expõe a necessidade de se empregar modelos minimalistas no estudo computacional de macromoléculas biológicas, em especial, as proteínas. Nesse contexto, são expostos os modelos baseados na estrutura e as respectivas representações que foram utilizadas na maior parte das simulações desta tese. Propõe-se o método para o cálculo do coeficiente de difusão dependente da coordenada de reação e do tempo de enovelamento. O Capítulo 3 apresenta as discussões dos resultados obtidos durante o trabalho de doutorado e publicados em dois artigos que se encontram anexos nos Apêndices A e B. O Capítulo 3 ainda discute os resultados do artigo, em preparação, anexo no Apêndice C. Finalmente, as conclusões desta tese envolvendo o processo de enovelamento de proteína, o coeficiente de difusão e a superfície de energia se encontram no Capítulo 4. O artigo do Apêndice A discute a dependência não-monotônica do coeficiente de difusão em função da posição. O artigo do Apêndice B mostra os resultados da dependência temporal e configuracional do coeficiente de difusão. A caracterização da densidade de estados proteicos é feita no Apêndice C. Espera-se que as técnicas desenvolvidas neste trabalho sejam úteis na compreensão de experimentos de uma única molécula e auxiliem novos ensaios experimentais.

Capítulo 1

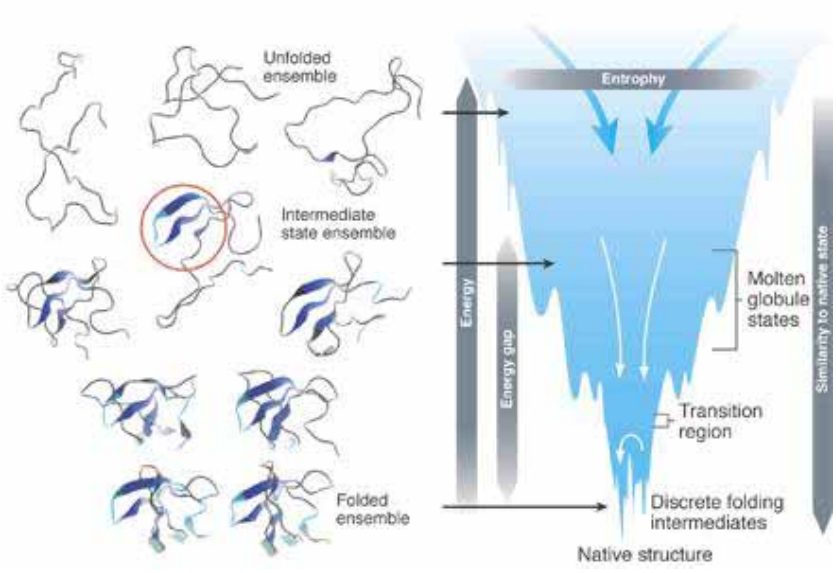
Introdução

1.1 O enovelamento de proteína

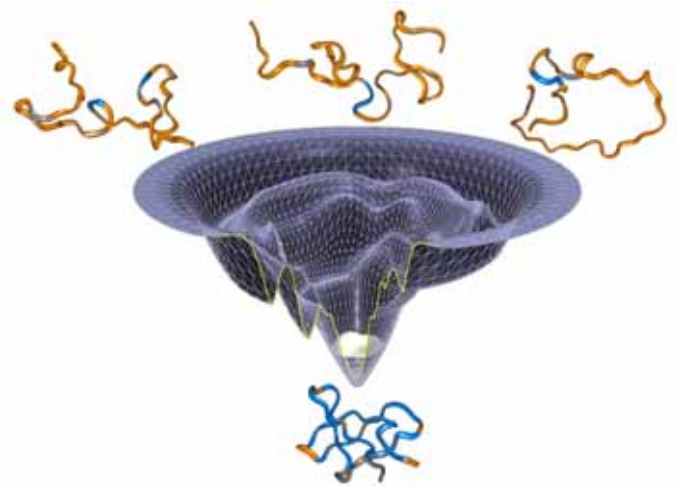
A compreensão dos mecanismos biomoleculares tem proporcionado um grande desafio intelectual. As proteínas pertencem a uma das mais importantes classes biomoleculares pelo fato de desempenhar funções vitais nos organismos. Compreender o mecanismo que permite uma proteína assumir sua forma funcional vem intrigando cientistas durante décadas. Ao que tudo indica, o estudo das proteínas continuará incentivando e gerando discussões entre pesquisadores de diversas áreas. O avanço das técnicas experimentais, sobretudo dos supercomputadores, e as recentes teorias parecem lançar uma nova luz sobre esse antigo problema, o enovelamento de proteína.

Na década de 60, Anfinsen, por meio de experimentos físico-químicos, afirmou que, conhecendo a sequência de aminoácidos, é possível determinar o estado compacto nativo de uma proteína, sendo esse, o mínimo global da energia livre de Gibbs [3, 4]. No final da mesma década, Levinthal propôs a seguinte questão: quantos microestados[1] uma proteína teria que vasculhar para alcançar seu estado de mínima energia [5]? Sabe-se que o número de estados de uma proteína, com aproximadamente 100 aminoácidos, é de ordem

[1]Maneiras diferentes de se arranjar uma cadeia polipeptídica no espaço.



(a)



(b)

Figura 1.1: a) Funil de estruturas de uma proteína proposto por Leopold [1] e um esboço qualitativo de sua superfície de energia (*energy landscape*). O eixo horizontal representa a entropia configuracional e o eixo vertical representa a energia total ou o parâmetro de ordem grau de similaridade de uma conformação em relação ao seu estado nativo. Também estão representados os estados intermediários que a proteína acessa a partir de seu estado desenovelado a caminho de seu estado nativo. As setas representam os múltiplos caminhos, rotas cinéticas, que a proteína pode percorrer durante a reação de enovelamento. b) Representação esquemática multidimensional da superfície de energia com a energia no eixo vertical e o espaço conformacional nos outros. *Adaptado de* [2].

astronômica. No entanto, a proteína converge para um estado particular nativo na escala de tempo geralmente de milissegundos.

Esse paradoxo proposto por Levinthal foi resolvido com os conceitos introduzidos pela recente teoria de superfície de energia[2] [1, 6–10] (Figura 1.1). Essa teoria propõe a existência de uma seleção natural de sequências que possuem uma superfície energética afunilada e rugosa, na qual a rugosidade é relativamente pequena se comparada à profundidade do mínimo global. Essa rugosidade se manifesta como armadilhas locais para o enovelamento da proteína. A reação de enovelamento deve acontecer a tempo da proteína exercer sua função biológica. Para isso, a inclinação do funil de energia deve ser grande o suficiente para vencer os mínimos locais. A teoria de superfície de energia se mostrou uma ferramenta fundamental para compreender qualitativamente e quantitativamente os resultados teóricos e experimentais envolvendo o enovelamento de proteína [2, 11–16].

1.2 O enovelamento como processo difusivo

O enovelamento pode ser descrito, estatisticamente, como a evolução progressiva de *ensembles* de estruturas parcialmente enoveladas, por meio do qual a proteína se enovela a caminho da sua conformação nativa pelo funil de energia [10]. Com isso, o gradiente de energia determina a média temporal com que a proteína percorre o funil até o seu estado enovelado. Isso é determinado por movimentos estocásticos cuja estatística depende dos saltos pelos mínimos locais. Numa primeira aproximação, esses movimentos podem ser considerados como uma difusão pelos ensembles como foi proposto por Bryngelson e Wolynes [6, 17]. Ao propor que a coordenada de reação capture as propriedades básicas de enovelamento e que a coordenada se mova difusamente pela superfície de energia, pode-se associar a equação de Fokker-Planck (relacionada aos movimentos estocásticos) para descrever a dinâmica por [18]

$$\frac{\partial P(Q,t)}{\partial t} = \frac{\partial}{\partial Q}\left\{D(Q)\left[\frac{\partial P(Q,t)}{\partial Q} + P(Q,t)\frac{\partial \beta F(Q)}{\partial Q}\right]\right\} \tag{1.1}$$

[2] Do inglês *energy landscape.*

sendo Q uma coordenada de reação arbitrária, $D(Q)$ o coeficiente de difusão configuracional local, $F(Q)$ a energia livre e β o inverso da temperatura (T) do sistema. A função $P(Q,t)$ é a densidade de probabilidade de uma população de estruturas estar em uma configuração (Q) em um determinado tempo (t).

Dessa forma, a multidimensionalidade da superfície de energia, que surge devido ao grande número de interações entre os aminoácidos da proteína e da proteína com o meio biológico, pode ser reduzida à potenciais efetivos como a energia livre em função de um único parâmetro de ordem que descreve a reação de enovelamento.

O coeficiente de difusão, presente na equação 1.1, depende das armadilhas dos mínimos locais que refletem a rugosidade da superfície de energia [19]. Por isso, as taxas de enovelamento estão intimamente relacionadas com D e com a altura da barreira energética global em F [20]. O tempo de enovelamento pode ser escrito como uma integral dupla [6]

$$\tau_f = \int_{Q_{unf}}^{Q_{fold}} dQ \int_0^Q dQ' \frac{e^{\beta\{F(Q)-F(Q')\}}}{D(Q)} \tag{1.2}$$

Para uma função de energia livre $F(Q)$ com uma barreira energética bem definida (Figura 1.2), em temperaturas próximas da temperatura de transição, Q_{unf} corresponde ao vale de $F(Q)$ onde a cadeia está desenovelada e Q_{fold} é o ponto da coordenada de reação no outro vale onde a cadeia está enovelada. A integral dupla, equação 1.2, pode ser aproximada pela lei de Kramers [18] por

$$\tau_f = \left(\frac{2\pi}{\beta}\right)^{\frac{1}{2}} \frac{1}{D_0 \omega_{unf} \bar{\omega}_{fold}} e^{\beta\{\bar{F}(Q_t)-F(Q_{unf})\}} \tag{1.3}$$

em que

$$\bar{F}(Q) = F(Q) - T\log\left[\frac{D(Q)}{D_0}\right] \tag{1.4}$$

com ω_{unf} sendo a curvatura em Q_{unf} e ω_{fold} a curvatura do topo da barreira de energia livre localizada em Q_t. D_0 é o coeficiente de difusão efetivo equivalente a uma superfície de energia lisa.

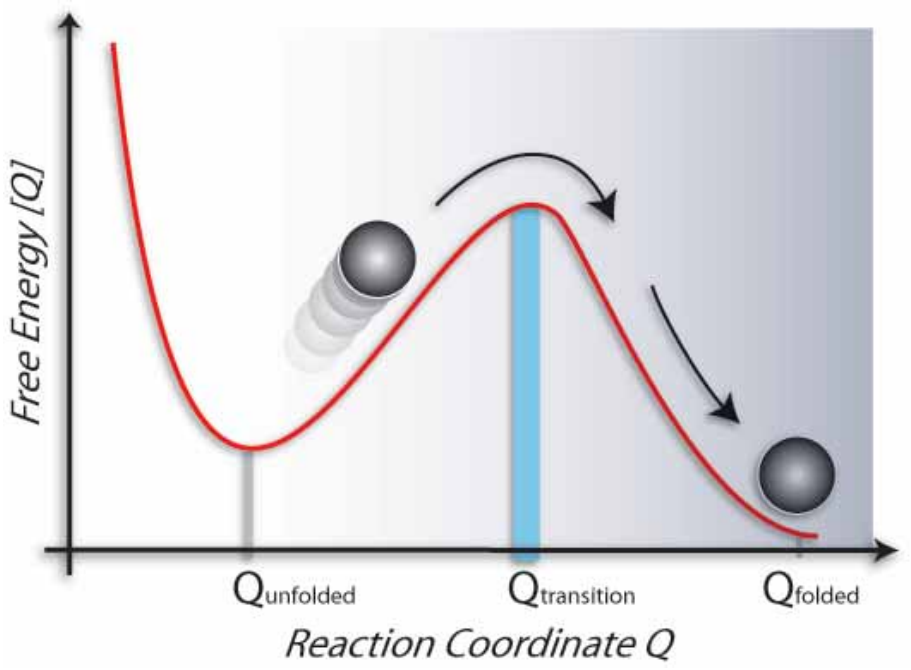


Figura 1.2: Esboço da energia livre ($F(Q)$) em função da coordenada de reação arbitrária Q próximo à temperatura de enovelamento T_f. A linha vertical em $Q_{transition}$ marca a região de barreira de energia que corresponde à transição do estado desenovelado (Q_{unf}) para o estado enovelado (Q_{fold}). Nesse potencial efetivo, a distribuição de probabilidades de uma determinada configuração no tempo ($P(Q,t)$) difunde-se pela barreira de energia livre por flutuações térmicas.

O coeficiente de difusão depende da superfície de energia e dos movimentos locais permitidos para a proteína. Bryngelson e Wolynes [6] analisaram a difusão configuracional assumindo o limite de uma superfície de energia rugosa descorrelacionada e as regras de Metropolis [21]. O coeficiente de difusão dependerá também da temperatura em que a proteína deverá escapar dos mínimos locais da superfície de energia. Para baixas temperaturas, T menor que T_g, o coeficiente de difusão é dado por

$$D(T,Q) = D_0 \exp[-S_0] \tag{1.5}$$

com S_0 a entropia numa superfície lisa. Para altas temperaturas, T maior que $2T_g(Q)$, D segue a lei de Ferry típica de vidros [6]

$$D(T,Q) = D_0 \exp[-\beta^2 \Delta E^2(Q)] \tag{1.6}$$

em que $\Delta E^2(Q)$ é a flutuação quadrática média da energia e $\beta = 1/T$. Em temperaturas intermediárias, T entre $T_g(Q)$ e $2T_g(Q)$, D pode ser aproximado por

$$D(T,Q) = D_0 \exp\left\{-S^*(Q) + [\beta_g(Q) - \beta]^2 \Delta E^2(Q)\right\} \tag{1.7}$$

em que $\beta_g = 1/T_g$.

A temperatura de transição para estado de vidro, dada por T_g, reflete a competição entre a rugosidade e a entropia na superfície de energia. A temperatura de transição de vidro, no equilíbrio, será dada por

$$T_g(Q) = \sqrt{\frac{\Delta E^2(Q)}{2S^*(Q)}} \tag{1.8}$$

em que $S^*(Q)$ é a entropia configuracional em Q. Em T_g, o coeficiente de difusão diminui com o número total de estados, sugerindo a possibilidade do paradoxo de Levinthal [5, 22]. Abaixo de T_g não se pode considerar a energia como a média dos ensembles, pois a flutuação na energia livre é considerável e dependerá fortemente do modelo e da seqüência do heteropolímero [19]. Sendo assim, o lento processo de enovelamento pode ser

melhor explicado por poucos caminhos cinéticos predominantes do que pelas leis de médias estatísticas para eventos rápidos. A dependência de D com a temperatura, da competição da entropia com a energia, leva a uma curva parabólica para os tempos de enovelamento em função de β. Leite [23, 24] mostrou que para uma superfície de energia de um sistema de polarização de solvente, as flutuações dominam gradualmente a cinética abaixo de T_g, análogo à teoria de Bryngelson e Wolynes para a superfície de energia de uma proteína.

Dessa forma, os tempos de enovelamento dependem da energia livre e do coeficiente de difusão que incorporam as múltiplas difusões através da barreira de energia. A escolha de uma coordenada de reação também poderá influenciar na dependência do coeficiente de difusão e na forma da energia livre de uma proteína [19].

1.3 Motivação

O coeficiente de difusão D faz parte da representação unidimensional por processo difusivo da teoria de superfície de energia para o enovelamento de proteína. A superfície de energia é multidimensional, porém a difusão pelo funil de energia pode ser investigada por um único parâmetro que capture as características básicas do enovelamento. Recentemente, diversos grupos de pesquisa teórica [25–32] e experimental [16, 20, 33–37] têm investigado o coeficiente de difusão e as taxas de enovelamento. As técnicas experimentais tiveram um grande avanço tecnológico e podem ser utilizadas no intuito de se obter D em função de uma coordenada de reação experimental bem como informações sobre rotas de enovelamento pela superfície de energia [16]. Sendo assim, existe um interesse em caracterizar D e, como consequência, os mecanismos que regem o enovelamento de proteína.

No trabalho de mestrado [29], desenvolvemos o método para calcular o coeficiente de difusão dependente da coordenada para o modelo computacional simples de rede cúbica [2, 38]. Encontramos que essa dependência modifica o estado de transição deslocando-o para a direita e para cima em alguns k_BT de energia, contribuição difusiva no enovelamento de proteína. O trabalho de mestrado motiva a continuação do estudo de D por modelos

no espaço contínuo que permitam uma relação mais próxima com técnicas experimentais recentemente desenvolvidas.

A relação direta de D com a teoria de superfície de energia, que aplicamos no estudo do enovelamento de proteína, nos motiva também a estudar esta superfície. Existe uma preocupação da comunidade experimental de como quantificar a superfície do funil de energia e como relacioná-lo com as medidas de cinética e de estabilidade termodinâmica. Os artigos científicos geralmente utilizam uma representação em diagrama da superfície de energia em forma de funil [39–43]. Por esse motivo, é comum calcular uma superfície de um potencial efetivo de fácil obtenção e que descreve a reação de enovelamento de forma satisfatória. Sendo assim, torna-se necessário o desenvolvimento de uma técnica para determinar quantitativamente a superfície de energia "real" de uma proteína e um parâmetro que correlacione as grandezas cinéticas com as termodinâmicas do enovelamento de proteína.

Capítulo 2

Modelo

Sabe-se que a descrição completa e realística de todas as interações de qualquer sistema físico proteico é praticamente impossível computacionalmente. Isso porque deve-se levar em conta as interações entre os resíduos de aminoácidos da cadeia peptídica e dela com o solvente, entre outras interações que possam existir. No entanto, pode-se abrir mão da descrição realística e utilizar modelos simples ou minimalistas sem perder as propriedades do sistema ou a capacidade da proteína de se enovelar. Para isso, é necessário construir modelos que descrevam as propriedades do enovelamento de interesse. Tem se tornado frequente o uso de modelos simples para o estudo computacional de sistemas complexos já que, com eles, pode-se executar simulações em um tempo relativamente curto e mensurável [44].

2.1 Modelo baseado na estrutura

A simulação computacional de modelos minimalistas tem contribuído para o entendimento do enovelamento de proteínas [8, 45–51], dimerização [52, 53], mudanças conformacionais funcionais e reações enzimáticas [14, 54–56], entre outros sistemas biomoleculares. Isso porque, a simulação de modelos mais realísticos de proteínas envolve

uma infinidade de parâmetros e de interações atômicas e, como consequência, o tempo computacional necessário para se extrair informações relevantes aumenta dramaticamente. Os modelos minimalistas, por utilizarem uma representação simplificada do problema, permitem explorar uma grande faixa de parâmetros dos sistemas em razoável tempo computacional.

No caso dos modelos minimalistas baseados na estrutura, a primeira simplificação se refere ao potencial de interação entre os componentes do sistema. A segunda se refere aos próprios componentes do sistema: todos os átomos ou somente carbonos alfas (C_α), detalhados a seguir. Aplica-se o modelo com todos os átomos quando as informações sobre as cadeias laterais são importantes para o entendimento do problema em estudo, como por exemplo, empacotamento e interação entre os resíduos de aminoácidos. Caso contrário, aplica-se o modelo C_α pois, além de ser o modelo melhor testado e utilizado pela comunidade científica, permite uma economia de tempo computacional de até 10 vezes se comparado ao modelo com todos os átomos.

O potencial do modelo baseado na estrutura é construído a partir da conformação nativa do monômero ou do dímero, ou seja, os parâmetros da expressão do potencial são obtidos por meio da estrutura nativa. O mínimo de energia corresponde à conformação da estrutura nativa da proteína depositada no Protein Data Bank (PDB) obtidas por técnicas experimentais como cristalografia de raio-X e ressonância magnética nuclear (RMN). Os modelos cujos potenciais são construídos a partir da estrutura nativa são conhecidos na literatura como modelos G$\bar{o}$ [57]. O potencial de uma determinada configuração Γ de uma proteína, tendo Γ_0 como sua configuração no estado nativo, será dado pela expressão

$$\begin{aligned} V(\Gamma,\Gamma_0) &= \sum_{bonds} \epsilon_r (r-r_o)^2 \\ &+ \sum_{angles} \epsilon_\theta (\theta-\theta_o)^2 \\ &+ \sum_{impropers/planar} \epsilon_\chi (\chi-\chi_o)^2 \\ &+ \sum_{backbone} \epsilon_{BB} F_D(\phi) \end{aligned}$$

$$
\begin{aligned}
&+ \sum_{sidechains} \epsilon_{SC} F_D(\phi) \\
&+ \sum_{contacts} \epsilon_C \left[\left(\frac{\sigma_{ij}}{r} \right)^{12} - 2 \left(\frac{\sigma_{ij}}{r} \right)^{6} \right] \\
&+ \sum_{non-contacts} \epsilon_{NN} \left(\frac{\sigma_{NN}}{r} \right)^{12}
\end{aligned} \tag{2.1}
$$

com

$$F(\phi) = [1 - \cos(\phi - \phi_o)] + \frac{1}{2}[1 - \cos(3(\phi - \phi_o))] \tag{2.2}$$

e $\epsilon_r = 100$, $\epsilon_\theta = 20$, $\epsilon_\chi = 40$ e $\epsilon_{NN} = 0.01$ em unidades de ϵ_c, a energia de interação entre contatos (aproximadamente 1 kcal/mol por ser um modelo reduzido). Nessa equação, r_0 representa a distância nativa entre dois átomos diretamente ligados entre si, θ_0 é o ângulo entre três átomos consecutivos da estrutura nativa e, analogamente, ϕ_0 é o ângulo diedral entre quatro átomos. O parâmetro σ_{ij} do termo de van der Waals (penúltima linha da equação 2.1) é determinado através do mapa de contatos entre os átomos na estrutura nativa da proteína. O mapa de contatos é obtido utilizando o algoritmo *Contact of Structural Units* (CSU) [58]. O parâmetro σ_{NN} está presente num termo repulsivo e serve para manter uma distância de máxima aproximação entre os átomos. Nesse caso, essa distância é maior que $\sigma_{NN} = 2.5$ Å. Esse número caracteriza o volume ocupado por um átomo. O arquivos de entrada para a simulação podem ser facilmente obtidos utilizando a interface gráfica do servidor SMOG@ctbp [51].

Com o objetivo de adicionar frustração energética não-específica aos contatos não-nativos, acrescentamos um potencial atrativo da forma de um poço gaussiano (Figura 2.1) dado pela expressão

$$V_f(r) = - \sum_{non-contacts} \epsilon_{NC} \exp \left\{ - \frac{(r - r_g)^2}{\sigma_g^2} \right\} \tag{2.3}$$

no qual r_g é a distância média dos contatos nativos em Å e $\sigma_g = 1.0$ Å. ϵ_{NC} determina o grau de frustração energética e utilizou-se, nas simulações desse trabalho, valores para ϵ_{NC} entre 0.1 e 0.7.

Como os parâmetros são obtidos a partir da estrutura depositada no PDB, o mínimo da energia potencial ocorrerá exatamente nessa estrutura. Portanto, ao iniciar

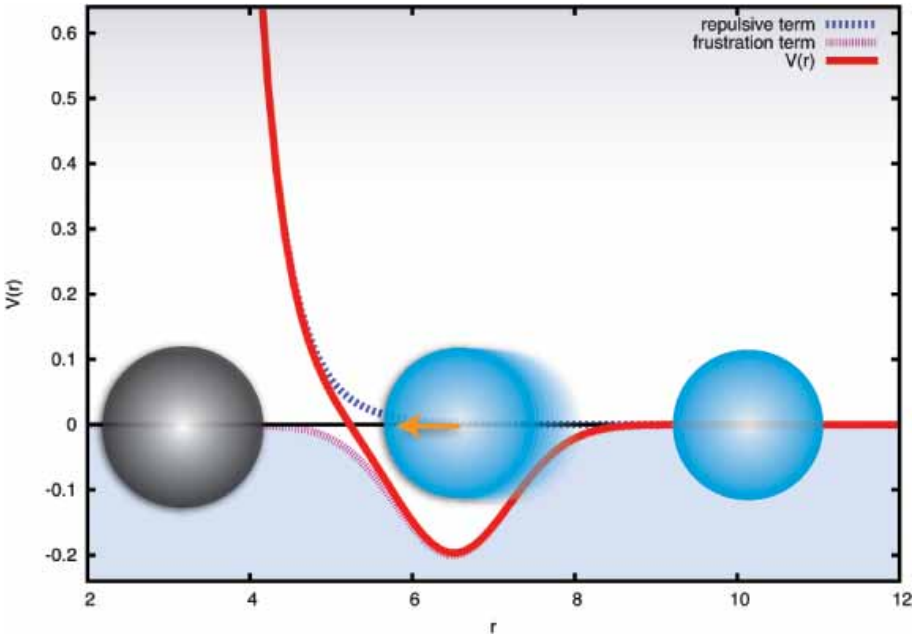


Figura 2.1: Potencial de frustração energética não-específica entre contatos não-nativos. A curva tracejada azul representa o termo de repulsão (última linha da equação 2.1) que impede a superposição entre átomos e entrelaçamento da cadeia na simulação. Nesse caso, empregamos $\sigma_{NN} = 4.0$ Å, raio comumente utilizado no modelo C_α. A curva tracejada vermelha representa o termo de atração (frustração) entre pares de átomos que não estão em contato na estrutura nativa (equação 2.3) e distantes pelo menos por quatro resíduos na cadeia. Este exemplo aplica $r_g = 6.5$ Å, média dos contatos nativos no caso da proteína *Tm*CSP. As esferas estão em tamanho reduzido para uma melhor visualização. A curva contínua vermelha representa a soma dos termos de repulsão e de atração. O termo de frustração cria uma pequena perturbação no potencial favorecendo levemente a interação entre esses pares somente quando muito próximos ($r \approx r_g$), já que a largura do poço (σ_g) restringe-se a aproximadamente 2 Å. A versão mais recente do servidor SMOG@ctbp [51] implementa o potencial gaussiano como o termo repulsivo da equação 2.1.

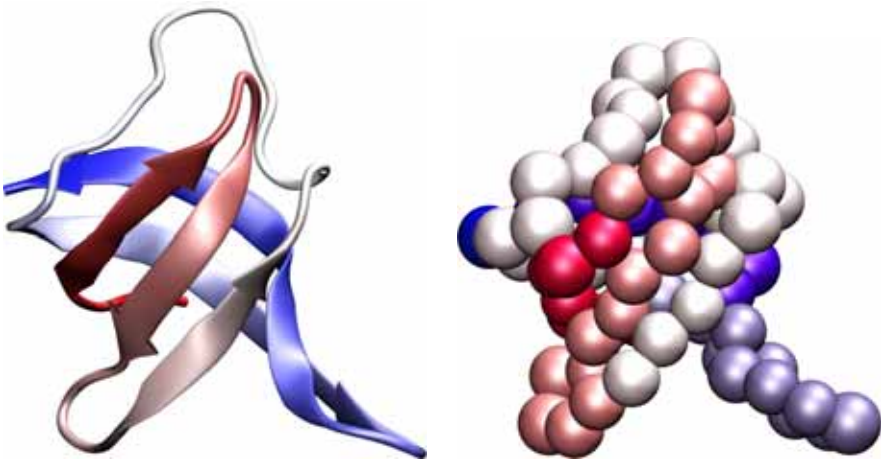
uma simulação com a cadeia completamente aberta, no final se obtém a estrutura nativa enovelada. No caso da mudança conformacional, pode-se incluir no potencial informações das estruturas aberta e fechada e a simulação terá dois mínimos para visitar. Isso permite estudar os mecanismos que governam o enovelamento ou uma mudança conformacional induzida por um ligante e, particularmente, o aspecto do estado de transição desses processos.

2.1.1 Modelo C_α

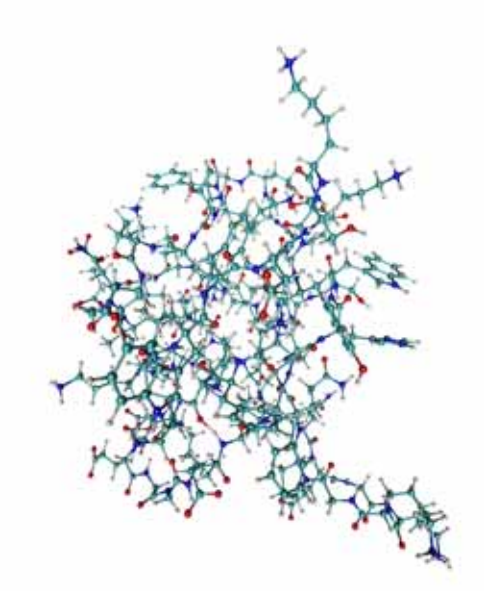
A simulação computacional baseada na dinâmica molecular do modelo C_α [49] utiliza uma simplificação fundamental. Todos os átomos de um aminoácido da cadeia polipeptídica são substituídos por uma esfera centrada na posição do carbono alfa (C_α) correspondente, mantendo seu raio de volume de exclusão. Dessa forma, apenas o C_α da cadeia principal é representado explicitamente e interagem entre si pelo potencial dado pela equação 2.1 com uma pequena adaptação. No modelo C_α, existem somente os termos de ligação covalente, angular e diedral para manter a geometria da cadeia principal. Os ângulos diedrais são formados entre quatro átomos C_α adjacentes e os contatos não-locais interagem via potencial com potências 12-10 ao invés das potências 12-6 do termo de Lennard-Jones típico do modelo com todos os átomos.

2.1.2 Modelo com todos os átomos

Ao aprimorar a resolução do sistema tem-se o modelo de proteína baseado na estrutura com todos os átomos [50], o qual todos os átomos pesados (não-hidrogênio) são incluídos. Cada átomo é representado por uma única esfera de massa unitária. Na equação 2.1, os termos de ligação covalente, angular, diedral impróprio e diedral planar são mantidos por potenciais harmônicos. Interações nativas não-locais, pares de átomos não-ligados covalentemente, que estão em contato no estado nativo entre resíduos i e j, com $i < j + 3$, são do tipo Lennard-Jones. Entretanto, todas as outras interações



(a) (b)



(c)

Figura 2.2: Proteína *cold-shock* da bactéria *Thermotoga maritima* (*Tm*CSP com o código 1G6P no Protein Data Bank [34]) a) Representação das estruturas secundárias da proteína. b) Modelo C_α de proteína com o tamanho das esferas correspondendo ao raio de volume de exclusão utilizado no modelo. As estruturas em a) e b) são coloridas por uma escala de vermelho (C-terminal) à azul (N-terminal). c) Modelo com todos os átomos de proteína com os carbonos (em verde), os oxigênios (em vermelho) e os nitrogênios (em azul). Os hidrogênios (em branco) da *Tm*CSP também são mostrados apesar de não estarem presentes no modelo com todos os átomos. As estruturas foram geradas com o programa VMD [59]. A *Tm*CSP é uma proteína globular de 66 aminoácidos do tipo barril-β, uma folha-β de 3 fitas anti-paralelas em contato com outra folha-β de 2 fitas também anti-paralelas ligadas por uma alça.

não-locais são consideradas repulsivas. Com relação às constantes diedrais, primeiramente são agrupados os ângulos diedrais que compartilham o mesmo átomo. Numa cadeia polipeptídica, podem ser definidos até quatro ângulos diedrais compartilhando a mesma ligação covalente $C - C_\alpha$ como ligação central. A cada diedro é atribuída uma energia de interação $1/N_D$, com N_D sendo o número de ângulos diedrais do grupo. ϵ_{BB} e ϵ_{SC} são tais que $R_{BB/SC} = \epsilon_{BB}/\epsilon_{SC} = 2$. Dessa forma, as constantes de interações diedrais e de contatos são escalonadas de forma que a razão entre a energia total dos contatos com relação à energia total diedral satisfaça $R_{C/D} = \sum \epsilon_C / (\sum \epsilon_{BB} + \sum \epsilon_{SC}) = 2$.

2.2 Simulação do coeficiente de difusão

A dinâmica molecular para o cálculo do coeficiente de difusão ($D(Q)$) é realizada restringindo as simulações para aumentar a amostragem ao redor de um ponto específico da coordenada Q. Para isso, é adicionado ao potencial original (equação 2.1) um termo harmônico de restrição similar ao usado na bem conhecida técnica amostragem de "guarda-chuva"[1] [60, 61]. O termo harmônico de restrição é dado por

$$V_{bias}(Q^*) = K_Q(Q - Q^*)^2 \tag{2.4}$$

com K_Q sendo a constante de mola de restrição (em unidades de k_BT) e Q^* é a região de interesse para a amostragem. Para cada Q^* fixo, D pode ser obtido como função de Q para investigar a dependência da coordenada de reação na difusão (Figura 2.3).

Além disso, Q somente pode assumir valores discretos devido ao fato de que contatos são formados ou não, assim qualquer função de Q será descontínua. Para evitar funções descontínuas durante a dinâmica molecular, Q é levemente modificado de modo a produzir o potencial de restrição da equação 2.4 com derivadas definidas. A redefinição de Q assume a forma de uma função do tipo

$$Q = \frac{1}{N_Q} \sum_{contacts} \frac{1}{2} \{1 - \tanh[10(r - 1.2\sigma_{ij})]\} \tag{2.5}$$

[1] Do inglês *umbrella sampling.*

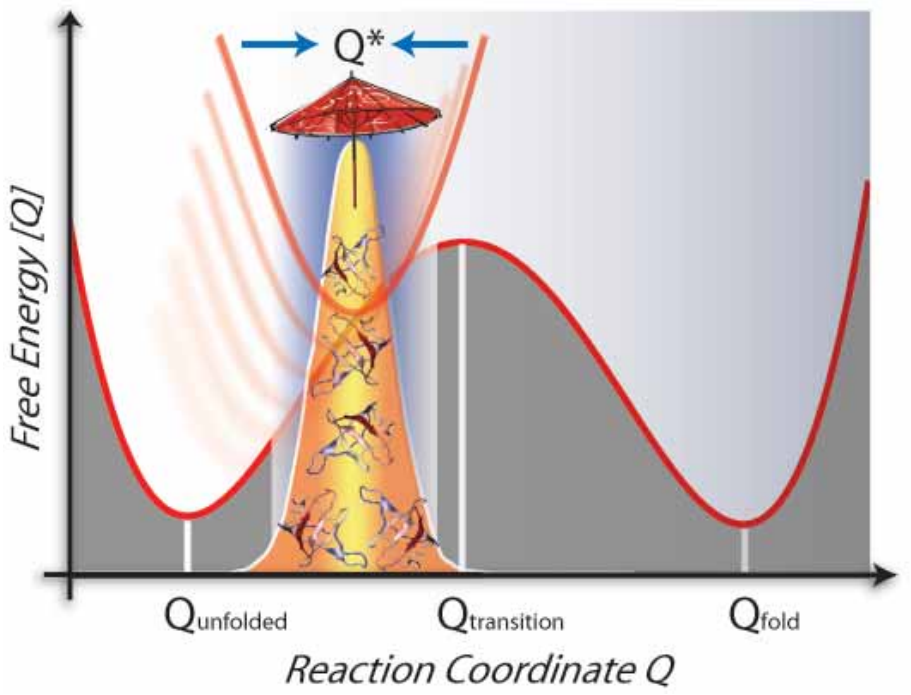


Figura 2.3: Energia livre (poço duplo) e potencial harmônico de restrição (poço único) em função da coordenada de reação Q. O potencial harmônico, dado pela equação 2.4, possui mínimo em $Q = Q^*$ que restringe o espaço conformacional aumentando a amostragem próximo a esse ponto da coordenada de reação e também desloca o máximo do histograma de estruturas acessadas dos mínimos do poço duplo para o mínimo em Q^*. A constante de força do potencial de restrição (K_Q) é escolhida numa faixa intermediária de modo a perturbar levemente a energia livre original (sem K_Q) e por outro lado criar um estado quasi-harmônico nas vizinhanças de Q^*. Assim, obtém-se o coeficiente de difusão ($D(Q^*)$) para um Q particular e, deslocando-se o "guarda-chuva", pode-se calcular o coeficiente de difusão em função de toda a coordenada. As linhas verticais marcam os estados desenovelado ($Q_{unfolded}$), transição ($Q_{transition}$) e nativo (Q_{folded}). Esse perfil de energia livre, com os dois poços no mesmo nível, pode ser obtido na temperatura de enovelamento (T_f).

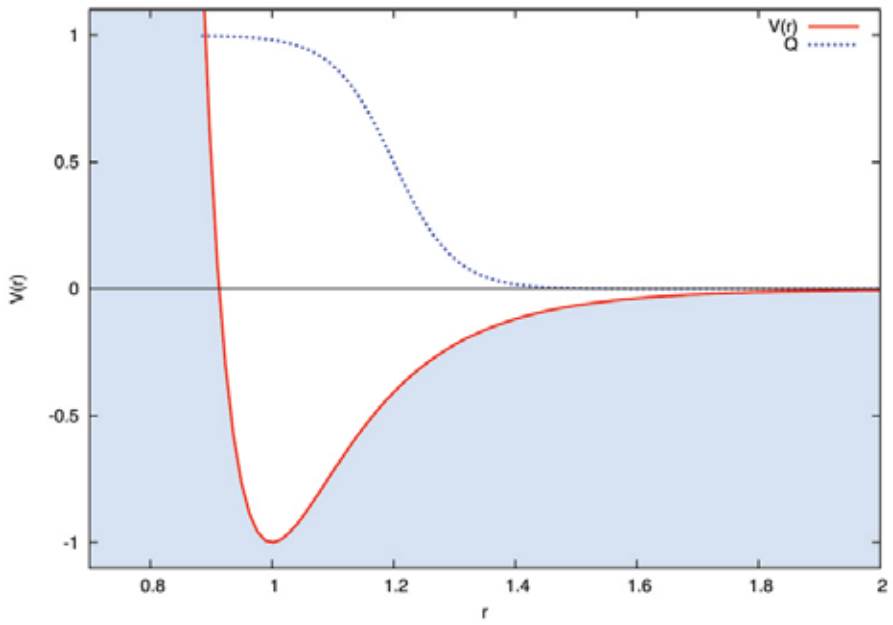


Figura 2.4: Redefinição do cálculo da coordenada de reação Q durante a simulação. As curvas contínua e tracejada representam, respectivamente, o termo de van der Waals (penúltima linha da equação 2.1) e a nova definição da coordenada de reação Q dada por uma função do tipo tanh (equação 2.5), ambos em função da distância r de separação entre pares de C_α. O termo de van der Waals atua sobre os contatos nativos na dinâmica; em seu mínimo, um contato é formado para um par $C_\alpha - C_\alpha$ ($Q \to 1$) e somado ao número de contatos nativos totais num instante t da simulação. Caso contrário, se a distância aumenta ($r \to \infty$), o contato não é formado ($Q \to 0$) o que mantém a noção discreta de contatos nativos sendo formados durante a simulação. Essa função de var der Waals é do tipo com potências 12-10 (ao invés das potências usuais 12-6) empregada em simulações do modelo C_α de proteína.

com σ_{ij} previamente definido como a distância nativa $C_\alpha - C_\alpha$ e N_Q o número total de contatos nativos. Tipicamente, um contato é definido como dois resíduos a uma distância de aproximadamente $1.2\sigma_{ij}$. Sendo assim, a função tanh atua efetivamente como uma função degrau o que não modifica a noção discreta de contatos sendo formados durante a simulação (Figura 2.4).

Os efeitos da dependência da posição no coeficiente de difusão são estudos variando a coordenada de reação pelo processo de enovelamento. Para cada Q^* fixo, o coeficiente de difusão é estimado usando a aproximação difusiva quasi-harmônica [19] dada por

$$D(Q^*) = \frac{\Delta Q^2(Q^*)}{2\tau_{corr}(Q^*)} \tag{2.6}$$

com ΔQ^2 sendo a flutuação média quadrada em Q e τ_{coor} sendo o tempo de relaxação diretamente associado com o decaimento da função correlação de Q

$$C(Q_0, \Delta) = \frac{< Q_0(t)Q_0(t+\Delta) > - < Q_0^2(t) >}{< Q_0^2(t) > - < Q_0(t) >^2} \tag{2.7}$$

Cada simulação, realizada numa temperatura T e Q^* fixos, a coordenada de reação é armazenada como função do tempo ($Q(t)$) para o cálculo das médias em ΔQ^2 e $C(Q_0, \Delta)$.

Além de $Q(t)$, também é possível extrair da trajetória de uma simulação outras coordenadas de reação mais intimamente relacionadas com medidas experimentais como o desvio médio quadrático ($RMSD(t)$) e o raio de giração ($R_g(t)$) como função do tempo (t) e determinar suas relações com a coordenada computacional $Q(t)$. Se o coeficiente de difusão é calculado para uma determinada coordenada, nesse caso Q, é possível obter uma aproximação para o coeficiente de difusão para uma outra coordenada de reação, por exemplo $D(R_g)$, usando a transformação [32]

$$D(R_g) = D(Q) \left(\frac{dR_g}{dQ} \right)^2 \tag{2.8}$$

com $|dR_g/dQ|$ sendo o jacobiano de transformação de variável. A relação $Q(R_g)$, calculada próxima da temperatura de transição de enovelamento (T_f), é uma função monotônica e

pode-se empregar o ajuste de uma função polinomial quadrática para se obter a função inversa $R_g(Q)$ e o respectivo jacobiano.

O perfil de energia livre termodinâmico é calculado combinando simulações executadas em múltiplas temperaturas usando o algoritmo WHAM (Weighted Histogram Analysis Method) [62, 63].

O tempo de enovelamento dado pela integral dupla (equação 1.2) assume uma forma discreta para o cálculo dos tempos dada por

$$\tau_f(T) = \sum_{Q=Q_u}^{Q_f} \sum_{Q'=0}^{Q-1} \frac{e^{\beta\{F[Q]-F[Q']\}}}{D(Q)} \tag{2.9}$$

A média do primeiro tempo de passagem (MFPT[2]), média sobre o tempo necessário para uma proteína passar de uma estrutura aberta para o seu estado enovelado, é comparada com os tempos analíticos dado pela equação 2.9, gráfico em "U" quando em função da temperatura.

[2]Do inglês *Mean First Passage Time*.

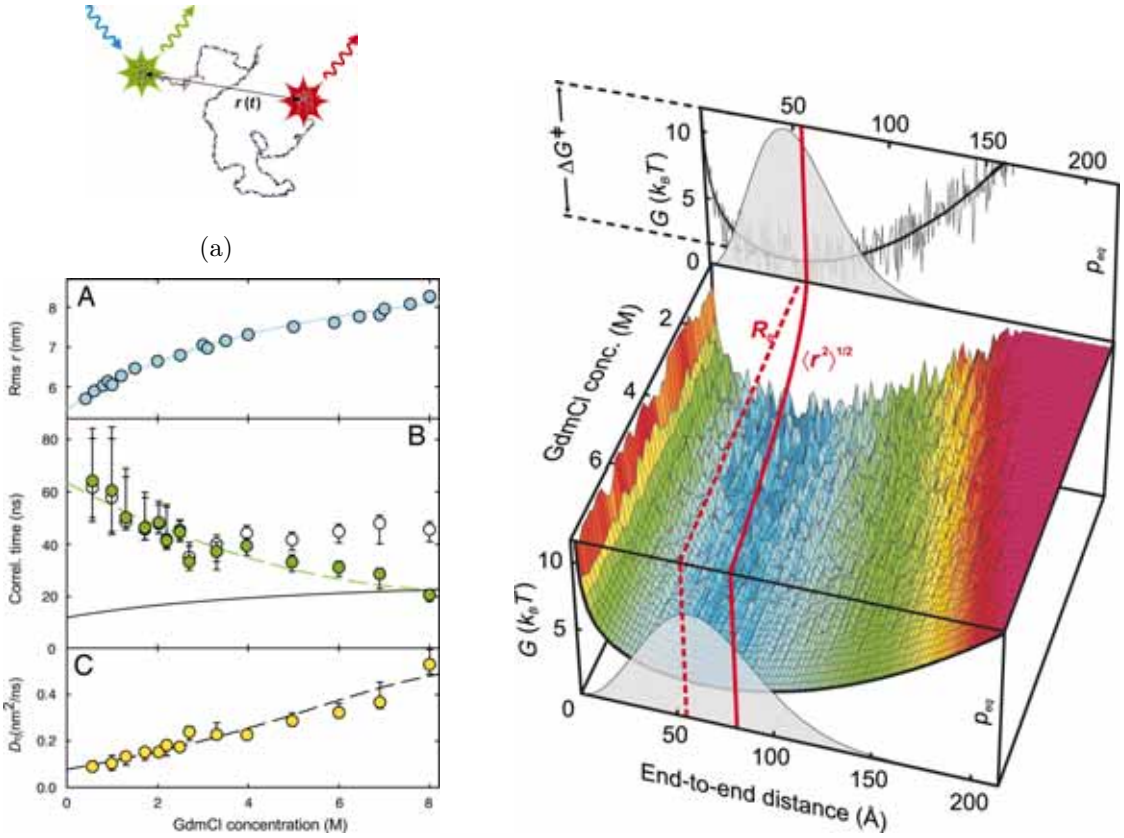
Capítulo 3

Resultados

No trabalho realizado durante o doutorado, estudou-se o efeito do coeficiente de difusão no problema de enovelamento de proteína [30, 31] que teve como motivação inicial o trabalho de mestrado [29]. No trabalho de mestrado empregou-se um modelo de rede cúbica e verificou-se que o coeficiente de difusão dependente da posição resulta em cálculos mais precisos dos tempos de enovelamento quando comparados com os cálculos existentes na literatura com o coeficiente de difusão constante durante toda a reação de enovelamento. Também verificou-se que a posição e a altura da barreira de energia livre, localizada no estado de transição, se modifica quando levado em conta a componente difusiva no processo de enovelamento.

Além disso, caracterizou-se a superfície de energia de uma coleção de proteínas para as quais é possível predizer, analítica e computacionalmente, taxas de enovelamento a partir das propriedades particulares de cada funil de energia (*manuscrito em preparação*).

Para isso, os modelos baseados na estrutura nativa com a simplificação de utilizar apenas carbonos alfas e todos os átomos pesados foram aplicados nestes estudos sobre o coeficiente de difusão e a superfície de energia. Os resultados teóricos computacionais estão em comum acordo com os resultados experimentais (Figura 3.1).



(a)

(b)

(c)

Figura 3.1: a) Experimentos de FRET realizados com a proteína TmCSP no qual a distância entre os marcadores de finais de cadeia ($r(t)$) é armazenada. No painel b) em função da concentração de desnaturante, A) flutuação da distância entre final de cadeia $\langle r^2 \rangle$ ($Rms\ r$), B) tempo de autocorrelação dos doares de intensidade em a) (círculos abertos) e tempo de autocorrelação com a correção do efeito da viscosidade no experimento τ_r (círculos verdes) e C) coeficientes de difusão efetivos dos finais de cadeia com a correção do efeito da viscosidade (D_η) usando a equação $D_\eta = \langle r^2 \rangle / 6\tau_r$. As linhas contínuas representam ajustes empíricos. c) Superfície de energia livre esquemática da TmCSP. A superfície apresenta um modelo para a rugosidade, em escala de cor, que está em torno de $1.3 \pm 0.2kT$. O valor teórico/analítico encontrado pela difusão foi $1.7 \pm 0.3kT$. Pode-se verificar que a altura da barreira de energia livre está entre 7 e 10 k_BT (dependendo da concentração de desnaturante) próximo do valor teórico obtido via simulação (7 k_BT). *Adaptado de* [16].

3.1 O coeficiente de difusão dependente da coordenada

Nesta seção, discutem-se os resultados obtidos no trabalho "*The Origin of Nonmonotonic Complex Behavior and the Effects of Nonnative Interactions on the Diffusive Properties of Protein Folding*", anexo no apêndice A, publicados na revista *Biophysical Journal*. Neste trabalho, demonstrou-se um método para se calcular o coeficiente de difusão em função de uma única coordenada ($D(Q)$).

A difusão pode ser estudada no enovelamento de proteína com modelos em que a superfície de energia é reduzida a poucos graus de dimensionalidade [19, 26, 29–32, 64–66]. Simplifica-se consideravelmente o problema ao obter as grandezas físicas e as propriedades cinéticas e termodinâmicas (como energia livre e coeficiente de difusão) em função de um único parâmetro de ordem. De acordo com a equação 1.1, $D(Q)$ é o coeficiente de difusão dependente de um único parâmetro de ordem, a coordenada de reação Q. Q pode ser definido de várias maneiras. Nesse trabalho, Q está relacionado com o grau de similaridade de uma determinada estrutura com a conformação de mais baixa energia (estado nativo). No caso aplicado de dinâmica molecular utilizando modelos computacionais, define-se Q como a fração de contatos nativos presentes em uma determinada estrutura com relação à estrutura nativa proteica.

O modelo baseado na estrutura (modelo Gō, como é conhecido na literatura [57, 67]), com a representação somente dos carbonos alfas da cadeia principal, foi empregado na dinâmica molecular [49]. O potencial, bem como as propriedades desse modelo, foram discutidos detalhadamente no capítulo 2 e na seção *Models and Methods* do artigo anexado no apêndice A. A proteína estudada foi a *cold-shock* da bactéria *Thermotoga maritima* (TmCSP).

Primeiramente, realizou-se uma análise cinética e termodinâmica da TmCSP procurando verificar a influência dos parâmetros do potencial harmônico de restrição (equação 2.4) no sistema. A Figura 3.2 apresenta os perfis de energia livre em função do parâmetro de ordem ($F(Q) \times Q$) para diferentes valores da constante de força do potencial

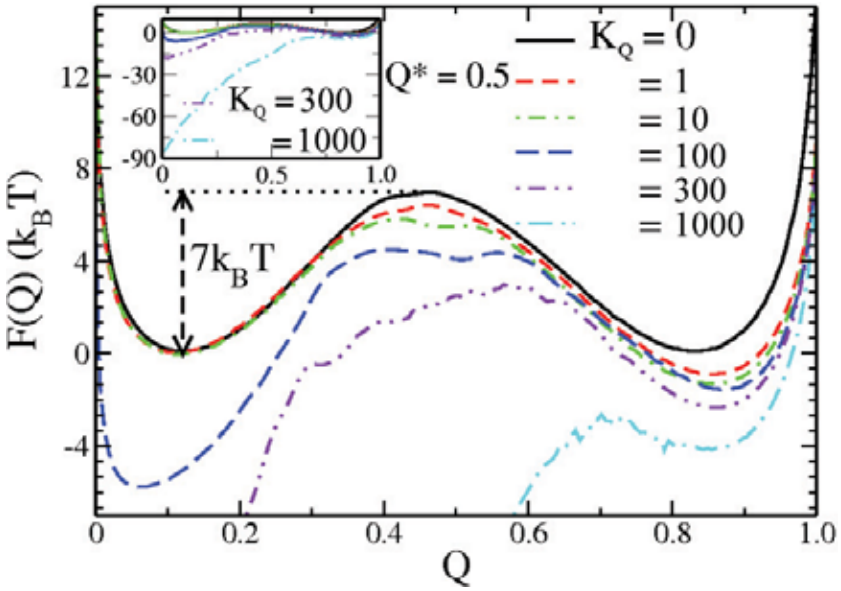


Figura 3.2: Energia livre em função da coordenada de reação Q (fração de contatos nativos formados) para as simulações sem restrição ($K_Q = 0$) e com restrição ($K_Q > 0$) para diferentes valores da constante de força de restrição K_Q. Com o aumento de K_Q, a função energia livre começa a desviar da função original (sem restrição) e alguns estados são privilegiados onde confinamos o sistema próximo de Q^* tornando esse estado quasi-harmônico. Esta figura exemplifica as simulações fixando o termo de restrição no estado de transição ($Q^* = 0.5$) e na temperatura de enovelamento (T_f) do sistema irrestrito. A energia está em unidades de $k_B T_f$.

de restrição (K_Q). A energia livre tem dois mínimos separados por uma barreira (estado de transição) por volta de $Q = 0,5$: um mínimo próximo em $Q = 0,9$ (estado enovelado) e outro em $Q = 0,1$ (estado desenovelado). A Figura 3.2 foi obtida fixando o termo de restrição em $Q^* = 0,5$ (estado de transição) na temperatura de enovelamento (T_f). T_f é a temperatura na qual metade da população está enovelada enquanto que a outra metade está em outra configuração qualquer. A energia livre com $K_Q = 0$ correspondem à simulação do sistema original sem o termo de restrição e pode ser considerada como o modelo Gō puro. Para esse caso, encontra-se uma barreira de ativação de aproximadamente 7 k_bT que é consistente com os resultados experimentais obtidos pelo grupo de Schuler (veja Figura 3.1) e Eaton [68, 69]. Variando o valor de K_Q, pode-se verificar que a forma da energia livre se modifica em torno de Q^*. Como nosso objetivo é tornar a região em torno de $Q = 0,5$ quasi-harmônica sem distorcer completamente o perfil de energia livre original, verificamos que a faixa de K_Q que satisfaz essa condição está em torno de $10 < K_Q < 300$. Nessa faixa de K_Q, o histograma de energia é deslocado para os estados em $Q^* = 0,5$ que são privilegiados quando comparados com a distribuição próxima da original (histograma da Figura 2.3). Dessa forma, deslocando-se o parâmetro Q^* pela coordenada de reação, pode-se obter o coeficiente de difusão em função de toda a coordenada.

Após a análise da intensidade da constante do potencial harmônico de restrição (K_Q) calcula-se o coeficiente de difusão configuracional $D(Q)$. A Figura 3.3 mostra os coeficientes de difusão (D) calculados pela equação 2.6 em função de Q^* em T_f e sem frustração no potencial ($\epsilon_{NC} = 0$). A Figura 3.3 apresenta o coeficiente de difusão em função de Q ($D(Q)$) para diferentes valores da constante de mola K_Q. Como verificado anteriormente, na faixa de valores intermediários de K_Q ($10 < K_Q < 300$), $D(Q)$ varia muito pouco comparando as curvas entre si o que significa que D será independente da escolha da constante de força de restrição se calculado nesta faixa de K_Q. $D(Q)$ depende de movimentos locais e mínimos existentes na superfície de energia. A baixa difusão no estado desenovelado mostra que há uma dificuldade na dinâmica da proteína em se difundir na direção do estado enovelado por flutuações aleatórias no modelo utilizado; existe uma dificuldade em atravessar a barreira de energia livre no estado de transição. $D(Q)$ volta a aumentar no estado enovelado (ainda relativamente baixo) devido a uma entropia residual

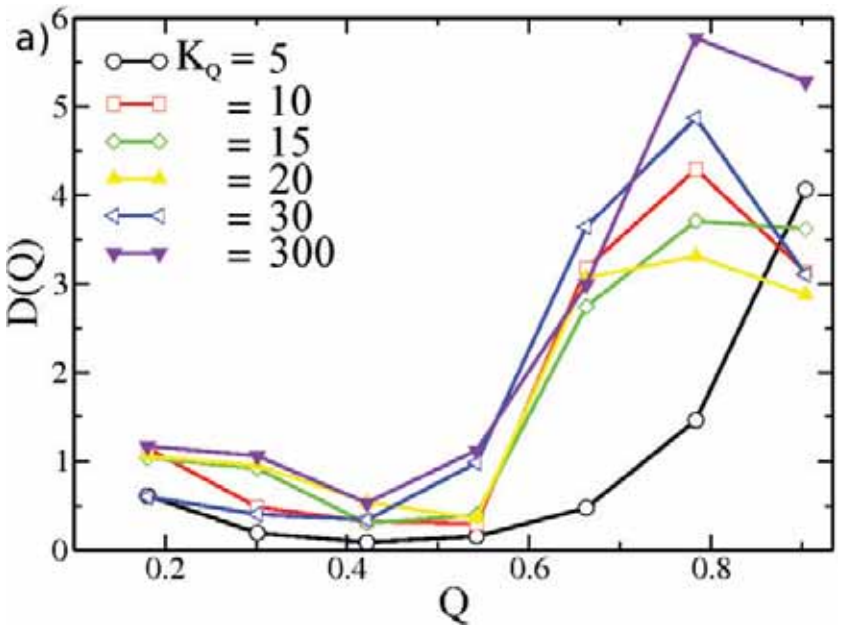


Figura 3.3: Coeficiente de difusão D em função da coordenada de reação Q obtido para cada Q^* escolhido como parâmetro do potencial harmônico de restrição para cada simulação realizada. É calculado $D(Q)$ para diferentes constantes de restrição K_Q: entre 5 e 300. Na faixa de valores intermediários de K_Q ($10 < K_Q < 300$), $D(Q)$ varia muito pouco comparando os valores das curvas entre si o que significa que D será independente da escolha da força de restrição. As simulações foram executadas na temperatura de enovelamento (T_f) e D tem unidade de inverso do tempo em unidades reduzidas.

nesse estado, a proteína ainda é capaz de realizar pequenos movimentos locais que não alteram Q. Yang e colaboradores [27, 70] calcularam $D(Q)$ por uma difusão da distribuição gaussiana de probabilidades em função de Q ($P(Q)$) e obtiveram um comportamento muito similar ao da Figura 3.3. Hummer e colaboradores [26, 66], aplicando a análise de Bayesian na simulação de réplicas em dinâmica molecular, também encontraram esse comportamento para $D(Q)$: baixo D na barreira de energia e relativamente maior nos vales de energia livre. No entanto, em trabalho posterior [32], Hummer encontrou um comportamento oposto para $D(Q)$ modificando o coeficiente de fricção da simulação.

Nesse trabalho de doutorado foi mostrado que o coeficiente de difusão é dependente da posição (coordenada de reação) e apresenta um comportamento não-monotônico em função da coordenada. Foram encontrados altos valores nos estados enovelado e desenovelado e baixos valores no estado de transição. Isso indica que a rugosidade da superfície de energia não é sempre a mesma num espaço de configuração unidimensional. Esses altos valores de D não são esperados intuitivamente no estado nativo. O comportamento de D, no estado nativo, foi compreendido pelo fato que esse estado é levemente degenerado. O estado nativo apresenta uma entropia residual permitindo que a proteína explore o espaço configuracional nessa região mais rapidamente. Consequentemente, D aumenta na direção do estado nativo.

O coeficiente de difusão também foi estimado com uma pequena quantidade de frustração energética introduzida no sistema. Essa frustração estabilizou o estado de transição e foi responsável pelo aumento de $D(Q)$ nesse estado, com relação ao caso não frustrado, na direção do estado enovelado. Esse resultado foi constatado com o cálculo dos valores-Φ por resíduo no estado de transição bem como por resíduo ao longo de Q. Os cálculos mostraram que os valores-Φ foram parcialmente homogeneizados na região do estado de transição que pode explicar o aumento de D com a presença de uma leve frustração energética.

Com isso, conclui-se que o coeficiente de difusão está intimamente ligado aos aspectos topológicos e energéticos da proteína e contribuiu consideravelmente para o entendimento do processo de enovelamento de pequenas proteínas globulares. Proteínas

maiores e com motivos mais complexos que a *Tm*CSP serão estudadas em trabalhos posteriores.

3.2 O coeficiente de difusão dependente da coordenada e do tempo

Nesta seção, serão discutidos os resultados do trabalho "*Coordinate and Time-Dependent Diffusion Dynamics in Protein Folding*", apêndice B, publicados na revista *Methods*. O objetivo principal desse trabalho foi desenvolver métodos analíticos e computacionais para explorar o enovelamento de proteína do ponto de vista difusivo.

O estudo da cinética do enovelamento de proteína é importante para compreender o seu mecanismo. O coeficiente de difusão D está estritamente relacionado com a cinética, pois reflete a dificuldade da proteína de vencer as barreiras energéticas locais impostas pela rugosidade do funil de energia. Na seção anterior (seção 3.1) constatou-se que D é dependente da coordenada de reação (Q), pois o sistema proteico possui barreiras locais ao longo de Q. Por outro lado, numa coordenada fixa (Q qualquer), o tempo de escape local dependerá da distribuição de barreiras energéticas ao seu redor, então D também dependerá do tempo.

A dependência espacial do coeficiente de difusão modifica a altura da barreira cinética em F resultando na alteração da posição do estado de transição [29]. Isso promove implicações no cálculo das taxas de enovelamento de proteína que dependem essencialmente da altura da barreira em F e do pré-fator D. Além disso, altera-se as rotas de enovelamento com a modificação do estado de transição.

A proteína analisada, a *Tm*CSP, é mencionada na seção anterior e nos artigos anexados nos apêndices A e B. A *Tm*CSP é alvo de pesquisadores como Schuler e colaboradores [16], que por meio de experimentos de uma única molécula, obtiveram o coeficientes de difusão e os perfis de energia livre em função de um parâmetro de ordem.

Nesse trabalho, verificou-se que o efeito em D, por ser dependente do tempo, pode resultar em cinéticas não-exponenciais e distribuições de tempo de enovelamento de estatística não-poissônica. Uma distribuição de barreiras energéticas e escalas de tempo,

que conduzirão ao coeficiente de difusão dependente do tempo e à cinética não-exponencial, surgem com a projeção do relevo de superfície de energia multidimensional em uma única ou poucas dimensões. A distribuição de coeficientes de difusão temporal é um reflexo da rugosidade local em uma determinada posição da superfície de energia.

Com isso, a teoria sobre o coeficiente de difusão temporal terá importância no entendimento de experimentos de única molécula. Nesses experimentos pode-se observar uma complexa cinética em que diferentes escalas de tempo podem coexistir. Se a proteína possuir uma baixa ou nenhuma barreira energética, comparável à flutuação termal $1kT$, a contribuição de D para a cinética é decisiva, pois o pré-fator D dominará completamente a dinâmica do sistema.

Em especial, Martin Gruebele e colaboradores têm investigado a cinética de enovelamento de proteínas super-rápidas como a λ_{6-85} *repressor* [71, 72] e a *WW domain* [73]. Essas proteínas possuem baixa barreira de ativação e, por isso, as taxas de enovelamento são diretamente proporcionais ao coeficiente de difusão [20]. Os experimentos e simulações de Gruebele, em superfícies de energia livre em duas dimensões (2D), sugerem que, devido à uma dinâmica não-exponencial, deve haver uma difusão ao longo da energia livre local da coordenada de reação (rugosidade longitudinal) e uma difusão numa superfície de energia livre multidimensional (rugosidade transversal) [71]. Com isso, sugere-se que existe a necessidade de análises do coeficiente de difusão em função de uma coordenada de reação e do tempo, já que as evidências teóricas e experimentais apontam para esse sentido.

3.3 Descritor de superfície de energia – *The Landscape Descriptor*

Nesta seção, discute-se o manuscrito em preparação que será submetido a uma revista em processo de escolha. O estágio atual do manuscrito está anexado no apêndice C com título inicialmente sugerido: "*Topography of Funneled Landscape Determines the Thermodynamics and Kinetics of Protein Folding*". Neste trabalho, quantificou-se o relevo de superfície de energia real de proteína e propõem-se uma grandeza física que relaciona as principais características do relevo do funil de energia como um novo método para classificar as proteínas.

Nas seções anteriores, mencionou-se que a principal teoria mais aceita atualmente, acerca do enovelamento de proteína, é a teoria de superfície de energia. Os grupos experimentais têm-se interessado em medir o funil de superfície de energia e relacioná-lo às grandezas físicas experimentais mensuráveis como temperatura de enovelamento, temperatura de vidro, taxas cinéticas, barreira de ativação, etc. Dessa forma, estabelece-se uma conexão da topografia da superfície de energia com a cinética e a termodinâmica, por meio de uma análise estatística do enovelamento de proteína.

Pode-se identificar três quantidades essenciais, baseadas na densidade de estados proteicos, para caracterizar a topografia da superfície de energia:

- a distância energética entre o estado nativo e a média dos estados não-nativos

$$\delta E = |E_n - \overline{E}_{non-native}|, \tag{3.1}$$

- a rugosidade ou a variância das energias dos estados não-nativos, medidos pelo desvio padrão das energias dos estados não-nativos

$$\Delta E = \sqrt{\langle E^2 \rangle - \langle E \rangle^2}, \tag{3.2}$$

- o tamanho do funil medido pela entropia do sistema

$$S = k_B ln[\Omega] \tag{3.3}$$

sendo k_B a constante de Boltzmann e Ω o número de estados. Estas quantidades podem ser identificadas na Figura 3.4 e, determinadas para a proteína TmCSP, na Figura 3.5. Também foram determinadas para uma gama de proteínas cujos resultados encontram-se no anexo C.

A razão entre as equações 3.1, 3.2 e 3.3 resulta numa grandeza adimensional

$$\Lambda = \frac{\delta E}{\Delta E \sqrt{2 S_0}} \tag{3.4}$$

que denominou-se Descritor de Superfície (LD[1]). A equação 3.4 pode ser deduzida, analiticamente, resolvendo a razão entre as expressões da temperatura de enovelamento (T_f) e de vidro (T_g), estabelecidas na comunidade científica desde o final da década de 80 [6, 10, 17, 40, 74–77].

Inicialmente, utilizou-se o modelo de rede [29, 78] com intuito de realizar testes do cálculo de LD. O próximo passo foi aprimorar o nível de refinamento do modelo, empregando o modelo baseado na estrutura, somente com carbonos alfas (modelo C_α-Gō) [30, 31, 49, 79]. Foram simuladas nove proteínas, organizadas por ordem de número de aminoácidos e diferentes motivos, com o modelo C_α-Gō. A fim de melhorar mais o refinamento dos modelos, foi empregado o modelo de dinâmica molecular com todos os átomos imersos em campo de força do pacote AMBER[2] [80]. Nesse modelo, a contribuição do solvente foi incluída implicitamente. Para o modelo com todos os átomos, simulou-se duas proteínas com objetivo de comparar com as simulações utilizando modelo C_α-Gō.

Observou-se que, para todos os modelos utilizados, LD está relacionado com as grandezas termodinâmicas e cinéticas de enovelamento. LD aumenta monotonicamente com a razão entre a estabilidade termodinâmica e a temperatura de armadilhamento

[1]Do inglês *Landscape Descriptor.*

[2]Do inglês *Assisted Model Building with Energy Refinement.*

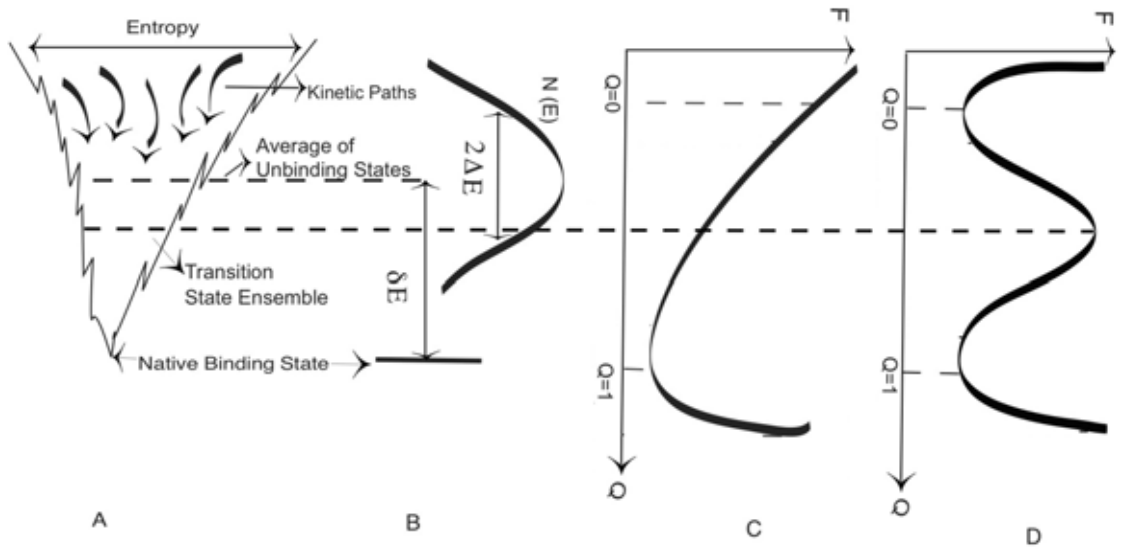


Figura 3.4: Esboço idealizado teoricamente para a) superfície de energia com energia total no eixo vertical e a entropia configuracional sendo a largura do funil para cada um dos planos na energia total b) densidade de estados ($n(E)$) e energia livre (F) em função do número de contatos nativos (Q) para o caso de um sistema de c) um estado e d) dois estados. Os painéis a) e b) estão alinhados pela mesma energia total no eixo vertical e os painéis c) e d) estão alinhados por Q no eixo vertical. δE é a diferença de energia do estado fundamental para a energia média e ΔE é a rugosidade na superfície de energia. As setas verticais em a) representam os múltiplos caminhos, rotas cinéticas, pelas quais a reação pode ocorrer a caminho do fundo do funil, o estado nativo. A linha tracejada horizontal que marca os painéis a) b) c) e d) representa o estado de transição. Este esboço representa a superfície de energia para o caso de um sistema de ligação biomolecular (dímero proteico, por exemplo) que pode ser estendida para o caso simples de uma única proteína transpondo os conceitos de estados ligado/desligado por estados enovelado/desenovelado. *Adaptado de* [40].

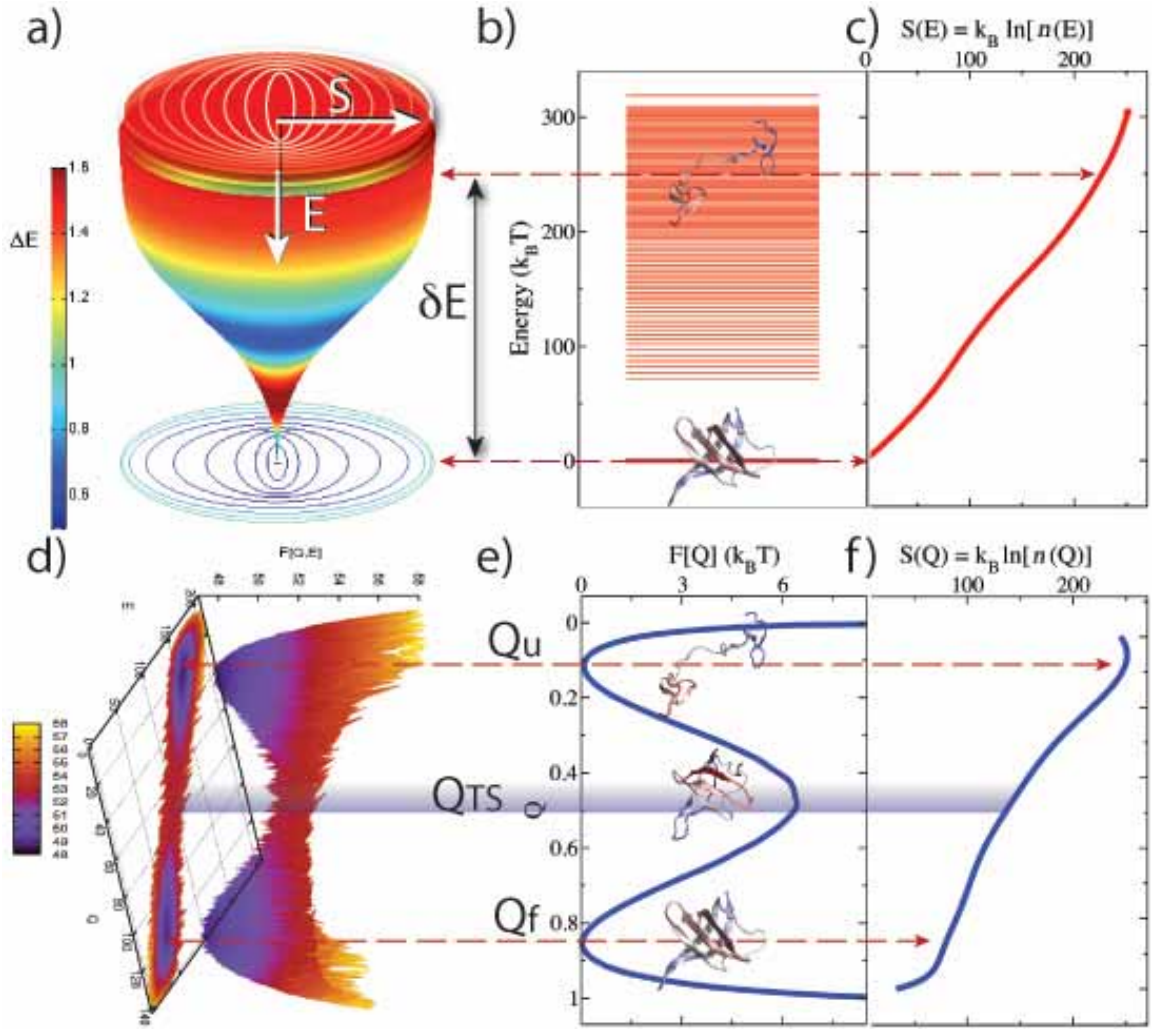


Figura 3.5: Resultados obtidos por simulação computacional da proteína *Tm*CSP. a) Superfície de energia afunilada com a energia total (E) no eixo vertical e a entropia configuracional (S) como a área do funil para cada um dos planos em E. A área do funil, extraída de f), escala com $(1-Q)^{\gamma}$, sendo γ arbitrário. δE é a diferença de energia do estado fundamental para a energia média e ΔE é a rugosidade na superfície de energia representada pela escala de cor. O estado de transição encontra-se aproximadamente na faixa azul escura. b) Espectro com os níveis de energia da proteína. c) Entropia em função da energia total ($S(E)$). Os painéis a) b) e c) estão alinhados pelo mesmo eixo vertical E e δE. d) Energia livre (F) em função da energia total (E) e do número de contatos nativos (Q). e) Energia livre ($F(Q)$) e f) Entropia ($S(Q)$) em função de Q. Os painéis e) e f) estão alinhados pelo mesmo eixo vertical Q e o painel d) está rotacionado de forma a superpor as setas dos painéis e) e f) que indicam os estados nativo (Q_f), desenovelado (Q_u) e de transição (Q_{TS}). d) e e) estão na temperatura de enovelamento T_f.

(T_f/T_g). Por outro lado, o LD decai monotonicamente com as taxas cinéticas moduladas pelo tamanho da proteína. Outro estudo realizado foi a introdução de frustração energética no sistema, por meio de um potencial perturbativo, e verificou-se a influência em LD. Utilizou-se novamente a *Tm*CSP. Com relação à adição crescente de frustração, notou-se a existência de um valor máximo de LD. Essa frustração ótima coincide com o valor no qual a proteína se enovela mais rapidamente: o ponto mínimo para a curva em "U" da média do primeiro tempo de passagem para o estado enovelado (MFPT[3]). Essa mesma frustração ótima também é encontrada na curva de T_f contra a frustração, correspondendo ao ponto máximo dessa curva (seção 3.1).

Os resultados foram discutidos mais detalhadamente no apêndice C, os quais podemos resumir na seguinte conclusão: LD é um parâmetro que descreve e classifica, satisfatoriamente, as proteínas bem como seu relevo de energia, relacionando-o com as taxas de enovelamento e estabilidade termodinâmica. Essa nova abordagem conecta a topografia do relevo de superfície com a cinética e termodinâmica do enovelamento de proteína, proporcionando aos grupos experimentais quantificarem um funil de energia real. O rápido desenvolvimento das técnicas experimentais de uma única molécula é fundamental para essa tarefa.

[3]Do inglês *Mean First Passage Time*.

Capítulo 4

Conclusão e Perspectivas Futuras

A teoria de superfície de energia mostrou ser uma ferramenta fundamental para se compreender complexos mecanismos de enovelamento de proteína. Aliado à teoria, o desenvolvimento de modelos computacionais minimalistas permite que seja acessível as escalas temporais que envolvem as reações de enovelamento de enovelamento de proteína e até outros problemas biomoleculares. No caso particular do estudo do coeficiente de difusão no enovelamento de proteína, com modelos simples como o C_α e todos os átomos com potencial baseado na estrutura (Gō), é possível amostrar exaustivamente o espaço de fase de pequenas proteínas globulares com o uso de recursos computacionais existentes [81].

Observações experimentais afirmam que a difusão torna-se extremamente importante no estado de transição de proteínas globulares de baixa barreira energética e em toda a coordenada de reação para proteínas sem barreira [82]. Durante o trabalho de doutorado, foi desenvolvida a teoria analítica e os métodos computacionais para explorar a dinâmica difusiva de uma proteína através de um potencial efetivo de energia. O estudo do enovelamento de proteína através de dinâmica molecular de modelo Gō é um método muito eficiente na determinação do coeficiente de difusão.

A proteína globular *Tm*CSP foi estudada por dinâmica molecular para o cálculo

do coeficiente de difusão (D) em função da coordenada de reação (Q e R_g). Verificou-se que D tem um dependência não-monotônica com Q com um mínimo no estado de transição. Esta dependência pode ser atribuída a uma entropia residual presente no estado nativo que faz aumentar D neste ensemble. Também, adicionando frustração energética, existe uma quantidade ótima que maximiza a estabilidade térmica, torna o estado de transição polarizado e aumenta a difusão para o estado enovelado. A rugosidade do relevo de energia da *Tm*CSP, determinada pelo coeficiente de difusão, e a barreira energética estão em comum acordo com resultados experimentais. Os aspectos topológicos e energéticos determinam a dinâmica difusiva no enovelamento de proteína.

Esse trabalho mostrou também que D está relacionado com a cinética e que, localmente, num ponto qualquer de Q, o tempo de escape depende da distribuição de barreiras energéticas na região e conclui-se que D tem uma dependência temporal. Esta dependência temporal remete às diferentes escalas temporais existentes na complexa reação de enovelamento.

Em trabalhos futuros será possível calcular explicitamente a dependência do tempo e da posição na difusão aplicando os conceitos e as técnicas estudas nesse presente trabalho e confrontá-las com outras técnicas, descritas na literatura, para calcular D [26–28]. Também será possível estender o cálculo de $D(Q)$ para proteínas sem barreira de rápido enovelamento. Pode-se propor experimentos de uma única molécula para estudar o coeficiente de difusão dependente da coordenada e do tempo e verificar computacionalmente a dinâmica através do relevo de energia. Desta forma, haverá uma melhor compreensão dos mecanismos de enovelamento de proteínas de um ou dois estados aliando-se estreitamente a relação entre grupos teóricos e experimentais.

Com a quantificação da superfície de energia de enovelamento, pode-se afirmar que, para todas as proteínas estudadas, todas apresentaram uma superfície afunilada como previsto pela teoria. A razão LD (Λ) que caracteriza a topologia do enovelamento pela superfície de energia proporciona um novo método para se classificar as proteínas. Mostrou-se que Λ está fortemente correlacionado com a cinética e termodinâmica do enovelamento, assim unindo estas duas características importantes de uma proteína.

As ferramentas analíticas e computacionais desenvolvidas nesta tese podem ser estendidas para sistemas biomoleculares maiores e mais complexos como dímeros, trímeros, proteína-ligante, DNA-proteína, mudança conformacional, efeito de solvente, etc.

Referências Bibliográficas

[1] P.E. Leopold, M. Montal, and J.N. Onuchic. Protein folding funnels - A kinetic approach to the sequence structure relationship. *Proc. Natl. Acad. Sci. USA*, 18:8721–8725, 1992.

[2] 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*, 21:167–195, 1995.

[3] C.B. Anfinsen; E. Harber; M. Sela and F. White Jr. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proc. Natl. Acad. Sci. USA*, 47:1309–1314, 1961.

[4] C.B. Anfinsen. Principles that govern the folding of proteins chains. *Science*, 181:223–230, 1973.

[5] C. Levinthal. Are there pathways for protein folding? *J. Chem. Phys.*, 65:44–45, 1968.

[6] J.D. Bryngelson and P.G. Wolynes. Intermediates and barrier crossing in a random energy-model (with applications to protein folding). *J. Phys. Chem.*, 93(19):6902–6915, 1989.

[7] R.L. Baldwin. The nature of protein folding pathways: The classical versus the new view. *J. Biomol. NMR*, 5:103–109, 1995.

[8] V.S. Pande, A.Yu. Grosberg, and T. Tanaka. On the theory of folding kinetics for short proteins. *Folding and Design*, 2(2):109–114, 1997.

[9] B.A. Shoemaker, J. Wang, and P.G. Wolynes. Structural correlations in protein folding funnels. *Proc. Natl. Acad. Sci. USA*, 94:777–782, 1997.

[10] J.N. Onuchi; H. Nymeyer; A.E. García; J. Chahine and N. D. Socci. The energy landscape theory of protein folding: Insights into folding mechanisms and scenarios. *Adv. Protein Chem.*, 53:87–152, 2000.

[11] 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.*, 126:8426–8432, 2004.

[12] C.L. Lee, G.S., and J. Wang. First-passage time distribution and non-markovian diffusion dynamics of protein folding. *J. Chem. Phys.*, 118(2):959–968, 2003.

[13] Y. Levy and J.N. Onuchic. Mechanisms of protein assembly: Lessons from minimalist models. *Acc. Chem. Res.*, 39:135–142, 2006.

[14] D.M. Zuckerman. Simulation of an ensemble of conformational transitions in a united-residue model of calmodulin. *J. Phys. Chem. B*, 108:5127–5137, 2004.

[15] L. Sutto, J. Lätzer, J.A. Hegler, D.U. Ferreiro, and P.G. Wolynes. Consequences of localized frustration for the folding mechanism of the IM7 protein. *Proc. Natl. Acad. Sci. USA*, 104(50):19825–19830, 2007.

[16] D. Nettels, I.V. Gopich, A. Hoffmann, and B. Schuler. Ultrafast dynamics of protein collapse from single-molecule photon statistics. *Proc. Natl. Acad. Sci. USA*, 104:2655–2660, 2007.

[17] J.D. Bryngelson and P.G. Wolynes. Spin-glasses and the statistical mechanics of protein folding. *Proc. Natl. Acad. Sci. USA*, 84:7524–7528, 1987.

[18] H.A. Kramers. Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica*, 7:284–304, 1940.

[19] N.D. Socci, J.N. Onuchic, and P.G. Wolynes. Diffusive dynamics of the reaction coordinate for protien folding funnels. *J. Chem. Phys.*, 104:5860–5868, 1996.

[20] J. Kubelka, J. Hofrichter, and W.A. Eaton. The protein folding "speed limit". *Curr. Opin. Struct. Biol.*, 14:76–88, 2004.

[21] N. Metropolis; A.W. Rosenbluth; M.N. Rosenbluth; A.N. Teller and E. Teller. Equation of state calculations by fast computing machines. *J. Chem. Phys.*, 21:1087–1092, 1953.

[22] K. A. Dill and H. S. Chan. From Levinthal to pathways to funnels. *Nat. Struct. Biol.*, 4:10–19, 1997.

[23] V.B.P. Leite and J.N. Onuchic. Structure and dynamics of solvent landscapes in charge-transfer reactions. *J. Phys. Chem.*, 100:7680–7690, 1996.

[24] V.B.P. Leite. Smooth landscape solvent dynamics in electron transfer reactions. *J. Chem. Phys.*, 110(20):10067–10075, 1999.

[25] A. Baumketner and Y. Hiwatari. Diffusive dynamics of protein folding studied by molecular dynamics simulations of an off-lattice model. *Phys. Rev. E*, 66(1):011905, 2002.

[26] G. Hummer. Position-dependent diffusion coefficients and free energies from bayesian analysis of equilibrium and replica molecular dynamics simulations. *New J. Phys.*, 7:34, 2005.

[27] S. Yang, J.N. Onuchic, and H. Levine. Effective stochastic dynamics on a protein folding energy landscape. *J. Chem. Phys.*, 125(5):054910–054918, 2006.

[28] A.K. Sangha and T. Keyes. Proteins fold by subdiffusion of the order parameter. 113(48):15886–15894, 2009.

[29] J. Chahine, R.J. Oliveira, V.B.P. Leite, and J. Wang. Configuration-dependent diffusion can shift the kinetic transition state and barrier height of protein folding. *Proc. Natl. Acad. Sci. USA*, 104(37):14646–14651, 2007.

[30] R.J. Oliveira, P.C. Whitford, J. Chahine, V.B.P. Leite, and J. Wang. Coordinate and time-dependent diffusion dynamics in protein folding. *Methods*, 52(1):91–98, 2010.

[31] R.J. Oliveira, P.C. Whitford, J. Chahine, J. Wang, J.N. Onuchic, and V.B.P. Leite. The origin of nonmonotonic complex behavior and the effects of nonnative interactions on the diffusive properties of protein folding. *Biophys. J.*, 99(2):600–608, 2010.

[32] R.B. Best and G. Hummer. Coordinate-dependent diffusion in protein folding. *Proc. Natl. Acad. Sci. USA*, 107(3):1088 –1093, 2010.

[33] M. Karplus and D. L. Weaver. Protein folding dynamics: the diffusion-collision model and experimental data. *Prot. Sci.*, 3(4):650–668, 1994.

[34] W. Kremer, B. Schuler, S. Harrieder, M. Geyer, W. Gronwald, C. Welker, R. Jaenicke, and H. R. Kalbitzer. Solution NMR structure of the cold-shock protein from the hyperthermophilic bacterium *Thermotoga maritima*. *Eur. J. Biochem.*, 268(9):2527–2539, 2001.

[35] B. Schuler. Single-Molecule fluorescence spectroscopy of protein folding. *ChemPhysChem*, 6(7):1206–1220, 2005.

[36] A. Hoffmann, A. Kane, D. Nettels, D.E. Hertzog, P. Baumgartel, J. Lengefeld, G. Reichardt, D.A. Horsley, R. Seckler, O. Bakajin, and B. Schuler. Mapping protein collapse with single-molecule fluorescence and kinetic synchrotron radiation circular dichroism spectroscopy. *Proc. Natl. Acad. Sci. USA*, 104:105–110, 2007.

[37] T. Cellmer, E.R. Henry, J.H., and W.A. Eaton. Measuring internal friction of an ultrafast-folding protein. *Proc. Natl. Acad. Sci. USA*, 105(47):18320 –18325, 2008.

[38] N.D. Socci and J.N. Onuchic. Folding kinetics of proteinlike heteropolymers. *J. Chem. Phys.*, 101:1519–1528, 1994.

[39] P.G. Wolynes, J.N. Onuchic, and D. Thirumalai. Navigating the folding routes. *Science*, 267(5204):1619–1620, 1995.

[40] J. Wang and G.M. Verkhivker. Energy landscape theory, funnels, specificity, and optimal criterion of biomolecular binding. *Phys. Rev. Lett.*, 90(18):188101, 2003.

[41] J.N. Onuchic and P.G. Wolynes. Theory of protein folding. *Curr. Opin. Struct. Biol.*, 14:70–75, 2004.

[42] A. Borgia, P.M. Williams, and J. Clarke. Single-Molecule studies of protein folding. *Annu. Rev. Biochem.*, 77(1):101–125, 2008.

[43] M. Karplus. Behind the folding funnel diagram. *Nat. Chem. Biol.*, 7:401–404, 2011.

[44] K.A. Dill; S. Bromberg; K. Yue; K.M. Fiebig; D.P. Yee; P.D. Thomas and H.S. Chan. Principal of protein folding — A perspective from simple exact models. *Prot. Sci.*, 4:561–602, 1995.

[45] DK. Klimov and D. Thirumalai. Mechanisms and kinetics of beta-hairpin formation. *Proc. Natl. Acad. Sci. USA*, 97:2544–2549, 2000.

[46] J. Shimada, E.L. Kussell, and E.I. Shakhnovich. The folding thermodynamics and kinetics of crambin using an all-atom monte carlo simulation. *J. Mol. Biol.*, 308:79–95, 2001.

[47] R. Du, V.S. Pande, A.Yu. Grosberg, T. Tanaka, and E.S. Shakhnovich. On the transition coordinate for protein folding. *J. Chem. Phys.*, 108(1):334–350, 1998.

[48] V.S. Pande and D.S. Rokhsar. Molecular dynamics simulations of unfolding and refolding of beta-hairpin fragment of protein G. *Proc. Natl. Acad. Sci. USA*, 96(16):9062–9067, 1999.

[49] 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(5):937–953, 2000.

[50] P.C. Whitford, J.K. Noel, S. Gosavi, A. Schug, K.Y. Sanbonmatsu, and J.N. Onuchic. An all-atom structure-based potential for proteins: Bridging minimal models with all-atom empirical forcefields. *Proteins: Struct. Func. Biolnf.*, 75(2):430–441, 2009.

[51] J.K. Noel, P.C. Whitford, K.Y. Sanbonmatsu, and J.N. Onuchic. Smog@ctbp: simplified deployment of structure-based models in GROMACS. *Nucleic Acids Res.*, 2010.

[52] Q. Lu, H.P. Lu, and J. Wang. Exploring the mechanism of flexible biomolecular recognition with single molecule dynamics. *Phys. Rev. Lett.*, 98(12):128105, 2007.

[53] A. Schug, P. C. Whitford, Y. Levy, and J. N. Onuchic. Mutations as trapdoors to two competing native conformations of the Rop-dimer. *Proc. Natl. Acad. Sci. USA*, 104(45):17674–17679, 2007.

[54] P.C. Whitford, O. Miyashita, Y. Levy, and J.N. Onuchic. Conformational transitions of adenylate kinase: Switching by cracking. *J. Mol. Biol.*, 366:1661–1671, 2007.

[55] R.B. Best, Y. Chen, and G. Hummer. Slow protein conformational dynamics from multiple experimental structures: The helix/sheet transition of arc repressor. *Structure*, 13:1755–1763, 2005.

[56] Q. Lu and J. Wang. Single molecule conformational dynamics of adenylate kinase: Energy landscape, structural correlations, and transition state ensembles. *J. Am. Chem. Soc.*, 130(14):4772–4783, 2008.

[57] Y. Ueda, H. Taketomi, and N. Gō. Studies on protein folding, unfolding and fluctuations by computer simulation. I. The effects of specific amino acid sequence represented by specific inter-unit interactions. *Int. J. Peptide Res.*, 7:445–459, 1975.

[58] V. Sobolev, R. Wade, G. Vried, and M. Edelman. Molecular docking using surface complementarity. *Proteins: Struct. Funct. Genet.*, 25:120–129, 1996.

[59] W. Humphrey, A. Dalke, and K. Schulten. VMD: Visual Molecular Dynamics. *J. Mol. Graphics*, 14(1):33–38, 1996.

[60] G. M. Torrie and J. P. Valleau. Nonphysical sampling distributions in Monte Carlo free-energy estimation - Umbrella sampling. *J. Comp. Phys.*, 23:187–199, 1977.

[61] C. Bartels and M. Karplus. Multidimensional adaptive umbrella sampling: Applications to main chain and side chain peptide conformations. *J. Comp. Phys.*, 18:1450–1462, 1997.

[62] Swendsen RH. Ferrenberg AM. New monte carlo technique for studying phase transitions. *Phys. Rev. Lett.*, 61:2635–2638, 1988.

[63] Swendsen RH. Ferrenberg AM. Optimized monte-carlo data analysis. *Phys. Rev. Lett.*, 63:1195–1198, 1989.

[64] M. Jacob and F.X. Schmid. Protein folding as a diffusional process. *Biochemistry*, 38(42):13773–13779, 1999.

[65] J. Kubelka, J. Hofrichter, and W.A. Eaton. The protein folding "speed limit". *Curr. Opin. Struct. Biol.*, 14(1):76–88, 2004.

[66] R.B. Best and G. Hummer. Diffusive model of protein folding dynamics with Kramers turnover in rate. *Phys. Rev. Lett.*, 96(22):228104, 2006.

[67] H. Nymeyer, A.E. Garcia, and J.N. Onuchic. Folding funnels and frustration in off-lattice minimalist protein landscapes. *Proc. Natl. Acad. Sci. USA*, 95:5921–5928, 1998.

[68] B. Schuler, W. Kremer, H.R. Kalbitzer, and R. Jaenicke. Role of entropy in protein thermostability: Folding kinetics of a hyperthermophilic cold shock protein at high temperatures using 19F NMR. *Biochemistry*, 41:11670–11680, 2002.

[69] B. Schuler, E.A. Lipman, and W.A. Eaton. Probing the free-energy surface for protein folding with single-molecule fluorescence spectroscopy. *Nature*, 429:743–747, 2002.

[70] S. Yang, J.N. Onuchic, A.E. Garcia, and H. Levine. Folding time predictions from all-atom replica exchange simulations. *J. Mol. Biol.*, 372(3):756–763, 2007.

[71] F. Liu and M. Gruebele. Downhill dynamics and the molecular rate of protein folding. *Chem. Phys. Lett.*, 461(1-3):1 – 8, 2008.

[72] H. Ma and M. Gruebele. Kinetics are probe-dependent during downhill folding of an engineered λ_{6-85} protein. *Proc. Natl. Acad. Sci. USA*, 102(7):2283–2287, 2005.

[73] F. Liu, D. Du, A.A. Fuller, J.E. Davoren, P. Wipf, J.W. Kelly, and M. Gruebele. An experimental survey of the transition between two-state and downhill protein folding scenarios. *Proc. Natl. Acad. Sci. USA*, 105(7):2369–2374, 2008.

[74] V. I. Abkevich, A. M. Gutin, and E. I. Shakhnovich. Free energy landscape for protein folding kinetics: Intermediates, traps, and multiple pathways in theory and lattice model simulations. *J. Chem. Phys.*, 101(3):6052–6062, 1994.

[75] J. Wang, W. Huang, H. Lu, and E. Wang. Downhill kinetics of biomolecular interface binding: Globally connected scenario. *Biophys. J.*, 87(4):2187 – 2194, 2004.

[76] J. Wang, C. Lee, and G. Stell. The cooperative nature of hydrophobic forces and protein folding kinetics. *Chem. Phys.*, 316(1-3):53 – 60, 2005.

[77] J. Wang, L. Xu, and E. Wang. Optimal specificity and function for flexible biomolecular recognition. *Biophys. J.*, 92(12):L109 – L111, 2007.

[78] V.B.P. Leite, J.N. Onuchic, and J. Wang. Probing the kinetics of single molecule protein folding. *Biophys. J.*, 87(6):3633–3641, 2004.

[79] L 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 for several reduced models. *J. Phys. Chem. B*, 2007, DOI: 10.1021/jp0769835.

[80] D.A. Case, T.E. Cheatham, T. Darden, H. Gohlke, R. Luo, K.M. Merz, A. Onufriev, C. Simmerling, B. Wang, and R.J. Woods. The amber biomolecular simulation programs. *J. Comp. Chem.*, 26(16):1668–1688, 2005.

[81] http://www.rocksclusters.org/rocks-register/details.php?id=1391.

[82] W.Y. Yang and M. Gruebele. Folding at the speed limit. *Nature*, 423(6936):193–197, May 2003.

Apêndice A

Artigo publicado na revista *Biophysical Journal*

O artigo, "*The Origin of Nonmonotonic Complex Behavior and the Effects of Nonnative Interactions on the Diffusive Properties of Protein Folding*", publicado na revista científica internacional *Biophysical Journal* refere-se aos resultados obtidos da colaboração de nosso grupo de pesquisa com o Prof. Dr. Jin Wang[1]. O artigo também apresenta parte dos resultados obtidos durante o trabalho de doutorado.

[1]Universidade e endereço para contato se encontram no artigo que segue nesse apêndice A.

The Origin of Nonmonotonic Complex Behavior and the Effects of Nonnative Interactions on the Diffusive Properties of Protein Folding

Ronaldo J. Oliveira,[†△] Paul C. Whitford,[‡§**△*] Jorge Chahine,[†] Jin Wang,[¶‖] José N. Onuchic,[**] and Vitor B. P. Leite[†*]

[†]Departamento de Física, Instituto de Biociências, Letras e Ciências Exatas, Universidade Estadual Paulista, São José do Rio Preto, Brazil; [‡]Theoretical Biology and Biophysics Group, Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico; [§]International Institute for Complex Adaptive Matter, University of California at Davis, Davis, California; [¶]Department of Chemistry and Department of Physics, State University of New York at Stony Brook, Stony Brook, New York; [‖]State Key Laboratory of Electroanalytical Chemistry, Changchun Institute of Applied Chemistry, Chinese Academy of Sciences, Changchun, Jilin, China; and **Center for Theoretical Biological Physics and Department of Physics, University of California at San Diego, San Diego, California

ABSTRACT We present a method for calculating the configurational-dependent diffusion coefficient of a globular protein as a function of the global folding process. Using a coarse-grained structure-based model, we determined the diffusion coefficient, in reaction coordinate space, as a function of the fraction of native contacts formed Q for the cold shock protein (*Tm*CSP). We find nonmonotonic behavior for the diffusion coefficient, with high values for the folded and unfolded ensembles and a lower range of values in the transition state ensemble. We also characterized the folding landscape associated with an energetically frustrated variant of the model. We find that a low-level of frustration can actually stabilize the native ensemble and increase the associated diffusion coefficient. These findings can be understood from a mechanistic standpoint, in that the transition state ensemble has a more homogeneous structural content when frustration is present. Additionally, these findings are consistent with earlier calculations based on lattice models of protein folding and more recent single-molecule fluorescence measurements.

INTRODUCTION

The energy landscape theory of protein folding (1–5) has been an invaluable theoretical framework for understanding protein folding (6–10), oligomerization (11–13), and functional transitions (14–18). According to the theory, the energy landscape associated with protein folding lacks large energetic traps and has an overall funnel shape where the native ensemble is the lowest energy state. These minimally frustrated landscapes can be idealized as being devoid of energetic roughness, which enables the use of structure-based (Gō-like) models (8,10,19–23) to study the thermodynamic and kinetic properties of the folding process. Because these structure-based models lack energetic trapping, they also provide a means to characterize the topological contributions to folding.

Although there is a strong correlation between simulated barrier heights and experimental folding times (24), rates are a consequence of both the free-energy profile and the diffusion coefficient (25,26). Accordingly, direct comparison between experiments and theory requires both quantities. In principle, one may circumvent the need for the diffusion coefficient by simulating many thousands of folding trajectories and calculating the mean first passage time of folding (27–29). Such approaches are often computationally intractable and they do not always advance our physical understanding of the process. Therefore, it is desirable to calculate both the free energy F and the diffusion coefficient D as functions of a global folding coordinate Q (30–34). If D is constant, then it only serves as a prefactor to the folding rate. However, when D is not constant, as we describe below, it can give rise to kinetic barriers in addition to the thermodynamic barriers (32).

The diffusion coefficient D is a result of the underlying energy landscape. As every conformation has a unique set of locally accessible interactions, D is a function Q. Although Q can be defined by a variety of measures, here we use the fraction of native contacts, as it has been shown to capture, accurately, the transition state ensemble of two-state proteins (35). Low values of Q correspond to the unfolded state and high values correspond to the folded ensemble. When Q is low, energetic contributions are largely from water-protein interactions. In the folded state (high Q), the burial of hydrophobic surface area can be the dominant energetic contribution. In these two regimes, the local energetic roughness can be quite different, which can lead to different diffusion coefficients.

Many recent efforts have attempted to characterize $D(Q)$ via experimental methods (36–42) and theoretical calculations (26,32–34,43–49). These studies have found that diffusion is not constant as a protein folds to the native state. This naturally leads to the question: Does diffusion vary with the degree of compactness because of energetic trapping, or topological constraints? To address this, we calculate D for a structure-based model that lacks energetic roughness and compare the findings to variants of the model that include energetic frustration.

In this article, we present the diffusive properties of a C_α structure-based model in molecular dynamics simulations.

Submitted November 19, 2009, and accepted for publication April 14, 2010.

[△]Ronaldo J. Oliveira and Paul C. Whitford contributed equally to this article.

*Correspondence: vleite@sjrp.unesp.br or whitford@lanl.gov

Editor: Feng Gai.

0006-3495/10/07/0600/9 $2.00

doi: 10.1016/j.bpj.2010.04.041

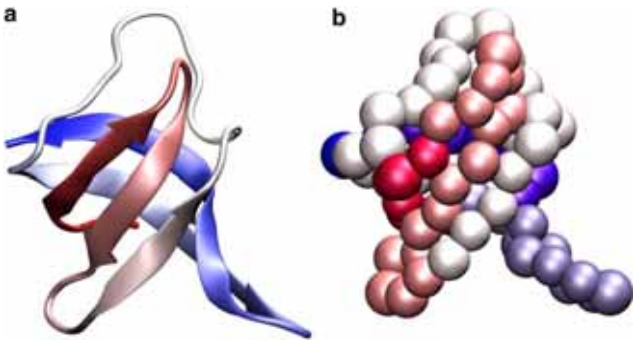


FIGURE 1 Cold-shock protein from *Thermotoga maritima* (*Tm*CSP Protein DataBank entry 1G6P (36)), shown in (*a*) cartoon representation and (*b*) C_α representation. The size of the atoms in panel *b* correspond to the excluded volume radii used in this model. The structures are colored from red (C-terminus) to blue (N-terminus) and were visualized with VMD (82). The *Tm*CSP is a small globular protein with 66 amino acids, molecular mass of 7.5 kDa, and a three-dimensional structure known as a Greek-key β-barrel (five β-strands divided in two antiparallel β -sheets).

We compare the results obtained from the unfrustrated model (i.e., only native interactions are stabilizing) with an energetically frustrated variant of the model, which allows for a quantitative, and qualitative, comparison of topological and energetic contributions to the diffusion coefficient. As there is a large body of experimental data available, including denaturant-dependent diffusion coefficient measurements, we chose to study the cold shock protein from the hyperthermophilic bacterium *Thermotoga maritima* known as *Tm*CSP (36) (Fig. 1 *a*). *Tm*CSP is a 66-amino-acid β-barrel protein that is known to have well-defined two-state folding behavior (50–53). Through comparison with previous computational, theoretical, and experimental results, we provide evidence of the degree of roughness present in *Tm*CSP.

MODELS AND METHODS

Structure-based C_α model

Here, we employ a well-studied coarse-grained structure-based model (8). In this model, each residue is represented as a single bead, located at the position of the C_α atom (Fig. 1 *b*). For unfrustrated simulations, only native interactions are stabilizing and all residue pairs not in contact in the native structure are given a repulsive interaction to prevent chain crossing. In this model, the native structure is the global energetic minimum and the landscape lacks energetic traps. Native contacts were determined by the Contact of Structural Units software package (54). The functional form of the potential is

$$V = \sum_{\text{bonds}} \epsilon_r (r - r_o)^2 + \sum_{\text{angles}} \epsilon_\theta (\theta - \theta_o)^2 + \sum_{\text{dihedrals}} \epsilon_\phi \{[1 - \cos(\phi - \phi_o)] + \tfrac{1}{2}[1 - \cos(3(\phi - \phi_o))]\} + \sum_{\text{contacts}} \epsilon_C \left[5\left(\frac{\sigma_{ij}}{r}\right)^{12} - 6\left(\frac{\sigma_{ij}}{r}\right)^{10}\right] + \sum_{\text{noncontacts}} \epsilon_{NN}\left(\frac{\sigma_{NN}}{r}\right)^{12}, \quad (1)$$

where $\epsilon_r = 100$, $\epsilon_\theta = 20$, $\epsilon_\phi = 1$, $\epsilon_C = 1$, $\epsilon_{NN} = 1$, and $\sigma_{NN} = 4.0$ Å. r_o, θ_o, ϕ_o, and σ_{ij} are given the values found in the native structure.

To model nonspecific energetic frustration, we introduced an additional attractive interaction between all residue pairs that are not in contact in the native state and are separated by at least four residues in sequence. The functional form of the nonnative interactions is

$$V_f(r) = -\epsilon_{NC} \exp\left\{-\frac{(r - r_g)^2}{\sigma_g^2}\right\}, \quad (2)$$

with $r_g = 6.5$ Å and $\sigma_g = 1.0$ Å. The degree of energetic frustration is determined by ϵ_{NC}. In this study, we performed simulations with $\epsilon_{NC} = 0.1$–0.7.

Biasing potential

To calculate the diffusion coefficient about a specific value of Q, we introduced umbrella potentials (55,56) that restrained each simulation to a specified range of Q values. See Supporting Material for technical details.

ϕ-values analysis

Experimentally, the structural content of the transition state ensemble in proteins is often studied by measuring changes in native stability and folding/unfolding rates upon point mutations. An approximate kinetic measure of the protein structure around a mutated residue is given by (57,58)

$$\phi \equiv \frac{-RT \ln k_{\text{mut}}/k_{\text{wt}}}{\Delta\Delta G^0}, \quad (3)$$

where k_{mut} and k_{wt} are the mutant and the wild-type folding rates, and $\Delta\Delta G^0$ is the change in stability of the folded state upon mutation.

From a simulation, one may also calculate ϕ-values by determining the change in the thermodynamic free energy barrier upon site mutation and comparing it to the change in native stability $\Delta\Delta G^{F-U}$. Computationally, this is less demanding than trying to determine differences in folding rates upon mutation. The ϕ-values from structure-based simulations for each native contact pair (residues i and j) can be further approximated as (8,59)

$$\phi_{ij} = \frac{\Delta\Delta G^{TS-U}}{\Delta\Delta G^{F-U}} \approx \frac{P_{ij}^{TS} - P_{ij}^{U}}{P_{ij}^{F} - P_{ij}^{U}}, \quad (4)$$

where P_{ij}^{X} is the probability of a contact between i and j being formed in state X (with X being F, TS, or U). For ease of discussion, here, we report ϕ_i-values averaged over all native contacts with residue i.

RESULTS

Diffusion coefficient is robust to changes in restraining potential

The primary objective of this study was to determine how the diffusion coefficient D, in reaction coordinate space, changes during the folding process of *Tm*CSP (Fig. 1 *a*). To calculate D, we employed a C_α structure-based model (Fig. 1 *b*) with a restraining potential to ensure that each simulation sampled the phase space local to a particular value of Q (see Models and Methods for full description). The restraining potential was harmonic, centered at Q^*, and was given a strength of K_Q. When adding such a restraint, one must first ensure that the quantities of interest are not dependent on the strength of the restraint. To ensure that the diffusion coefficients are

a result of the underlying energy landscape, and not the biasing potential, several sets of simulations were performed, each with a different strength of the restraint.

To calculate the diffusion coefficient from a simulation, we employed a quasiharmonic diffusive approximation (25)

$$D = \frac{\Delta Q(T)^2}{2\tau(T)}, \quad (5)$$

where $\Delta Q(T)^2$ is the mean-squared fluctuations in Q, and $\tau(T)$ is the relaxation time associated with the decay of the autocorrelation function of Q, i.e., $C_Q(t)$. Here, $Q(t)$ is defined as the fraction of native C_α-C_α contacts formed as a function of time (see Models and Methods). To use Eq. 5, the value of K_Q must be in a range for which a quasiharmonic approximation is warranted and D is not dependent on K_Q.

To determine values of K_Q for which the quasiharmonic approximation is valid, we compared the probability distributions in Q for a variety of K_Q values. Fig. 2 *a* shows the probability distributions for several values of K_Q (where the harmonic restraint is centered at $Q^* = 0.5$), each at the folding temperature in the unrestrained case. For $K_Q = 10$, the probability distribution is clearly bimodal, with one peak corresponding to nativelike structures ($Q \approx 0.8$) and one peak corresponding to the unfolded ensemble ($Q \approx 0.2$). For $K_Q = 50$, the probability distribution possesses a single peak near the minimum of the restraining potential $Q = 0.5$. For $K_Q = 100$, the width of the distribution is further reduced. This additional reduction of the width is undesirable. Because the diffusion coefficient describes the multidimensional process of the protein escaping from local energetic/topological minima, an overly-strong restraint may lead to artifacts by disallowing some possible routes of escape. In that scenario, our calculations of D could probe the restraining potential and not the underlying energy landscape.

In addition to identifying a range of value of K_Q for which a quasiharmonic approximation is valid, we also determined a range of K_Q-values for which the calculated diffusion coefficients are not K_Q-dependent. Fig. 2 *b* shows D as a function of K_Q for a wide range of Q values. For low values of K_Q (< 10), all calculated values of D increase with K_Q. As discussed above, this is due to the probability distribution being altered from a bimodal distribution to a distribution centered about the Q value of interest. For $10 < K_Q < 50$ the calculated D is nearly constant for all Q values. Above $K_Q = 50$, the values of D again increase for $Q > 0.7$. Based on these data, we concluded that $K_Q = 50$ will provide reliable values for the position-dependent diffusion.

Diffusion coefficient dependence on *Q*

To understand the origins of the Q-dependence of the diffusion coefficient, one must consider the fluctuations in Q and the decay time of these fluctuations τ_Q. As our calculated values of D are not sensitive to K_Q at $K_Q \sim 50$, all further values are reported for simulations performed at the folding temperature of the unrestrained simulations T_f^0 with $K_Q = 50$. Fig. 3 *a* shows the time autocorrelation functions of Q for a variety of Q^* values. As Q^* is increased from 0.2 to 0.5 the characteristic decay time, τ_Q, increases. At higher Q values (>0.5), the decay time decreases to a value smaller than in the unfolded ensemble (larger $1/\tau_Q$ values in Fig. 3 *b*). The dispersion in Q (ΔQ^2) also displays a nonlinear dependence on Q (Fig. 3 *b*). Similar to τ_Q, ΔQ^2 initially increases with increasing Q (0.2–0.5) and then decreases as the native ensemble is reached ($Q = 0.8$). Fluctuations in Q, shown in Fig. 3 *b*, rise considerably near the transition state due to the intrinsic instability of the transition state ensemble (TSE). In other words, Q exhibits large amplitude fluctuations as it overcomes the free energy barrier. For high and low Q values (the folded and unfolded ensembles), Q is

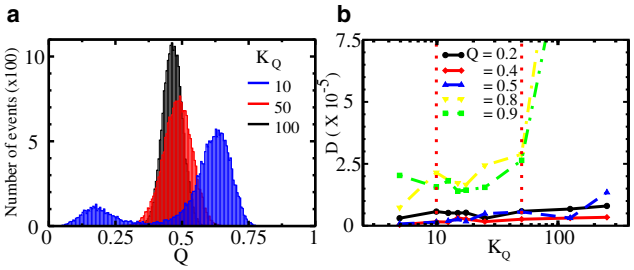


FIGURE 2 (*a*) Probability distributions in Q for biased ($K_Q > 0$, $Q^* = 0.5$) simulations with different strengths of the restraining potential K_Q. As K_Q increases, the distribution changes from a bimodal distribution, with peaks corresponding to the native and unfolded ensembles, to a single peak centered about the minimum of the restraining potential Q^*. The distribution is quasiharmonic for $K_Q > 10$. (*b*) The diffusion coefficient D is shown, on a semilog plot, as a function of K_Q for five values of Q^*. For $10 < K_Q < 50$ (region delimited by the *vertical dashed lines*) $D(Q)$ is relatively constant, demonstrating that estimates of D will be independent of K_Q over this interval. Simulations were performed at the folding temperature of the pure structure-based model T_f^0.

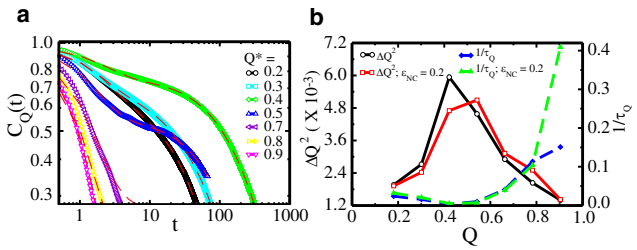


FIGURE 3 (*a*) Normalized correlation functions of Q, $C_Q(t)$, shown on a log-log plot (time in reduced units) for different values of Q^*. Because a single exponential did not always fit well, each curve was fit to the sum of three exponentials to obtain an average decay time τ_Q. $K_Q = 50$ was used and the temperature was the folding temperature of the unbiased simulations T_f^0. The characteristic decay time τ_Q is used to calculate the diffusion coefficient D. (*b*)The dispersion of the reaction coordinate ΔQ^2 as a function of the reaction coordinate Q (*left axis*), with and without energetic frustration. The inverse correlation time of Q ($1/\tau_Q$) shown as a function of Q (*right axis*), with and without the frustration term. Calculations are shown at T_f of each ϵ_{NC}.

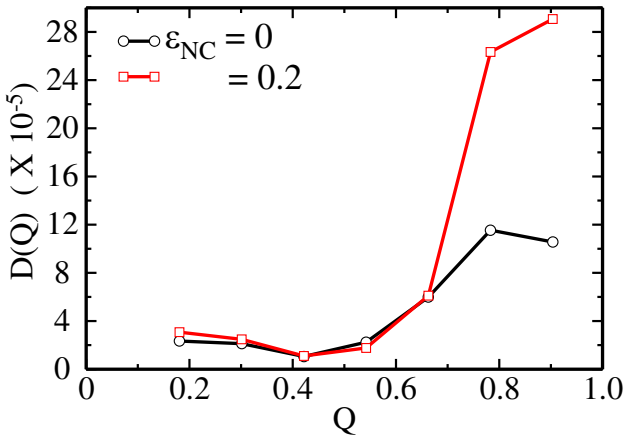


FIGURE 4 Comparison of the diffusion coefficients calculated without energetic frustration (*black circles*) and with energetic frustration ($\epsilon_{NC} = 0.2$, *red squares*) as functions of Q. All results were calculated with a restraining potential of strength $K_Q = 50$ and at the T_f of each ϵ_{NC}.

highly localized, which results in a small dispersion in Q. While ΔQ^2 is roughly symmetric about the TSE, the autocorrelation time, as well as its inverse, is asymmetric. This symmetry-breaking in τ_Q leads to an asymmetric relationship between D and Q.

The diffusion coefficient as a function of Q, $D(Q)$, is proportional to the product of ΔQ^2 and $1/\tau_Q$, and is shown in Fig. 4. We find D has large variations as a function of the folding reaction, which is in agreement with earlier studies on lattice models (32) and analytic studies (30,31,33,43). As discussed above, $D(Q)$ (Fig. 4) largely follows $1/\tau_Q$ (Fig. 3 *b*). Fluctuations of the reaction coordinate ΔQ^2 appear to have less influence on $D(Q)$, as ΔQ^2 changes only modestly with Q. The fact that $D(Q)$ reaches a minimum around the TSE suggests the presence of a kinetic barrier, in addition to a thermodynamic one. After the protein moves from the TSE to the folded state, $D(Q)$ once again increases and eventually reaches values that are 10-times larger than those corresponding to the unfolded ensemble.

The one-dimensional position-dependent diffusion coefficient variations indicate that the ruggedness of the energy landscape is not the same over the one-dimensional configuration space. $D(Q)$ describes the local moves over microscopic barriers that connect states with similar values of Q. If the microstate is deep, it acts like a speed bump slowing both the drift and the superimposed Brownian movement (60) (i.e., the diffusion coefficient becomes small and escape-time from traps increases (7)). Because our energetically unfrustrated model gives rise to values of D that vary with Q, our results clearly indicate that the topology of the ensemble about a particular value of Q is inextricably linked to the diffusive dynamics. In other words, each configuration of the protein has a particular set of accessible escape routes, independent of the energetic roughness, which lead to the nonconstant form of $D(Q)$.

Energetic frustration alters the diffusive dynamics

Due to the funnel-like nature of protein-folding energy landscapes, completely unfrustrated models, such as the one employed in this study, are sufficient to capture many aspects of protein folding (7,9,59,62–73). However, there is mounting evidence that a low degree of frustration can lead to accelerated folding rates (63) and provide a more accurate description of the unfolded ensemble (74). Such findings suggest a potential influence of energetic frustration on the diffusive properties associated with protein folding. To investigate this further, we employed a modified structure-based model in which the degree of frustration may be controlled. Specifically, we used the structure-based C_α-model and added nonspecific attractive interactions between all nonnative atom pairs, where the functional form is a Gaussian with an energetic weight ϵ_{NC} (see Models and Methods). Accordingly, $\epsilon_{NC} = 0$ corresponds to the purely structure-based model.

Thermodynamic quantities were calculated for each frustrated system ($\epsilon_{NC} > 0.0$) with $K_Q = 0$. For each parameter set, the fraction of native proteins $f_N(T)$ was defined as

$$f_N(T) = \frac{\int_{\text{native}} \exp[-F(Q)/k_B T]dQ}{\int_0^1 \exp[-F(Q)/k_B T]dQ}, \quad (6)$$

where $F(Q)$ is the free energy as a function of Q, the integral in the numerator is over all native conformations, and the denominator is over all possible Q values. We define the folding temperature T_f as the temperature where $f_N = 0.5$ (*dotted horizontal line* in Fig. 5 *a*). As the degree of frustration is increased from 0, T_f initially increases and reaches its maximum at $\epsilon_{NC} = 0.2$. Because T_f measures thermodynamic stability, an increase in native-state stability with increased nonnative interaction strength may be surprising. This feature has two origins. First, in the native state ensemble, proteins are constantly fluctuating (75), which allows nonnative residue pairs to fluctuate toward and away from each other and form transient nonnative interactions (10). When nonspecific interactions are stabilizing, these transient nonnative interactions increase the stability of near-native conformations. The second contribution to

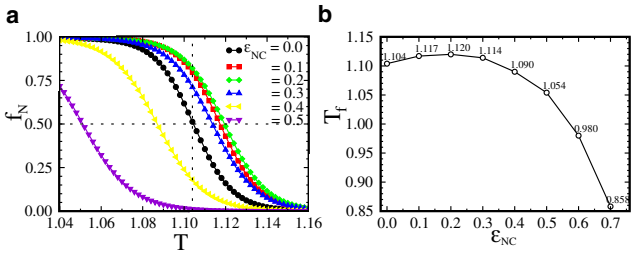


FIGURE 5 (*a*) Fraction of native protein f_N versus temperature T for different degrees of frustration ϵ_{NC}. The value f_N was calculated using Eq. 6. (*b*) Folding temperature T_f as a function of ϵ_{NC} for $K_Q = 0$. The folding temperature T_f has a maximum at $\epsilon_{NC} = 0.2$.

the increased native-state stability may be due to using a C_α representation. All nonnative interactions were given energetic minima at 6.5 Å. When coarse-graining, C_α pairs may be within that distance, but the side-chain configurations may lead to these pairs being considered not-in-contact. Thus, the noncontacting residues may stabilize these native configurations via the nonspecific interactions. Above $\epsilon_{NC} = 0.2$, the energetic frustration stabilizes the unfolded ensemble more than the folded ensemble and the folding temperature decreases, as expected. These findings are consistent with experimental results indicating that weakly attractive nonspecific interactions can increase the stability of Src homology 3 domain (70). In addition to affecting native stability, these experiments also revealed variations in the thermodynamic properties of the transition state ensemble, which were manifested as increased unfolding and refolding rates.

Increased levels of frustration also have direct effects on the calculated $D(Q)$ profiles (Fig. 4). Similar to the unfrustrated simulations, the majority of the changes in $D(Q)$ may be attributed to fluctuations in τ_Q (Fig. 3). When energetic frustration is introduced, ΔQ^2 is only marginally perturbed while $1/\tau_Q$ exhibits substantial deviations (Fig. 3). Comparison of the frustrated and unfrustrated simulations (Fig. 3) indicates that frustration has little effect on the ΔQ^2 and $1/\tau_Q$ values associated with the unfolded ensemble ($Q < 0.5$). After the protein has reached the folding transition state ($Q \approx 0.5$) and moves to higher Q values, $1/\tau_Q$ increases for both the unfrustrated and $\epsilon_{NC} = 0.2$ simulations, although there is a larger increase in $1/\tau_Q$ for the frustrated simulations than the unfrustrated ones. This finding may be counterintuitive, but it shows that a low degree of frustration can actually reduce the height of the microscopic barriers that are described by the diffusion coefficient.

Residual entropy of the native state ensemble

The fact that $D(Q)$ reaches a maximum in the folded ensemble can be understood by analyzing the density of states as a function of Q. Fig. 6 shows the density of states as a function Q for a lattice model (32) and the presented C_α model. Although there is an increase in $D(Q)$ as Q goes to 1, $D(Q)$ remains on the same scale for large Q as for small Q. This is due to the ensemble nature of the native state. That is, in the C_α model, the protein may interconvert between local structures without changing the value of Q, even when all native contacts are formed ($Q = 1$). This leads to a degenerate native state, residual entropy, and nonzero correlation times. In contrast, in the lattice model, every possible move from the $Q = 1$ state results in a decrease in Q. This nondegenerate native state leads to very low correlation times, and hence very large diffusion coefficients for the native state. Additionally, in the lattice model, single rearrangements can result in multiple contacts being formed or broken simultaneously. This lack of residual entropy in the lattice model has prevented previous evaluation of $D(Q = 1)$. Here, by using an off-lattice C_α model, we are able to calculate $D(Q)$ for the full range of Q.

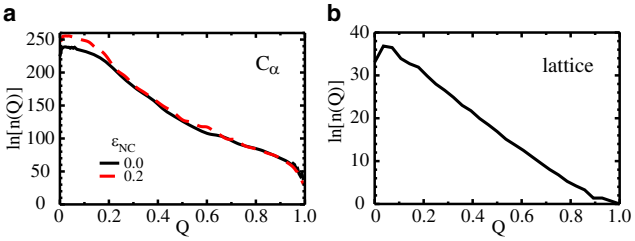


FIGURE 6 Density of states n as a function of Q for (a) the C_α model and (b) a lattice model (32). High Q represents the folded ensemble and low Q is the unfolded ensemble. Panel a shows the density of states for the unfrustrated ($\epsilon_{NC} = 0.0$) and frustrated ($\epsilon_{NC} = 0.2$) systems. The C_α model has a highly degenerate folded state, indicating the presence of residual entropy which allows for an increase in D after the protein passes through the TSE. For the lattice model there is a rapid decrease in the density of states with a nondegenerate folded state ($n(E_{\text{folded}}) = n(Q_{\text{folded}}) = 1$) and monotonically decreasing values of D (32).

Folding mechanism and ϕ-values analysis

The introduction of attractive nonnative interactions, or energetic frustration, changes the folding energy landscape (62,63,76) and can alter the structural content of the transition state ensemble. The ϵ-values are commonly used experimentally to measure the degree of native structural content in the TSE about each residue. Computationally, ϕ-values can also be determined, where a value of 0 indicates no native structural content and 1 indicates full structural content in the TSE. Fig. 7 shows ϕ_i (ϕ-value for each residue i) obtained from simulations with no frustration ($\epsilon_{NC} = 0.0$) and a low degree of frustration ($\epsilon_{NC} = 0.2$). When $\epsilon_{NC} = 0.2$,

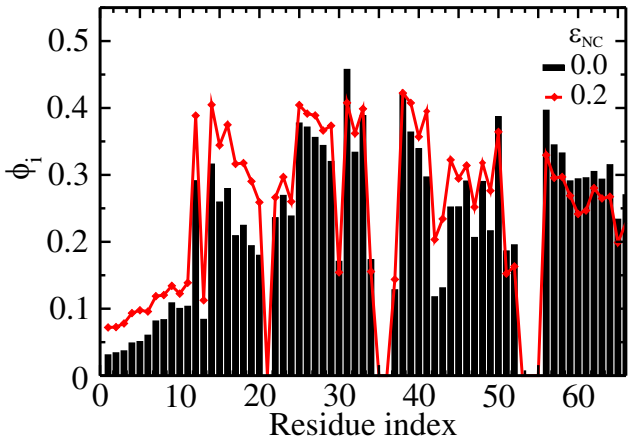


FIGURE 7 The ϕ-values calculated for the unfrustrated simulations ($\epsilon_{NC} = 0.0$, *black bars*) and a weakly-frustrated system ($\epsilon_{NC} = 0.2$, *red diamonds*). Many residues with lower ϕ_i-values increase, and high values decrease, upon the addition of frustration. This indicates a more homogeneous TSE when a low level of frustration is introduced. Simulations were performed without a restraining potential ($K_Q = 0$) and at $T_f(\epsilon_{NC})$.

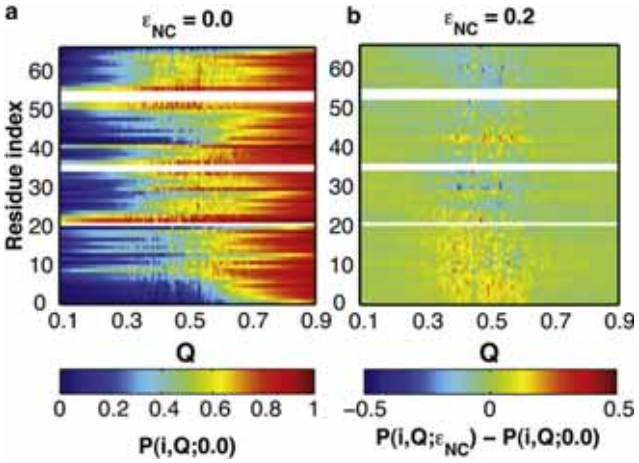


FIGURE 8 Probability of contacts being formed for each residue $P(i, Q, \epsilon_{NC})$ as a function of the reaction coordinate Q, for the unfrustrated and weakly frustrated system. (*a*) $P(i, Q, 0)$ increases (*blue* to *red*) as the protein folds. (*b*) Difference between the probabilities for the frustrated and unfrustrated simulations, $P(i, Q, \epsilon_{NC}) - P(i, Q, 0.0)$, where $\epsilon_{NC} = 0.2$. The simulations were performed at $T = T_f$. (*Open bars*) Residues lacking any native contacts.

there is a shift in structural content from the C-terminus to the N-terminus. Specifically, when ϵ_{NC} is increased from 0.0 to 0.2 the ϕ-values for residues 0–20 increase and residues 55–65 decrease. Additionally, other regions with low ϕ-values, such as residues 42–45, increase when frustration is introduced.

Overall, introducing a low-degree of frustration appears to (partially) homogenize the ϕ-values. Because residues that are less native (low ϕ-values) are not surrounded by formed native interactions, they are more exposed to nonnative interactions. Accordingly, nonspecific stabilizing interactions are more accessible to less-native residues than highly-native ones. These nonspecific attractive interactions can then localize the residues involved in native-contacts, which results in additional native structure formation and a shift in the ϕ-values.

To characterize the effects of frustration on the unfolded and folded basins, in addition to the TSE, we calculated the probability of contacts being formed with each residue, as functions of Q (Fig. 8). The probabilities for the unfrustrated case are shown in Fig. 8 *a*. For low Q, the probabilities are not homogeneously distributed, but are high around residues 22, 53, and 56, and nearly zero for all other residues. At approximately the transition state ($Q = 0.5$), the probabilities follow the ϕ-values, and have peaks around residues 22, 35, and 55. After passing the transition state, these regions may be considered nucleation sites, about which the rest of the protein's native structure is formed. Fig. 8 *b* shows the changes in the probabilities when frustration is introduced. Blue corresponds to decreased structure and red indicates increased structure formation. Similar to the ϕ-values, in the TSE there is a shift in probabilities from the C-terminal residues to the N-terminal residues. Surprisingly, the effects of the energetic frustration appear to be isolated to the TSE. One explanation for this feature is that frustration in the unfolded ensemble may not be well described by Gaussian potentials, as we have employed here. Instead, longer-range, screened-electrostatic interactions may be a larger contributor to frustration in the unfolded ensemble (74). In contrast, using a coarse-grained structure-based model, Das et al. (77) showed that introducing nonnative interactions and energetic heterogeneity has a large effect on the TSE, and improves agreement between experimental and theoretical ϕ-values for Src homology 3 domain. Although our finding suggest short-range frustration is most important in the TSE, real proteins likely exhibit a combination of short-range and long-range nonnative interactions. Further investigation will be necessary to untangle the relationship among different types of frustration, the folding mechanism, and the diffusive dynamics of the folding process.

Perl et al. (78) explored the role of the chain termini residues on the folding stability by comparing the cold shock proteins *Bc*CSP from the thermophile *Bacillus caldolyticus* with its homolog *Bs*CSPB *Bacillus subtilis*. These two cold shock proteins have nativelike activated states of folding, similar to that of the hyperthermophilic *Thermotoga maritima Tm*CSP (50) studied in this work. Their studies illustrate that major contributors to the difference in stability are residue 3 (which takes on nativelike structure in the TSE) and the C-terminal residue 66 (which forms late in the folding process) (78). Despite the fact that the C-terminal residues have high ϕ-values in experiments and low ϕ-values for the unfrustrated model, as discussed above, the ϕ-values of the termini increase with increased energetic roughness. As suggested by our analysis of stability as a function of roughness, this comparison also demonstrates that cold shock protein likely has a modest degree of energetic roughness, though the exact degree, and type, of frustration cannot be unambiguously determined from the presented simulations.

CONCLUSIONS

In this work, we have studied the folding of *Tm*CSP using a coarse-grained structure-based model and we calculated the diffusion coefficient as a function of a reaction coordinate Q. Our main results can be outlined as follows: The diffusion coefficient displays nonmonotonic behavior as a function of Q, which can be attributed to a residual entropy of the native state ensemble. A role of residual entropy has been suggested previously (33), though here we explicitly calculate it and show its relationship to the diffusive dynamics. By introducing varied degrees of energetic roughness, we have shown that for low levels of frustration, *Tm*CSP displays increased thermal stability and diffusion coefficients, relative to the unfrustrated regime, which agrees with previous findings (62,63). As frustration is increased, the stability reaches

a maximum, after which increased frustration leads to a less stable protein. At this optimum degree of frustration, the transition state is characterized by a more homogeneous distribution of ϕ-values, relative to the unfrustrated case. In addition to changes in the structural content of the TSE, the diffusion coefficient is also affected by a low-degree of energetic frustration.

This work has shown that the diffusive dynamics are intimately linked to the topological and energetic aspects of a protein, and lays a foundation for understanding the diffusive properties of protein folding. Many examples can be found where the diffusion coefficient provides a nontrivial contribution to the folding dynamics. For example, as the presence of additional small free-energy barriers can actually accelerate folding rates (79,80), there must be a balance between folding barriers and diffusion along the reaction coordinate. The folding of proteins with smaller free-energy barriers, such as BBL (81), will also depend more on the precise structure of the diffusion coefficient (44,41,42), such that the nonmonotonic behavior of the diffusion may be the limiting factor that determines folding rates. With the presented framework, further investigation will explore the details of how the diffusive dynamics contributes to the folding of these and other systems.

SUPPORTING MATERIAL

Additional text and one reference is available at http://www.biophysj.org/biophysj/supplemental/S0006-3495(10)00539-4.

We thank the Center for Theoretical Biological Physics for use of its computing facilities.

This work was supported by the Center for Theoretical Biological Physics sponsored by the National Science Foundation (grant No. PHY-0822283, with additional support from NSF grant No. MCB-0543906). R.J.O., J.C., and V.B.P.L. were supported by Fundação de Amparo à Pesquisa do Estado dãe So Paulo and Conselho Nacional de Desnvolvimento Científico e Tecnológico, Brazil. R.J.O. and J.C. have also been supported by the Fundacao Coordenação de Aperfeiçoamento de Pessoal de Nivel Superior, Brazilian Ministry of Education. P.C.W. thanks the U. S. National Science Foundation for its I2CAM International Materials Institute Award (grant No. DMR-0645461) to fund this international collaboration. J.W. thanks the National Science Foundation for its Career Award.

REFERENCES

1. Leopold, P. E., M. Montal, and J. N. Onuchic. 1992. Protein folding funnels—a kinetic approach to the sequence structure relationship. *Proc. Natl. Acad. Sci. USA.* 18:8721–8725.
2. Bryngelson, J. D., J. N. Onuchic, …, P. G. Wolynes. 1995. Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins.* 21:167–195.
3. Bryngelson, J. D., and P. G. Wolynes. 1987. Spin glasses and the statistical mechanics of protein folding. *Proc. Natl. Acad. Sci. USA.* 84:7524–7528.
4. Onuchic, J. N., and P. G. Wolynes. 2004. Theory of protein folding. *Curr. Opin. Struct. Biol.* 14:70–75.
5. Wang, J., J. N. Onuchic, and P. G. Wolynes. 1996. Statistics of kinetic pathways on biased rough energy landscapes with applications to protein folding. *Phys. Rev. Lett.* 76:4861–4864.
6. Shoemaker, B. A., J. Wang, and P. G. Wolynes. 1997. Structural correlations in protein folding funnels. *Proc. Natl. Acad. Sci. USA.* 94: 777–782.
7. Nymeyer, H., A. E. García, and J. N. Onuchic. 1998. Folding funnels and frustration in off-lattice minimalist protein landscapes. *Proc. Natl. Acad. Sci. USA.* 95:5921–5928.
8. Clementi, C., H. Nymeyer, and J. N. Onuchic. 2000. 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.* 298:937–953.
9. Gosavi, S., L. L. Chavez, …, J. N. Onuchic. 2006. Topological frustration and the folding of interleukin-1 β. *J. Mol. Biol.* 357:986–996.
10. Whitford, P. C., J. K. Noel, …, J. N. Onuchic. 2009. An all-atom structure-based potential for proteins: bridging minimal models with all-atom empirical forcefields. *Proteins Struct. Funct. Bioinf.* 75:430–441.
11. Levy, Y., and J. N. Onuchic. 2006. Mechanisms of protein assembly: lessons from minimalist models. *Acc. Chem. Res.* 39:135–142.
12. Yang, S. C., S. S. Cho, …, J. N. Onuchic. 2004. Domain swapping is a consequence of minimal frustration. *Proc. Natl. Acad. Sci. USA.* 101:13786–13791.
13. Lu, Q., H. P. Lu, and J. Wang. 2007. Exploring the mechanism of flexible biomolecular recognition with single molecule dynamics. *Phys. Rev. Lett.* 98:128105.
14. Whitford, P. C., O. Miyashita, …, J. N. Onuchic. 2007. Conformational transitions of adenylate kinase: switching by cracking. *J. Mol. Biol.* 366:1661–1671.
15. Schug, A., P. C. Whitford, …, J. N. Onuchic. 2007. Mutations as trapdoors to two competing native conformations of the Rop-dimer. *Proc. Natl. Acad. Sci. USA.* 104:17674–17679.
16. Best, R. B., Y. G. Chen, and G. Hummer. 2005. Slow protein conformational dynamics from multiple experimental structures: the helix/sheet transition of arc repressor. *Structure.* 13:1755–1763.
17. Zuckerman, D. M. 2004. Simulation of an ensemble of conformational transitions in a united-residue model of calmodulin. *J. Phys. Chem. B.* 108:5127–5137.
18. Lu, Q., and J. Wang. 2008. Single molecule conformational dynamics of adenylate kinase: energy landscape, structural correlations, and transition state ensembles. *J. Am. Chem. Soc.* 130:4772–4783.
19. Pincus, D. L., S. S. Cho, …, D. Thirumalai. 2008. Minimal models for proteins and RNA: from folding to function. *In* Molecular Biology of Protein Folding, Pt B, Vol. 84, Progress in Molecular Biology and Translational Science.. Elsevier/Academic Press, Amsterdam, The Netherlands.
20. Mickler, M., R. I. Dima, …, M. Rief. 2007. Revealing the bifurcation in the unfolding pathways of GFP by using single-molecule experiments and simulations. *Proc. Natl. Acad. Sci. USA.* 104:20268–20273.
21. Hills, Jr., R. D., and C. L. Brooks, III. 2009. Insights from coarse-grained Gō models for protein folding and dynamics. *Int. J. Mol. Sci.* 10:889–905.
22. Karanicolas, J., and C. L. Brooks, 3rd. 2003. Improved Gō-like models demonstrate the robustness of protein folding mechanisms towards non-native interactions. *J. Mol. Biol.* 334:309–325.
23. Ueda, Y., H. Taketomi, and N. Gō. 1975. Studies on protein folding, unfolding and fluctuations by computer simulation. I. The effects of specific amino acid sequence represented by specific inter-unit interactions. *Int. J. Peptide Res.* 7:445–459.
24. Chavez, L. L., J. N. Onuchic, and C. Clementi. 2004. Quantifying the roughness on the free energy landscape: entropic bottlenecks and protein folding rates. *J. Am. Chem. Soc.* 126:8426–8432.
25. Socci, N. D., J. N. Onuchic, and P. G. Wolynes. 1996. Diffusive dynamics of the reaction coordinate for protein folding funnels. *J. Chem. Phys.* 104:5860–5868.

26. Baumketner, A., and Y. Hiwatari. 2002. Diffusive dynamics of protein folding studied by molecular dynamics simulations of an off-lattice model. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 66:011905.
27. Pande, V. S., A. Y. Grosberg, and T. Tanaka. 1997. On the theory of folding kinetics for short proteins. *Fold. Des.* 2:109–114.
28. Du, R., V. S. Pande, …, E. S. Shakhnovich. 1998. On the transition coordinate for protein folding. *J. Chem. Phys.* 108:334–350.
29. Pande, V. S., and D. S. Rokhsar. 1999. Molecular dynamics simulations of unfolding and refolding of a β-hairpin fragment of protein G. *Proc. Natl. Acad. Sci. USA.* 96:9062–9067.
30. Lee, C. L., G. Stell, and J. Wang. 2003. First-passage time distribution and non-Markovian diffusion dynamics of protein folding. *J. Chem. Phys.* 118:959–968.
31. Lee, C. L., C. T. Lin, …, J. Wang. 2003. Diffusion dynamics, moments, and distribution of first-passage time on the protein-folding energy landscape, with applications to single molecules. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 67:041905.
32. Chahine, J., R. J. Oliveira, …, J. Wang. 2007. Configuration-dependent diffusion can shift the kinetic transition state and barrier height of protein folding. *Proc. Natl. Acad. Sci. USA.* 104:14646–14651.
33. Yang, S., J. N. Onuchic, and H. Levine. 2006. Effective stochastic dynamics on a protein folding energy landscape. *J. Chem. Phys.* 125:054910–054918.
34. Yang, S., J. N. Onuchic, …, H. Levine. 2007. Folding time predictions from all-atom replica exchange simulations. *J. Mol. Biol.* 372:756–763.
35. Cho, S. S., Y. Levy, and P. G. Wolynes. 2006. P versus Q: structural reaction coordinates capture protein folding on smooth landscapes. *Proc. Natl. Acad. Sci. USA.* 103:586–591.
36. Kremer, W., B. Schuler, …, H. R. Kalbitzer. 2001. Solution NMR structure of the cold-shock protein from the hyperthermophilic bacterium *Thermotoga maritima*. *Eur. J. Biochem.* 268:2527–2539.
37. Nettels, D., I. V. Gopich, …, B. Schuler. 2007. Ultrafast dynamics of protein collapse from single-molecule photon statistics. *Proc. Natl. Acad. Sci. USA.* 104:2655–2660.
38. Hoffmann, A., A. Kane, …, B. Schuler. 2007. Mapping protein collapse with single-molecule fluorescence and kinetic synchrotron radiation circular dichroism spectroscopy. *Proc. Natl. Acad. Sci. USA.* 104:105–110.
39. Nettels, D., I. V. Gopich, …, B. Schuler. 2008. Unfolded protein and peptide dynamics investigated with single-molecule FRET and correlation spectroscopy from picoseconds to seconds. *J. Phys. Chem.* 112:6137–6146.
40. Karplus, M., and D. L. Weaver. 1994. Protein folding dynamics: the diffusion-collision model and experimental data. *Protein Sci.* 3:650–668.
41. Kubelka, J., T. K. Chiu, …, J. Hofrichter. 2006. Sub-microsecond protein folding. *J. Mol. Biol.* 359:546–553.
42. Gruebele, M. 2005. Downhill protein folding: evolution meets physics. *C. R. Biol.* 328:701–712.
43. Bryngelson, J. D., and P. G. Wolynes. 1989. Intermediates and barrier crossing in a random energy-model (with applications to protein folding). *J. Phys. Chem.* 93:6902–6915.
44. Kubelka, J., J. Hofrichter, and W. A. Eaton. 2004. The protein folding 'speed limit'. *Curr. Opin. Struct. Biol.* 14:76–88.
45. Oliveberg, M., and P. G. Wolynes. 2005. The experimental survey of protein-folding energy landscapes. *Q. Rev. Biophys.* 38:245–288.
46. Hummer, G. 2005. Position-dependent diffusion coefficients and free energies from Bayesian analysis of equilibrium and replica molecular dynamics simulations. *N. J. Phys.* 7:34–48.
47. Best, R. B., and G. Hummer. 2010. Coordinate-dependent diffusion in protein folding. *Proc. Natl. Acad. Sci. USA.* 107:1088–1093.
48. Nettels, D., S. Müller-Späth, …, B. Schuler. 2009. Single-molecule spectroscopy of the temperature-induced collapse of unfolded proteins. *Proc. Natl. Acad. Sci. USA.* 106:20740–20745.
49. Krivov, S. V., and M. Karplus. 2008. Diffusive reaction dynamics on invariant free energy profiles. *Proc. Natl. Acad. Sci. USA.* 105: 13841–13846.
50. Perl, D., C. Welker, …, F. X. Schmid. 1998. Conservation of rapid two-state folding in mesophilic, thermophilic and hyperthermophilic cold shock proteins. *Nat. Struct. Biol.* 5:229–235.
51. Wassenberg, D., C. Welker, and R. Jaenicke. 1999. Thermodynamics of the unfolding of the cold-shock protein from *Thermotoga maritima*. *J. Mol. Biol.* 289:187–193.
52. Schuler, B., W. Kremer, …, R. Jaenicke. 2002. Role of entropy in protein thermostability: folding kinetics of a hyperthermophilic cold shock protein at high temperatures using ^{19}F NMR. *Biochemistry.* 41:11670–11680.
53. Schuler, B., E. A. Lipman, and W. A. Eaton. 2002. Probing the free-energy surface for protein folding with single-molecule fluorescence spectroscopy. *Nature.* 429:743–747.
54. Sobolev, V., R. Wade, …, M. Edelman. 1996. Molecular docking using surface complementarity. *Proteins Struct. Funct. Genet.* 25:120–129.
55. Torrie, G. M., and J. P. Valleau. 1977. Nonphysical sampling distributions in Monte Carlo free-energy estimation—umbrella sampling. *J. Comput. Phys.* 23:187–199.
56. Bartels, C., and M. Karplus. 1997. Multidimensional adaptive umbrella sampling: applications to main chain and side chain peptide conformations. *J. Comput. Phys.* 18:1450–1462.
57. Matouschek, A., J. T. Kellis, Jr., …, A. R. Fersht. 1989. Mapping the transition state and pathway of protein folding by protein engineering. *Nature.* 340:122–126.
58. Fersht, A. R. 1995. Characterizing transition states in protein folding: an essential step in the puzzle. *Curr. Opin. Struct. Biol.* 5:79–84.
59. Nymeyer, H., N. D. Socci, and J. N. Onuchic. 2000. Landscape approaches for determining the ensemble of folding transition states: success and failure hinge on the degree of frustration. *Proc. Natl. Acad. Sci. USA.* 97:634–639.
60. Onuchic, J. N., Z. Luthey-Schulten, and P. G. Wolynes. 1997. Theory of protein folding: the energy landscape perspective. *Annu. Rev. Phys. Chem.* 48:545–600.
61. Reference deleted in proof.
62. Plotkin, S. S. 2001. Speeding protein folding beyond the Gō model: how a little frustration sometimes helps. *Proteins Struct. Funct. Genet.* 45:337–345.
63. Clementi, C., and S. S. Plotkin. 2004. The effects of nonnative interactions on protein folding rates: theory and simulation. *Protein Sci.* 13:1750–1766.
64. Fan, K., J. Wang, and W. Wang. 2002. Folding of lattice protein chains with modified Gō potential. *Eur. Phys. J. B.* 30:381–391.
65. Li, L., L. A. Mirny, and E. I. Shakhnovich. 2000. Kinetics, thermodynamics and evolution of non-native interactions in a protein folding nucleus. *Nature.* 7:336–342.
66. Treptow, W. L., M. A. A. Barbosa, …, A. F. P. de Araújo. 2002. Non-native interactions, effective contact order, and protein folding: a mutational investigation with the energetically frustrated hydrophobic model. *Proteins Struct. Funct. Bioinf.* 49:167–180.
67. Garcia, L. G., and A. F. P. de Araújo. 2006. Folding pathway dependence on energetic frustration and interaction heterogeneity for a three-dimensional hydrophobic protein model. *Proteins Struct. Funct. Bioinf.* 62:46–63.
68. Morton, V. L., C. T. Friel, …, S. E. Radford. 2007. The effect of increasing the stability of non-native interactions on the folding landscape of the bacterial immunity protein Im9. *J. Mol. Biol.* 371:554–568.
69. Hamada, D., and Y. Goto. 1997. The equilibrium intermediate of β-lactoglobulin with non-native α-helical structure. *J. Mol. Biol.* 269:479–487.
70. Viguera, A. R., C. Vega, and L. Serrano. 2002. Unspecific hydrophobic stabilization of folding transition states. *Proc. Natl. Acad. Sci. USA.* 99:5349–5354.

71. Di Nardo, A. A., D. M. Korzhnev, …, A. R. Davidson. 2004. Dramatic acceleration of protein folding by stabilization of a nonnative backbone conformation. *Proc. Natl. Acad. Sci. USA.* 101:7954–7959.

72. Neudecker, P., A. Zarrine-Afsar, …, L. E. Kay. 2006. Identification of a collapsed intermediate with non-native long-range interactions on the folding pathway of a pair of Fyn SH3 domain mutants by NMR relaxation dispersion spectroscopy. *J. Mol. Biol.* 363:958–976.

73. Shan, B., D. Eliezer, and D. P. Raleigh. 2009. The unfolded state of the C-terminal domain of the ribosomal protein L9 contains both native and non-native structure. *Biochemistry.* 48:4707–4719.

74. Weinkam, P., E. V. Pletneva, …, P. G. Wolynes. 2009. Electrostatic effects on funneled landscapes and structural diversity in denatured protein ensembles. *Proc. Natl. Acad. Sci. USA.* 106:1796–1801.

75. Frauenfelder, H., S. G. Sligar, and P. G. Wolynes. 1991. The energy landscapes and motions of proteins. *Science.* 254:1598–1603.

76. Paci, E., M. Vendruscolo, and M. Karplus. 2002. Native and non-native interactions along protein folding and unfolding pathways. *Proteins Struct. Funct. Genet.* 47:379–392.

77. Das, P., S. Matysiak, and C. Clementi. 2005. Balancing energy and entropy: a minimalist model for the characterization of protein folding landscapes. *Proc. Natl. Acad. Sci. USA.* 102:10141–10146.

78. Perl, D., G. Holtermann, and F. X. Schmid. 2001. Role of the chain termini for the folding transition state of the cold shock protein. *Biochemistry.* 40:15501–15511.

79. Wagner, C., and T. Kiefhaber. 1999. Intermediates can accelerate protein folding. *Proc. Natl. Acad. Sci. USA.* 96:6716–6721.

80. Sinha, K. K., and J. B. Udgaonkar. 2008. Barrierless evolution of structure during the submillisecond refolding reaction of a small protein. *Proc. Natl. Acad. Sci. USA.* 105:7998–8003.

81. Cho, S. S., P. Weinkam, and P. G. Wolynes. 2008. Origins of barriers and barrierless folding in BBL. *Proc. Natl. Acad. Sci. USA.* 105: 118–123.

82. Humphrey, W., A. Dalke, and K. Schulten. 1996. VMD: visual molecular dynamics. *J. Mol. Graph.* 14: 33–38, 27–28.

Biophysical Journal, Volume 99

Supporting Material

The Origin of Nonmonotonic Complex Behavior and the Effects of Nonnative Interactions on the Diffusive Properties of Protein Folding

Ronaldo J. Oliveira, Paul C. Whitford, Jorge Chahine, Jin Wang, José N. Onuchic, and Vitor B. P. Leite

Supporting information for "Exploring the origin of non-monotonic complex behavior and the effects of non-native interactions on the diffusive properties of protein folding". Oliveira et al.

Biasing potential

For small, single-domain, proteins it has been shown that the fraction of native contacts (Q) is a reliable reaction coordinate for probing the folding dynamics [1]. Since each contact is either formed, or not, Q can only take a discrete set of values and any function of Q will be discontinuous. Since molecular dynamics simulations require that all energetic terms have a defined first derivative, the umbrella potential was introduced in terms of a modified definition of Q, Q_{mod}. We defined Q_{mod} as

$$Q_{mod} = \frac{1}{N_Q} \sum_{contacts} \frac{1}{2} \{1 - \tanh[C(r - 1.2\sigma_{ij})]\} \quad (1)$$

where σ_{ij} represents the native $C_\alpha - C_\alpha$ distances, C=10Å^{-1} and N_Q is the total number of native contacts. The tanh function is effectively a step-function centered about $1.2\sigma_{ij}$ (which is commonly designated as the cut-off distance which defines residues as "in-contact"). The correlation between the continuous Q and discrete Q is high, with a correlation coefficient $r > 0.999$ (not shown). With a continuous definition of Q, the biasing potential was included as

$$V_{bias}(Q_{mod}) = K_Q(Q_{mod} - Q^*)^2 \quad (2)$$

where K_Q is the strength of the bias included and Q^* is the region of interest.

References

[1] S.S. Cho, Y. Levy, and P.G. Wolynes. P versus Q: Structural reaction coordinates capture protein folding on smooth landscapes. *Proc. Natl. Acad. Sci. USA*, 103(3):586–591, 2006.

Apêndice B

Artigo publicado na revista *Methods*

O artigo, "*Coordinate and time-dependent diffusion dynamics in protein folding*", publicado na revista científica *Methods*, refere-se aos resultados obtidos durante o estágio no exterior com o grupo do Prof. Dr. Jin Wang[1]. O artigo também apresenta parte dos resultados obtidos durante o trabalho de doutorado no Brasil.

[1]Universidade e endereço para contato se encontram no artigo que segue nesse apêndice B.

Methods 52 (2010) 91–98

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Review Article

Coordinate and time-dependent diffusion dynamics in protein folding

Ronaldo J. Oliveira [a], Paul C. Whitford [b,c], Jorge Chahine [a], Vitor B.P. Leite [a], Jin Wang [d,e,*]

[a] Departamento de Física – Instituto de Biociências, Letras e Ciências Exatas, Universidade Estadual Paulista, São José do Rio Preto 15054-000, Brazil
[b] Theoretical Biology and Biophysics Group, Theoretical Division, Los Alamos National Laboratory, Mail Stop K710, T-6, Los Alamos, NM 87545, USA
[c] International Institute for Complex Adaptive Matter, University of California at Davis, Davis, CA 95616, USA
[d] Department of Chemistry and Department of Physics, State University of New York at Stony Brook, Stony Brook, NY 11794, USA
[e] State Key Laboratory of Electroanalytical Chemistry, Changchun Institute of Applied Chemistry, Chinese Academy of Sciences, Changchun, Jilin 130021, People's Republic of China

ARTICLE INFO

Article history:
Accepted 28 April 2010
Available online 11 May 2010

Keywords:
Position dependent diffusion
Time-dependent diffusion
Transition state
Mean first-passage time
Cold shock protein
Single molecule
Molecular dynamic simulation

ABSTRACT

We developed both analytical and simulation methods to explore the diffusion dynamics in protein folding. We found the diffusion as a quantitative measure of escape from local traps along the protein folding funnel with chosen reaction coordinates has two remarkable effects on kinetics. At a fixed coordinate, local escape time depends on the distribution of barriers around it, therefore the diffusion is often time distributed. On the other hand, the environments (local escape barriers) change along the coordinates, therefore diffusion is coordinate dependent. The effects of time-dependent diffusion on folding can lead to non-exponential kinetics and non-Poisson statistics of folding time distribution. The effects of coordinate dependent diffusion on folding can lead to the change of the kinetic barrier height as well as the position of the corresponding transition state and therefore modify the folding kinetic rates as well as the kinetic routes. Our analytical models for folding are based on a generalized Fokker–Planck diffusion equation with diffusion coefficient both dependent on coordinate and time. Our simulation for folding are based on structure-based folding models with a specific fast folding protein CspTm studied experimentally on diffusion and folding with single molecules. The coordinate and time-dependent diffusion are especially important to be considered in fast folding and single molecule studies, when there is a small or no free energy barrier and kinetics is controlled by diffusion while underlying statistics of kinetics become important. Including the coordinate dependence of diffusion will challenge the transition state theory of protein folding. The classical transition state theory will have to be modified to be consistent. The more detailed folding mechanistic studies involving phi value analysis based on the classical transition state theory will also have to be quantitatively modified. Complex kinetics with multiple time scales may allow us not only to explore the folding kinetics but also probe the local landscape and barrier height distribution with single-molecule experiments.

1. Introduction

Studying the kinetics of protein folding is very important to understanding the underlying mechanism. The crucial question posted by Levinthal in 1969 is how the many possible configurations of a polypeptide chain rapidly converge to one particular folded state (on the timescale of milliseconds) [1]. The issue has been resolved by the energy landscape theory of folding [2–4]. According to the theory, nature has selected for sequences where the energetic roughness of the landscape is small relative to the depth of the global basin. Though, in the funnel are the bumps and wiggles which can form local traps. For folding to be completed in a biological time scale under physiological temperature (300 K), the slope of the funnel must be steep enough to overcome the local traps. The energy landscape theory has been successful in explaining qualitatively, and quantitatively, many folding experiments [2–5].

This thermodynamic description of folding allows the kinetics to be described as diffusive motion along an order parameter that represents the progress towards the native state. While the folding occurs in a multi-dimensional space where the substates are all locally connected, any given order parameter may or may not have local connectivity. When the kinetic process is fast, either because of a large thermodynamic driving force or because the process is activationless, the native state can either be reached in one shot or through intermediates that are formed rapidly and unravel *en route* to reach the native state. In this case, the states in order-parameter space can move globally from one to another in a discontinuous fashion. On the other hand, if the kinetics are relatively

* Corresponding author at: Department of Chemistry and Department of Physics, State University of New York at Stony Brook, Stony Brook, NY 11794, USA. Fax: +1 631 632 7690.
E-mail address: jin.wang.1@stonybrook.edu (J. Wang).

1046-2023/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.ymeth.2010.04.016

slow due to the nature of the activation folding process, then in general, the states are locally connected in order-parameter space. Accordingly, the dynamic process can be described by a kinetic master equation [6,7] which, in the local connectivity limit, reduces to the diffusion equation [2,8–15].

The kinetics of folding governed by the diffusion equation is determined by both the thermodynamic driving force and local escape capability through diffusion. The diffusion coefficient can depend both on the coordinate and time. The origin of the configuration, or reaction coordinate, dependent diffusion is the fact that the underlying protein folding energy landscape is multidimensional in nature. In real experiments, we can only probe, or trace, finite degrees of freedom. When we project the multidimensional landscape into a few dimensions, or coordinates (for example, Q, fraction of native contacts; root mean square displacement RMSD; radius of gyration Rg, etc.), different coordinates will describe different local conformational landscapes. Therefore the local escape time or diffusion is, in general, coordinate or position dependent [2,16,10,11,13,17,14,18–24]. On the other hand, at each position of the reduced coordinate, there is a local character described by all the coordinates (itself and the rest of the coordinates) around it. This is a result of the roughness of the local energy landscape, which can lead to different time scales. When we project the multidimensional landscape into one, or a few dimensions, the distribution of energy barriers and time scales often emerges, where the diffusion is time dependent which leads to non-exponential kinetics [2,10–15].

Time-dependent diffusion can be understood in terms of the local energetic content. If the underlying landscape is smooth, then there is likely a uniform barrier, which gives single timescale kinetics. If the underlying local landscape is rough, then there is a distribution of barriers, giving rise to multiple timescale kinetics. The local barriers are determined by the distributions of the local energy landscape. Therefore, the local escape time is distributed according to the local energy landscape, or barrier distribution (roughness). The distribution of diffusion coefficients in time is therefore a reflection of the local roughness of the underlying energy landscape.

As mentioned, during the dynamical process, folding can probe different parts of the underlying energy landscape and detect different local barriers. Therefore many possible time scales may coexist, and the kinetics can become non-exponential or multi-exponential. In bulk measurements, it is often difficult to distinguish whether the observed non-exponential kinetics is intrinsic or due to the inhomogeneous distribution of single exponential processes. With recent technological advances, however, single-molecule detection has become possible. Single molecule studies employ probes that are sensitive to the local environments and are therefore ideal tools for understanding the structure of the energy landscape of the proteins [25–28]. Lately, a number of remarkable initial single-molecule folding experiments have been undertaken [29–36] but their interpretation is nontrivial. Statistical fluctuations are intrinsic to single molecules, and since they are not statistically weighted by the number of the molecules as in bulk studies, they can be directly measured. Single-molecule data are essentially sequences of "on-and-off" spikes as a function of time. Therefore determining the nature of complex kinetics with multiple time scales is challenging but they can help us to reveal the local underlying landscape of protein folding.

In the case of coordinate dependent diffusion, the transition state theory will likely be modified. In the kinetics of protein folding, it is conventionally expected that the free energy barrier is the thermodynamic bottleneck in reaching the folded state from the unfolded state [37,2,16,8,9]. From the free energy profile, we can locate the position and height of the barrier, or the transition state, by free energy optimization in the reaction coordinate space. The position of the transition state in the reaction coordinate determines how close the transition state (or the nucleation seed) is to the folded (or unfolded) state. Thus the kinetic rate is determined by the free energy difference between the transition state and the reactant state. Characterizing the transition state ensemble is important in determining the underlying kinetic mechanism and identifying the nucleation seeds from unfolded to folded state [38,2,10–15,21].

For coordinate dependent diffusion, although the expression and functional form of the transition state theory may or may not change significantly, the effective location and the height of the transition state barrier can change. In other words, the presence of the spatial dependence of the diffusion coefficient effectively contributes to the free energy in the kinetic sense so that the height and the position of the effective barrier are changed. Thus although the thermodynamics is not influenced by the coordinate dependent diffusion, the kinetics is controlled by both the thermodynamic free energy and the diffusion. The diffusion acts as an effective driving force in addition to the thermodynamic free energy to contribute to kinetics. It is also possible that the actual kinetic paths may not go through the thermodynamic transition state, but instead pass through the effective transition state determined by both thermodynamics and diffusion.

Including the configurational dependence will therefore challenge the transition state theory of protein folding and it will have to be modified accordingly. The more detailed folding mechanistic studies involving phi value analysis based on the classical transition state theory [39] will have to be quantitatively modified also.

We will examine the diffusion dynamics in protein folding by developing both analytical and simulation models. We will show that diffusion plays an important role in protein folding kinetics. We will first present analytical models for diffusion dynamics of protein folding [10,11,13,14]. Then we will present simulations with structural based model. [16,12,15,21]. We found that diffusion is often coordinate and time dependent. We find that the position dependence of the diffusion coefficient on the reaction coordinate can have a significant contribution to the kinetics in addition to the thermodynamic free energy barrier. It changes the effective free energy barrier and can modify the folding kinetic routes compared with the estimation from transition state theory with constant diffusion. The time-dependent diffusion can lead to non-exponential and multiple time scale kinetics, which reflects the local roughness of the underlying folding landscape. The coordinate and time-dependent diffusion are especially important to consider for fast folding process in single-molecule experiments where there is a small, or no, free energy barrier and the kinetics is controlled by the diffusion both on average and at the statistical distribution level.

Diffusion for specific fast folding protein such as the λ repressor and its fast mutant as well as Villin head piece, WW domain, BBL, CspTm, etc. have been experimentally extensively explored [40–45,32–36]. For fast folding, since the inherent thermodynamic barrier is low, or comparable, to thermal energy k_BT, the effect of diffusion on the kinetic barrier can be significant. On the other hand, the time-dependent diffusion can reveal the local landscape distribution through the multiple time scale kinetics, especially the single-molecule experiments. Thus, the theoretical explorations presented here will contribute to a more complete understanding of the interplay between thermodynamics and diffusion on protein folding kinetics, both at the average and statistical distribution level.

2. Materials and methods

In this work, we used a well studied structure-based model [5], to study the dynamical properties of the protein CspTm through

the ensemble of structures that determine the energy landscape of protein folding. Each residue is represented by a single atom located at the position of the C_α position, with an excluded volume that prevents chain crossing. In this C_α model [5], adjacent residues interact via harmonic bonds. Residues i, $i+1$ and $i+2$ are connected by harmonic bond angles. Native backbone structure is accounted for by a torsional term and non-local interactions are included via a 10–12 potential. The native structure is the global energy minimum and only native interactions are stabilized. Contacts were determined by the Contact of Structural Units software package (CSU) [46]. The functional form of the potential is

$$V = \sum_{\text{bonds}} \varepsilon_r (r - r_o)^2 + \sum_{\text{angles}} \varepsilon_\theta (\theta - \theta_o)^2 + \sum_{\text{backbone}} \varepsilon_\phi \left\{ [1 - cos(\phi - \phi_o)] + \frac{1}{2}[1 - cos(3(\phi - \phi_o))] \right\} + \sum_{\text{contacts}} \varepsilon_C [5(\frac{\sigma_{ij}}{r})^{12} - 6(\frac{\sigma_{ij}}{r})^{10}] + \sum_{\text{non-contacts}} \varepsilon_{NN} V_{NN}(r) \quad (1)$$

where

$$V_{NN}(r) = (\frac{\sigma_{NN}}{r})^{12}, \quad (2)$$

and $\varepsilon_r = 100$, $\varepsilon_\theta = 20$, $\varepsilon_\phi = 1$, $\varepsilon_C = 1$, $\varepsilon_{NN} = 1$ and $\sigma_{NN} = 4.0$ Å. The potential parameters r_o, θ_o, χ_o, $_{\phi o}$ (all dependent on i) and σ_{ij} (the native distance between the alpha carbons) are derived from the structures in the native state obtained by using techniques like Nuclear Magnetic Resonance (NMR). The small-angle X-ray scattering experiments can give reasonable information about the protein radius of gyration Rg.

From the molecular dynamic trajectories we measure the mean first-passage time of folding and the free-energy profiles are obtained via the Weighted Histogram Analysis Method [54]. By repeating the simulations with different initial conditions, we obtained information about the statistical distributions of these folding times and therefore determined the high order-moments as well as the distribution of the first-passage time.

The ratio of the n-th moment and the mean to the n-th power

$$R_n = \frac{\overline{\tau^n}}{\overline{\tau}^n} \quad (3)$$

can be easily computed from the simulations. The farther the ratio $R_n/n!$ is from unity, the greater the deviation from a single exponential Poisson process [15].

In this study we also describe the effects of a varying diffusion coefficient as a function of the order parameter on the folding process. The diffusion coefficient is given as [16]

$$D = \frac{\Delta Q(T)^2}{\tau_{\text{corr}}(T)}. \quad (4)$$

The numerator is the mean square dispersion of the reaction coordinate fluctuations from the dynamic trajectories as a result of the structure-based molecular dynamics simulations, and the denominator is the autocorrelation time of the reaction coordinate Q (number of native contacts) that characterizes the decay of the correlation function defined as:

$$C(Q_0, \Delta) = \frac{\langle Q_0(t) Q_0(t+\Delta) - \langle Q_0^2(t) \rangle \rangle}{\langle Q_0^2(t) \rangle - \langle Q_0(t) \rangle^2}. \quad (5)$$

In order to calculate the local diffusion coefficient at a specific position of the reaction coordinate, we included a harmonic umbrella potential [47,48] to restrain each simulation to a specified range of Q values [18]. A commonly employed reaction coordinate for protein folding is the fraction of native contacts formed in structure Q. Since a contact is either formed, or not formed, Q can only take a discrete set of values and any function of Q, and is therefore discontinuous. When including a biasing potential as a function of a reaction coordinate in molecular dynamics simulations the reaction coordinate must have a defined first derivative. To avoid this discontinuity in Q, we redefine Q to be of the form of a continuous tanh function

$$Q = \sum_{\text{contacts}} \frac{1}{2}\{1 + \tanh[10(r - 1.2\sigma_{ij})]\} \quad (6)$$

with σ_{ij} previously defined. This assumption does not change the nature of the dynamic or the usual contact formation counts due to the fact that this redefinition is essentially a step function at a distance of $1.2\sigma_{ij}$, which is a typical definition of a contact being formed.With this continuous definition of the reaction coordinate, a biasing potential [18] is included as

$$V_{\text{bias}}(Q) = K_Q (Q - Q^*)^2 \quad (7)$$

where Q is the reaction coordinate, K_Q is the strength of the bias included (in units of kT) and Q^* is the probed region of Q where the diffusion coefficient will be studied.

For a given free energy as a function of the reaction coordinate for biased and an unbiased simulations, the biased and unbiased simulations produce similar values of the diffusion coefficient for $K_Q < 50$, calculated by Eq. (4) which means that the calculated value of the fluctuations $\Delta Q(T)^2$, which is smaller than the unbiased case, scale in a similar manner to that of $\tau_{\text{corr}}(T)$, which is also smaller than the unbiased case. While it is not possible to ascertain if this will happen over the entire range of Q, we will make this assumption nevertheless.

From the trajectories, we can also determine the relationship between RMSD(t), Rg(t) and Q(t). Since we have calculated the diffusion coefficient $D(Q)$, we can derive approximate values for the diffusion coefficient along RMSD and Rg (D(RMSD) and D(Rg)) though the transformation $D(\text{Rg}) = D(Q)(d\text{Rg}/dQ)^2$ [24]. Here, we provide a more detailed look at diffusion in Rg, since that is more closely related to experimental measurements on CspTm [32–35].

3. Theory/calculations of diffusion dynamics of protein folding

The kinetics of folding process can be described as processes which obey Metropolis dynamics [2]:

$$R(E_1 \to E_2) = \begin{cases} R_0 exp[\frac{-(E_2 - E_1)}{T}] & for E_2 > E_1 \\ R_0 & for E_2 < E_1. \end{cases} \quad (8)$$

where $R(E_1 \to E_2)$ represents the transition rate for a single polypeptide chain from state 1 to 2 with total energies E_1 and E_2, respectively. R_0 is a overall constant describing the inverse time scale for the transition process between configurations (usually R_0 is on the order of inverse nanoseconds). Therefore the transition rate from one conformational state to a neighboring state is determined by the energy difference of these two states. Further analytic treatment to this problem is made by utilizing the continuous time random walk (CTRW) formalism. By construction, one is able to reduce the multi-dimensional random walk problem to a one-dimensional CTRW, resulting in a generalized master equation. Schematically, one first categorize the energy landscape by the order parameter Q, along which an energy distribution function $P(E,Q)$ [2,10,11]. With the use of Metropolis dynamics one can calculate the associated transition rate distribution function $P(R,Q)$, which specifies the jumping rate R for a molecule at a state with order parameter Q to its neighboring states. From this, one obtains the waiting-time distribution $\Psi(\tau,Q)$ for a molecule to stay at a conformational state for time τ before it leaves. A CTRW can be constructed by knowing both the waiting-time distribution for the system and the jumping probabilities between successive Q's. The latter is approximated to

be time-independent, which is equivalent to the quasi-equilibrium assumption. By this assumption one can calculate these probabilities utilizing the asymptotic distribution:

$$\lim_{t\to\infty} G(Q,\tau) \propto e^{-\beta F(Q)}, \tag{9}$$

where $G(Q,\tau)$ is the probability distribution function for the polypeptide chain at time τ.

In the local connectivity case, the kinetics of folding can be approximated by a generalized Fokker–Planck equation in the Laplace-transformed space [2,10,11,13,14](By generalized, we mean instead of the usual Fokker–Planck equation where diffusion is a constant in time and space representing a typical kinetic Markovian behavior, here we obtain a non-Markovian diffusion kernel in time and space due to the dimensional reduction from multiple configurational degrees of freedom to a single Q):

$$s\tilde{G}(Q,s) - n_i(Q) = \frac{\partial}{\partial Q}\left\{D(Q,s)\left[\tilde{G}(Q,s)\frac{\partial}{\partial Q}U(Q,s) + \frac{\partial}{\partial Q}\tilde{G}(Q,s)\right]\right\}, \tag{10}$$

where

$$U(Q,s) \equiv \frac{F(Q)}{T} + log\frac{D(Q,s)}{D(Q,0)}, \tag{11}$$

and $\tilde{G}(Q,s)$ is the Laplace transform of $G(Q,\tau)$, which is the probability density function and $G(Q,\tau)dQ$ is the probability for a protein to stay between Q and $Q+dQ$ at time τ. $n_i(Q)$ is the initial condition for $G(Q,\tau)$. s, which has an unit of inverse time, is the Laplace transform variable over time τ and $D(Q,s)$ is the frequency and spatial dependent diffusion coefficient. $F(Q)$ is the average free energy. The explicit expression for $D(Q,s)$ is given below.

Here we given an explicit expression for the frequency-dependent diffusion parameter $D(Q,s)$:

$$D(Q,s) \equiv \left\langle \frac{R}{R+s} \right\rangle_R (Q) \Big/ \left\langle \frac{1}{R+s} \right\rangle_R (Q). \tag{12}$$

The average $\langle\rangle_R$ is taken over $P(R,Q)$, the probability distribution function of the transition rate R from one state with order parameter Q to its neighboring states. [2,10,11]. The dependence of the diffusion coefficient D on T and the roughness when $s=0$ and $s>0$ is discussed in [2,28,10,11]. The analytical estimation of diffusion coefficient is based on the assumption of continuous random walk on a biased random energy landscape [2,10,11]. It is evaluated by averaging over the transition rate distributions. The transition rate is defined as the transition or escape rate from a particular state to its neighboring state (not folding transition rate). For practical purpose, we can extract the $D(Q,s)$ directly from the time dependence of the correlation functions of the simulation trajectories. We will explore in details in future studies. At $s=0$, which corresponds to infinite time: $D=D_0\ \exp[-S_0]$ for $T<T_g$; $D=D_0\ \exp[-\beta^2\Delta E^2]$ for $2T_g<T$; $D=D_0\ \exp[-S_0+(\beta_g-\beta)^2\Delta E^2]$ for $T_g<T<2T_g$. T_g is the glass transition temperature which reflects the competition between roughness and entropy of the energy landscape. $T_g=\sqrt{\frac{\Delta E^2}{2S_0}}$ where ΔE^2 is the roughness of the protein energy landscape and S_0 is the configurational entropy of the protein. We have carried out the simulations of CspTm and obtained roughness of the folding landscape $\Delta E = 2.81$ kT and entropy $S_0 = 224.56$ k which leads to trapping temperature $T_g = 0.13$ (low value of T_g is reasonable here because Go model only reflects the topological rather than energetic roughness). $\Delta E = 2.81$ kT is comparable to later estimation of the roughness of energy landscape from coordinate dependent kinetic diffusion in Rg which is around 2 kT. If we obtain the information of diffusion, we obtain the information on the roughness of the local landscape. The zero-frequency ($s=0$) configurational dependent diffusion $D(Q)$ and the constant diffusion coefficient as a function of temperature is explored on the Supplementary Information.

The boundary condition for the generalized Fokker–Planck equation is reflective at $Q=0$, where all the residues are in their non-native states:

$$\left[\tilde{G}(Q,s)\frac{\partial}{\partial Q}U(Q,s) + \frac{\partial}{\partial Q}\tilde{G}(Q,s)\right]\Big|_{Q=0} = 0,$$

and an absorbing one at $Q=Q_f$, where most of the residues are in the native states:

$$\tilde{G}(Q,s) = 0$$

The choice of an absorbing boundary condition at $Q=Q_f$ enables our calculation of the first-passage time distribution. The folding time distribution is less sensitive when the absorbing point is beyond the transition state (on the right of the transition state) and more sensitive when the absorbing point is near or on the left of the transition state.

The first-passage time (FPT) to reach Q_f (that is, the time required for the random walker to visit for the first time) is used as a typical or representative time scale for folding. One has the following relation for the FPT distribution function $P_{FPT}(\tau)$:

$$P_{FPT}(\tau) = \frac{d}{d\tau}(1-\Sigma) = -\frac{d\Sigma}{d\tau} \tag{13}$$

where

$$\sum(\tau) \equiv \int_0^{Q_f} dQ\, G(Q,\tau). \tag{14}$$

The moments of the FPT distribution function are calculated from the following relation:

$$\begin{aligned}\langle\tau^n\rangle &\equiv \int_0^\infty d\tau\,\tau^n P_{FPT}(\tau) \\ &= \left[n(-1)^{n-1}\int_0^\infty d(Q)(\frac{\partial}{\partial s})^{n-1}\tilde{G}(Q,s)\right]\Big|_{s=0}\end{aligned} \tag{15}$$

The $\tilde{G}(Q,\tau)$ and the corresponding moments and distributions of the first-passage times can be solved numerically through matrix inversions [10,11,13,14].

4. Results and discussions

4.1. Time-dependent diffusion: Fluctuations and distributions of first-passage time of folding

We can think of the dynamics of protein folding from an ensemble of unfolded states to the folded state as a chemical reaction in a low concentration of chemical denaturant or appropriate salt or pH conditions favorable for folding. From the energy landscape analysis we address whether there is an ensemble of transition states, or just a single transition state, separating the unfolded and folded ensembles.

We can take a simple example of a chemical reaction process to illustrate the relevant statistics. Let us assume the underline energy landscape for folding is smooth, which results in a two-state-like case (unfolded and folded). The probability of a transition (thermal barrier crossing) from the unfolded states to the folded state is assumed to be $\Delta(\Delta \sim \exp[-\frac{F_T-F_U}{kT}])$, where F_T is the free energy of transition state ensemble and F_U is the free energy of the ensemble of unfolded states) and there are no back reactions for simplicity. Then the survival probability, or kinetic population, of unfolded states is given as $S = \langle(1-\Delta)^N\rangle$ where N is the number of transition (barrier crossing) events between unfolded states to folded state. The average is over different realizations of the folding

events. As we can see already, the kinetic population for unfolding is an exponential function of time only for a statistically independent distribution (Poisson) of (barrier crossing) folding events, where the probability to have N crossing events (folding) during time t is equal to:

$$\rho(N) = (t/\tau)^N exp(-t/\tau) \quad (16)$$

where τ is the average time between two events. In this case, the survival probability is given by:

$$S(t) = \exp(-\Delta t/\tau) \quad (17)$$

In general, if the barrier crossing events are correlated with one another, there will be a non-Poisson distribution of reaction events, which leads to non-exponential kinetics. On the other hand, non-exponential kinetics often implies underlying non-Poisson kinetics for homogeneous systems. This kind of non-exponential and non-Possionian statistics can be quantified by calculating high order statistical correlations and moments or distribution of the whole kinetic populations as pointed out by Wang and Wolynes [25,26,28].

4.2. Temperature dependence of exponential vs non-exponential and Poisson vs non-Poisson statistics of kinetics

Here, we study the cold shock protein from the hyperthermophilic bacterium *Thermotoga maritma* known as CspTm [49] shown in Fig. 1. It is a small protein with 66 aminoacids and molecular mass of 7.5 kDa. The CspTm is a β-barrel protein with its 3*D* structure known as Greek-key with five β strands divided in two antiparallel β sheets. The protein exhibits well-defined two-state behavior [50,51,30,32,33] with rapid folding kinetics. CspTm is an ideal candidate for this study for several reasons. First, CspTm is beta-barrel protein. Barrels are considered topologically complex, and the dynamics of topologically complex proteins are more easily captured by coarse-grained structure-based models than topologically simpler folds [53,55]. Second, CspTm is a relatively small and fast folding protein. This alleviates the need for highly sophisticated sampling protocols and allows more direct calculations of the diffusion coefficients. Finally, there is vast experimental data on CspTm, including denaturant- and temperature-dependent diffusion measurements, hinting at a coordinate- and possibly time-dependent diffusion in CspTm [32–34].

The results of analytical models are described in detail elsewhere [2,16,10,11,13,14]. Our structure-based simulations are consistent with the results of these analytical studies [2,16,10,11,13,14,17] and previous simulation studies [16,52]. From Fig. 2a, we can see that the kinetics in terms of mean first-passage time from our simulations has a U shape dependence on temperature. This is due to the fact at high temperature, the folded state is unstable and the folding kinetics slow down. As temperature is decreased, the kinetic traps are more prominent, and the kinetics is also slowed down significantly. We further calculated the mean first-passage time according to the analytical Eq. 15 at several temperatures and the results are in reasonable agreements with simulations which is also shown in Fig. 2a as diamonds. In Fig. 2b, the fluctuations in kinetics, as measured by the moment ratio, are shown in circle from the simulation results compared with the analytical form from Eq. 15 in diamond. We also see reasonable agreements between analytical results and simulations. At high temperature, the ratio is close to one, which implies a Poisson distribution. At high temperature the kinetics are single exponential (Fig. 3, survival probability with high temperature single exponential decay in long time). As the temperature is decreased, the first-passage time exhibits larger fluctuations and its distribution begins to develop more extended tails.

The U shape average kinetic behavior in temperature was predicted by analytical models and simulations: [2,16,52,10–13,15]. Such kinetic behavior has also been measured experimentally for the lambda repressor protein [42]. Also as in Fig. 2a, position dependent diffusion models reproducing the folding time (or folding rates) upon solvent friction constant with good agreement has been shown by others works as well [18,23].

In the extreme, one can see a power law tail in long times as shown in Fig. 3 (survival probability with relative low temperature power law decay in long time). This is due to the fact that by lowering temperature, one starts to probe the roughness of the underlying energy landscape and the system becomes trapped in local minima. This leads to non-Poissonian statistics and non-exponential kinetics. The power-law behavior seen here can be understood in terms of the density of state of the underlying landscape. If the density of states approaches an exponential distribution $\exp(E/T_c)$ where T_c is a constant, because the transition state for folding is exponentially related to the energy barrier $\exp(-E/T)$, a power-law distribution is observed for both the transition rate and the

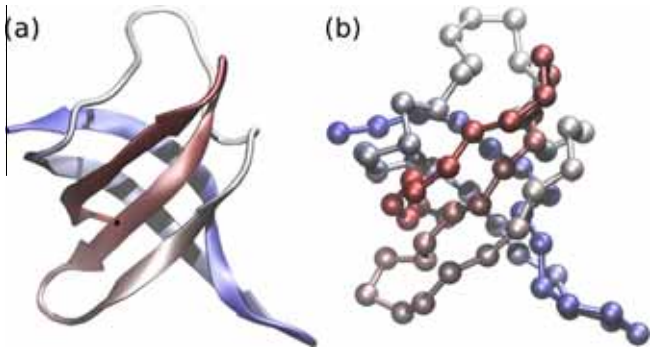


Fig. 1. (a) The structure of CspTm (pdb code 1G6P). (b) Residue level representation (C_α) of CspTm structure.

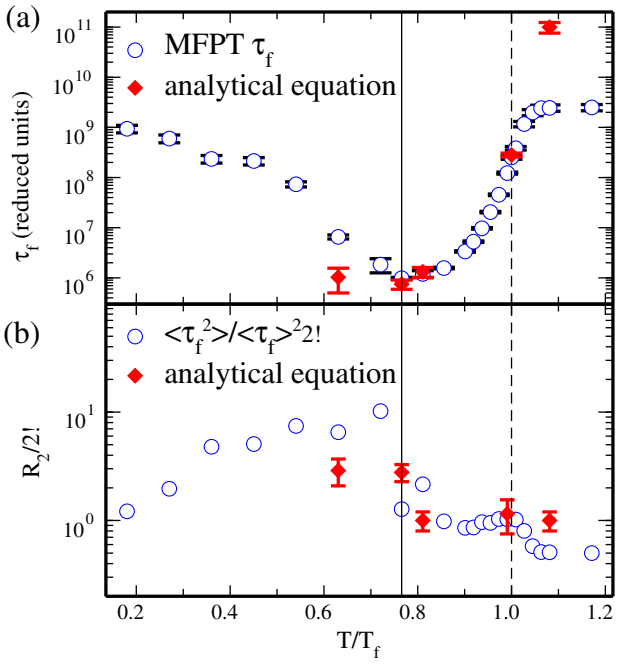


Fig. 2. (a) Mean first-passage time versus temperature. (b) In logarithmic scale, the moment ratio (second order moment divided by average of first-passage time) versus temperature. The open circles are from kinetic simulation runs and the diamonds using the analytical expression from equation 15. The vertical solid line is at the optimal temperature T^0 and the broken line marks the folding temperature T_f.

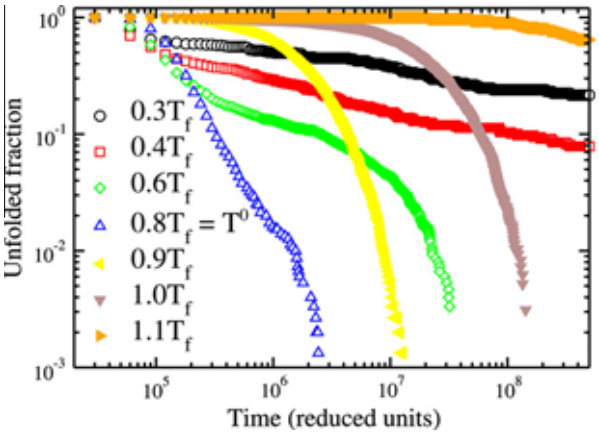


Fig. 3. Survival probability of unfolding (time evolution from unfolding to folding) versus time t in different temperatures *T*.

folding time $f(\tau) \sim 1/\tau^{(T/T_c)+1}$. As the temperature is lowered, $(T/T_c)+1$ decreases and the power law tail becomes more pronounced. By measuring the distribution of kinetics of folding, one can probe the density of states of the underlying folding landscape. When the distribution of first-passage time has long tails, there exists intermittence, where rare events can give a significant contribution to the folding statistics. Due to the ruggedness (local minima) of the underlying energy landscape at low temperatures, relative to the thermal fluctuations, specific discrete paths are energetically distinct and they provide distinguishable contributions to the kinetics. The full distribution of first-passage time, not only its mean, is needed to characterize the dynamics of the system. In the simulation studies [12,15,16], there is a regime of temperature where the population is almost power law in time for several orders of magnitude. This is the computational evidence of low temperature kinetic behavior. The density of states near to low energy end is also near exponential. This seems to support the proposed relationship between kinetics and density of states. Further detailed studies are needed to explore quantitative relationships between the two.

At even lower temperatures, there are again Poissonian, as shown in Fig. 2b, which indicates exponential kinetics and the dominance of a single trap.

The high temperature and low temperature two-state single exponential and intermediate-temperature multi-state non-exponential kinetic behavior were first predicted by analytical models [10,11,13,14]. Similar phenomena have also been reported in previous simulation studies [12,15]. Such exponential-non-exponential kinetic behavior is seen in experiments by temperature dependence measurements of the downhill and activated folding dynamics of several proteins as a result of high temperature activation and low temperature trapping from cold denaturation. [43] We also predicted the Poisson and non-Poisson statistics from the single exponential and exponential kinetics which can be tested by analyzing the experimental data in depth.

One of the advantages of this approach is that it provides a link between theory, simulations, and experiments. In the theoretical approach, the first-passage time and its statistical properties can be easily obtained from the simulations following the procedures outlined in this paper. In experiments, information about first passage time properties can be obtained from the kinetic folding trajectories (for example, from single molecule fluorescence signals). This will stimulate the current and the next round of single-molecule experiments and more detailed simulations to study the full range of the kinetic behavior in temperature. As mentioned, the fluctuations and distribution of folding kinetics provide clues about the density of states of the underlying landscape of protein folding.

4.3. Coordinate-dependent diffusion: shift of position of transition state and barrier height

For CspTm, we calculated the diffusion coefficient as a function of the fraction of native contact *Q* from the correlation functions of *Q* (Fig. 4a) according to the Eq. 4 (Fig. 4c), and from the relationships between *Q* and radius of gyration Rg (Fig. 4b and 4d), we also obtain the diffusion as a function of Rg. We found the diffusion coefficient in Rg monotonically decreases as Rg moves to native values, as shown clearly in Fig. 4d. This is likely due to the confinement and restraints of the configurational space to search through as more compact structures are adopted. We can study the mean-first-passage time for folding τ from any particular coordinate, from energy landscape theory once the diffusion coefficient is given [2,10,11]:

$$\tau(Q) = \int_{Q}^{Q_f} \exp[F(Q')/\mathrm{kT}]/D(Q')dQ' \int_{Q_u}^{Q'} \times \exp[-F(Q'')/\mathrm{kT}]dQ'' \quad (18)$$

where $D(Q)$ is the diffusion coefficient as a function of the reaction coordinate and $F(Q)$ (Fig. 4e and f) is the thermodynamic free energy. Q_u and Q_f represent unfolded and folded state, respectively.

It is important to point out, that for quasi-equilibrium conditions, the friction and diffusion coefficient are related through $D\zeta = \mathrm{kT}$ (via the fluctuation–dissipation theorem) where ζ is the friction coefficient. The resulting steady probability distribution is not dependent on the diffusion coefficient *D*. However, the kinetics is dependent on the diffusion coefficient explicitly. In other words, configurational diffusion will not influence the equilibrium probability distribution, but will influence the flux or kinetic rate [2,10,11].

We are interested in effects of diffusion in Rg on kinetics due to the experimental measurements on Rg [32–35]. From Fig. 4f, the inner integral of τ is dominated by the minimum of $F(\mathrm{Rg}'')$ (in unfolded denatured state in this case). When *D* is a constant, the outer integral is dominated by maximum of $F(\mathrm{Rg}')$. Then we recover the usual transition state expression for τ. The kinetic time is mainly determined by the thermodynamic free energy barrier height at the transition state. When diffusion coefficient is Rg dependent $D = D(Rg)$, then the outer integral is dominated by maximum of $\exp[F(\mathrm{Rg}')/\mathrm{kT}]/D(\mathrm{Rg}')$. Since $D(\mathrm{Rg})$ is monotonically decreasing as a function of Rg towards native state from the results of our simulations, it is obvious that τ with $D = D(\mathrm{Rg})$ is slower than τ with $D = D_0$ (D_0 is the diffusion coefficient at unfolded denatured state in this case). On the other hand, it is not hard to see that the position of the maximum of $\exp[F(Q)/\mathrm{kT}]/D(Q)$ is slightly right shifted (closer towards the native state) relative to the maximum of $\exp[F(Q)/\mathrm{kT}]$ (that is the kinetic transition state is right shifted relative to the thermodynamic transition state) while no significant changes of position of transition state from kinetic diffusion $D(\mathrm{Rg})$ on thermodynamic free energy profile $F(\mathrm{Rg})$.

As shown in Fig. 4f, the thermodynamic transition state is at $\mathrm{Rg} \sim \mathrm{RgM}$. Using the relationship from Eq. 18, $\exp[F(\mathrm{Rg})/\mathrm{kT}]/D(\mathrm{Rg}) = \exp[F(\mathrm{Rg})/\mathrm{kT} - \ln[D(\mathrm{Rg})/D_0]]$, which determines the kinetic time, has an effective barrier height of 5 – 7 kT (depending on the chosen initial value of the unfolding coordinate Rg), which is a higher kinetic barrier compared to the thermodynamic one which is 3 kT from Fig. 4f. Therefore, the contribution to the effective barrier which is purely from diffusion is (5 – 7)kT – 3kT = (2 – 4)kT, which is significant relative to the thermodynamic barrier height. This demonstrates that configuration-dependent diffusion can play a significant role in kinetics especially when the thermodynamic barrier is relatively small (fast folding proteins). The overall (5 – 7)kT barrier is consistent with description in *Q* (Fig. 4e) where thermodynamic barrier (7 kT) dominates and the diffusion only gives moderate contributions to the kinetics (1 kT). Both the description in Rg and *Q* give consistent kinetic results. On the other hand, we can

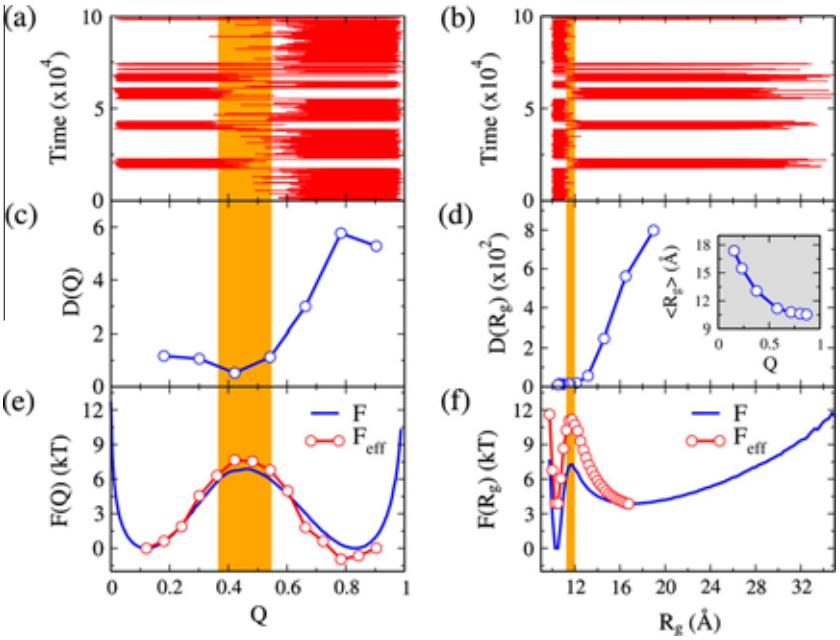


Fig. 4. (a) Trajectories in Q (number of native contacts). (b) Trajectories in Rg (Radius of gyration). (c) Diffusion in Q: $D(Q)$ in units of inverse of time. (d) Diffusion coefficient in Rg and relationship between Rg and Q. D(Rg) is in units of Å^2/time. (e) Thermodynamic free energy profile $F(Q)$ and effective free energy profile after the correction from $D(Q)$ (through $-\ln D(Q)/D_0$). (f) Thermodynamic free energy profile F(Rg) and effective free energy profile after the correction from $D(Q)$ (through $-\ln D(\mathrm{Rg})/D_0$). The orange areas represent the transition state defined as the region between the extremes 1 kT bellow the thermodynamic free energy barrier (free energy maximum). Time is shown in reduced units and the temperature is the folding temperature T_f.

see that the maximum of $\exp[F(Q)/\mathrm{kT}]/D(Q) = \exp[F(Q)/\mathrm{kT} - \ln[D(Q)/D_0]]$ is slightly shifted relative to the thermodynamic transition state from smaller Q to larger Q towards native state. Thus the kinetic route, or path, does not have to follow the equilibrium path as dictated by the underlying thermodynamics and may not go through the thermodynamic transition state. Instead, the kinetic path can go through a short cut, the possible projection to several more coordinates rather than one coordinate. In that situation, the short cut would be apparent. Although configuration-dependent diffusion does not alter the equilibrium distribution, it modifies the kinetic rate, or flux, by increasing the kinetic barrier height, and the kinetic route, through a right shift of the kinetic barrier position in Q.

Our findings are consistent with the results obtained on CspTm from the Schuler Laboratory [32–34]. The overall barrier of 5–7 kT is consistent with experimental measurements at non-zero GdmCl concentrations (around 2 M. A concentration of zero yields a barrier of about 10 kT). They found from single molecule studies, the diffusion coefficient decreases as [GdmCl] is decreased. Decreasing GdmCl favors collapse and therefore decreases Rg as shown experimentally by the Eaton group [35]. These experiments hinted that D(Rg) decreases as Rg decreases (collapse). The changes in the diffusion coefficient from unfolded to collapsed states are similar to our theoretical predictions. As mentioned this coordinate dependence of diffusion coefficient will slow the kinetics due to a shift in the barrier height of collapse, which originates from the roughness of the underlying landscape. The experimental estimate of the roughness for slowing down the kinetics is approximately 1.3(+0.1/−0.2) kT [32,33]. Our estimate from analytical result described earlier on the connection of D versus roughness is about $\Delta\varepsilon = \mathrm{kT}\sqrt{\Delta F} = \mathrm{kT}\sqrt{ln[D_0/D]} = 1.7 \pm 0.3\mathrm{kT}$ which is fairly close to experimental estimations. As seen, the roughness is significant when the folding barrier is low or none.

5. Conclusions

We developed both analytical and simulation methods to explore the diffusive dynamics of protein folding. We found that the diffusion is a quantitative measure of escape from local traps along the protein folding funnel which has two remarkable effects on kinetics. At a fixed coordinate, local escape time depends on the distribution of barriers around it, and the diffusion is therefore time-distributed with non-exponential kinetics and non-Poisson statistics. One of the advantages of this approach is that it provides a link among theory, simulations, and experiments. In the theoretical approach, the first-passage time and its statistical properties can be obtained from the simulations following the procedures outlined in this paper. In the experiments, information about first-passage time properties can be obtained from the kinetic folding trajectories (for example, single molecule fluorescence). This will stimulate the current and the next round of single-molecule experiments as well as more detailed simulations, which will include all-atom models [53], to study the full range of the kinetic behavior in temperature. As mentioned, the fluctuations and distribution of folding kinetics provide direct evidence to the nature of the density of states of the underlying landscape of protein folding.

On the other hand, the energetic character of the local environments (local escape barriers) change along the coordinates, and the diffusion is therefore coordinate dependent. The effects of time-dependent diffusion on folding can lead to non-exponential kinetics and non-Poissonian folding time distributions. The effects of coordinate dependent diffusion on folding can lead to the changes in the kinetic barrier height as well as the position of the corresponding transition state and it therefore can modify the folding rates as well as the kinetic routes. This is especially important for fast folding process where the thermodynamic free energy barrier is either small or zero (downhill process). The kinetics are thus largely determined by the diffusion, which reflects the ability of escaping from the local free energy landscape. Through the experimental and theoretical studies, we can detect and map more details of the local intrinsic features and topography of the underlying energy landscape.

Our findings that the effective free energy barrier shifts both in height and position with the configurational dependent diffusion challenges the transition state theory of protein folding. The

transition state theory works well for constant diffusion, but needs to be modified to describe the kinetics of protein folding when configurational diffusion is taken into account. When folding thermodynamic barriers are high, the diffusion will play a smaller role in determining the kinetics and the classical transition state theory will accurately describe the dynamic process. On the other hand, when the folding thermodynamic barrier is small or comparable to the thermal energy, such as in fast folding proteins, the configurational dependent diffusion can play a significant role in determining the kinetics of folding process.

In addition, some our conclusions remarked here are coherent with our own previous studies [21] and studies from Best and Hummer [24] in terms of position dependence in Q versus Cartesian-like coordinates Rg; shifts in the transition state in Q, definition of new transformed variables Rg in which the position dependence has been eliminated, and effects on the free energy surface and implications on the interpretation of single-molecule experiments by Schuler et al. [30–36]. The diffusion dependence in Q has a different trend then Fig. 4c but has a small contribution to $D(Q)$ as pointed here in the low free energy barrier change in Q (Fig. 4e) when compared with Rg (Fig. 4f). The Cartesian-like coordinates (Rg or RMSD) seem to play a more important role to the protein folding diffusion processes since they capture higher fluctuations in the unfolded ensemble when compared with Q. Moreover, it is possible to correlate new cartesian transformed variables (as in [24]) where the position dependence can be eliminated using Rg obtained from Q transformation deriving the new free energy surface $F(\mathrm{Rg})$. This has importance in experimental measurable quantities like end-to-end distance and Rg as well that can be directly related with theoretical approaches where one would like to obtain experimentally position dependent D using variable transformations. The spatial dependence of diffusion contributes to the effective free energy and gives estimation of the roughness of the underlying landscape of the folding [30–36].

Furthermore, for detailed folding mechanistic studies, the phi value analysis based on the transition state theory may need to be quantitatively modified [21,39]. The phi value is determined by the ratio of free energy change between transition state and unfolded state upon mutations versus free energy change between folded and unfolded state. So it is sensitively dependent on the kinetic barrier at the transition state and its associated changes. If the effective kinetic barrier is changed, then the phi value will also be changed. Since the position of the effective kinetic barrier may also be shifted, then the average phi value which is often correlated with the position can also be shifted.

The theory and methodology outlined in this study provide a basis for comparing and connecting models/simulations [2,10–12,15,13,14,21] with experiments [40–45,32–36] and can be applied to a wide variety of other biological, as well as condensed phase, systems and problems.

Acknowledgments

The authors thank Prof. Peter G. Wolynes, Prof. Jose N. Onuchic, Prof. Martin Gruebele, and Prof. Ben Schuler for helpful discussions. J.C and R.J.O were supported by CAPES, Brazil. R.J.O, V.B.P.L and J.C were partially supported by the Brazilian agency CNPq. V.B.P.L and R.J.O were supported by FAPESP, Brazil. J.W was partially supported by NSF Career Award, and NSFC (China). P.C.W thanks US National Science Foundation I2CAM International Materials Institute Award, Grant DMR-0645461, for funding this international collaboration. P.C.W is funded by a LANL Director's Fellowship. This work was also supported by the Center for Theoretical Biological Physics sponsored by the NSF (Grant PHY-0822283) with additional support from NSF – MCB-0543906.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ymeth.2010.04.016.

References

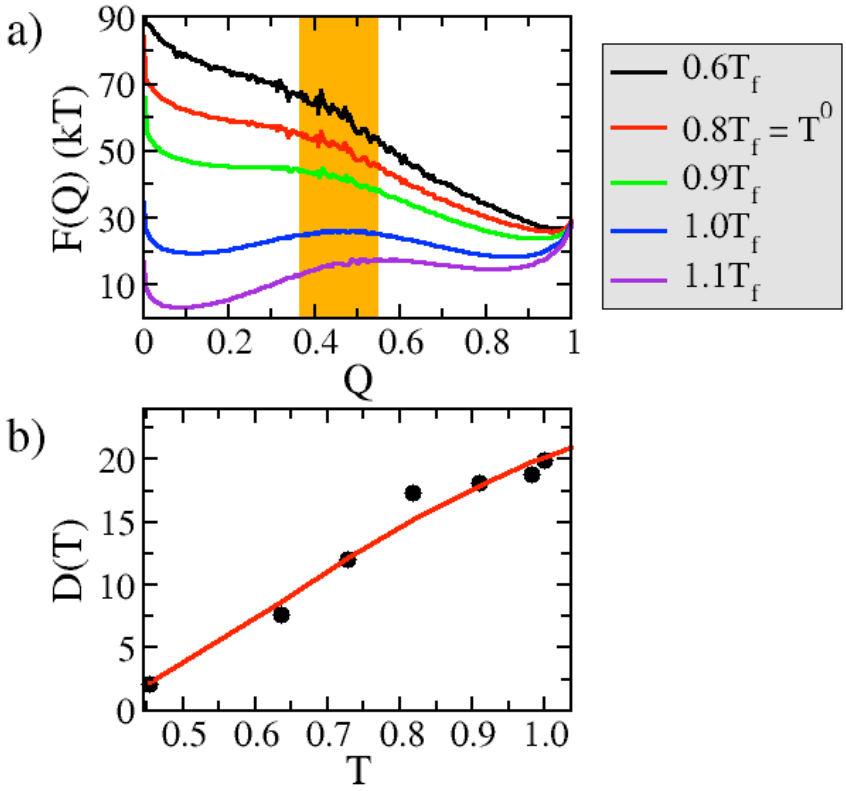
[1] C. Levinthal, in: P. Debrunner, J. Tsibris, E. Munch (Eds.), Proceedings in Mossbauer Spectroscopy in Biological Systems, University of Illinois Press, Urbana, 1969. pp. 22.
[2] J.D. Bryngelson, P.G. Wolynes, J. Phys. Chem. 93 (1989) 6902–6915.
[3] J.D. Bryngelson, J.N. Onuchic, N.D. Socci, P.G. Wolynes, Proteins: Struct. Funct. Genet. 21 (1995) 167–195.
[4] J. Wang, J. Onuchic, P. Wolynes, Phys. Rev. Lett. 76 (1996) 4861–4864.
[5] C. Clementi, H. Nymeyer, J.N. Onuchic, J. Mol. Biol. 298 (2000) 937–953.
[6] J. Wang, J.G. Saven, P.G. Wolynes, J. Chem. Phys. 105 (1996) 11276–11284.
[7] J.G. Saven, J. Wang, P.G. Wolynes, J. Chem. Phys. 101 (1994) 11037–11043.
[8] J. Wang, S.S. Plotkin, P.G. Wolynes, J. de. Physique. 7 (1997) 395.
[9] S.S. Plotkin, P.G. Wolynes, Phys. Rev. Lett. 80 (1998) 5015–5018.
[10] C.L. Lee, C.T. Lin, G. Stell, J. Wang, Phys. Rev. E. 67 (2003) 041905–041910.
[11] C.L. Lee, G. Stell, J. Wang, J. Chem. Phys. 118 (2003) 959–968.
[12] Y. Zhou, C. Zhang, G. Stell, J. Wang, J. Am. Chem. Soc. 125 (2003) 6300–6305.
[13] J. Wang, Biophys. J. 87 (2004) 2164–2171.
[14] J. Wang, C. Lee, G. Stell, Chem. Phys. 316 (2005) 53–60.
[15] V.B.P. Leite, J.N. Onuchic, G. Stell, J. Wang, Biophys. J. 87 (2004) 3633–3641.
[16] N.D. Socci, J.N. Onuchic, P.G. Wolynes, J. Chem. Phys. 104 (1996) 5860–5868.
[17] T.V. Pogorelov, Z. Luthey-Schulten, Biophys. J. 87 (2004) 207–214.
[18] G. Hummer, New J. Phys 7 (2005) 34–48.
[19] R.B. Best, G. Hummer, Phys. Rev. Lett. 96 (2006) 228104–228108.
[20] S. Yang, J.N. Onuchic, H. Levine, J. Chem. Phys. 125 (2006) 54910–054918.
[21] J. Chahine, R.J. Oliveira, V.B.P. Leite, J. Wang, Proc. Natl. Acad. Sci. USA 104 (2007) 14646–14651.
[22] S. Yang, J.N. Onuchic, A.E. Garcia, H. Levine, J. Mol. Biol. 372 (2007) 756–763.
[23] A.K. Sangha, T. Keyes, J. Phys. Chem. B 113 (2009) 15886–15894.
[24] R.B. Best, G. Hummer, Proc. Natl. Acad. Sci. USA 107 (2010) 1088–1093.
[25] J. Wang, P.G. Wolynes, Phys. Rev. Lett. 74 (1995) 4317–4320.
[26] J. Wang, P.G. Wolynes, J. Chem. Phys. 110 (1999) 4812–4819.
[27] J.N. Onuchic, J. Wang, P.G. Wolynes, Chem. Phys. 247 (1999) 175–184.
[28] J. Wang, J. Chem. Phys. 118 (2003) 952–958.
[29] A.A. Deniz, T.A. Laurence, G.S. Beligere, M. Dahan, A.B. Martin, D.S. Chemla, P.E. Dawson, P.G. Schultz, S. Weiss, Proc. Natl. Acad. Sci. USA 97 (2000) 5179–5184.
[30] B. Schuler, E.A. Lipman, W.A. Eaton, Nature 429 (2002) 743–747.
[31] E.A. Lipman, B. Schuler, O. Bakajin, W.A. Eaton, Science 301 (2003) 1233–1235.
[32] D. Nettels, I.V. Gopich, A. Hoffmann, B. Schuler, Proc. Natl. Acad. Sci. USA 104 (2007) 2655–2660.
[33] A. Hoffmann, A. Kane, D. Nettels, D.E. Hertzog, P. Baumgartel, J. Lengefeld, G. Reichardt, D.A. Horsley, R. Seckler, O. Bakajin, B. Schuler, Proc. Natl. Acad. Sci. USA 104 (2007) 105–110.
[34] D. Nettels, S. Muller-Spath, F. Kuster, H. Hofmann, D. Haenni, S. Ruegger, L. Reymond, A. Hoffmann, J. Kubelka, B. Heinz, K. Gast, R.B. Best, B. Schuler, Proc. Natl. Acad. Sci. USA 106 (2009) 20740–20745.
[35] K.A. Merchant, R.B. Best, J.M. Louis, I.V. Gopich, W.A. Eaton, Proc. Natl. Acad. Sci. USA 104 (2007) 1528–1533.
[36] T. Cellmer, E.R. Henry, J. Hofrichter, W.A. Eaton, Proc. Natl. Acad. Sci. USA 105 (2008) 18320–18325.
[37] H. Eyring, J. Chem. Phys. 3 (1935) 107–115.
[38] H.A. Kramers, Physica 7 (1940) 284–304.
[39] L.S. Itzhaki, D.E. Otzen, A.R. Fersht, J. Mol. Biol. 254 (1995) 260–288.
[40] J. Sabelko, J. Ervin, M. Gruebele, Proc. Natl. Acad. Sci. USA 96 (1999) 6031–6036.
[41] H. Nguyen, M. Jager, A. Moretto, M. Gruebele, J.W. Kelly, Proc. Natl. Acad. Sci. USA 100 (2003) 3948–3953.
[42] W.Y. Yang, M. Gruebele, Biochemistry 43 (2004) 13018–13025.
[43] F. Liu, D. Du, A.A. Fuller, J.E. Davoren, P. Wipf, J.W. Kelly, M. Gruebele, Proc. Natl. Acad. Sci. USA 105 (2008) 2369–2374.
[44] J. Kubelka, T.K. Chiu, D.R. Davies, W.A. Eaton, J. Hofrichter, J. Mol. Biol. 359 (2006) 546–553.
[45] M.M. Garcia-Mira, M. Sadqi, N. Fischer, J.M. Sanchez-Ruiz, V. Munoz, Science 298 (2002) 2191–2195.
[46] V. Sobolev, R. Wade, G. Vried, M. Edelman, Proteins: Struct. Funct. Genet. 25 (1996) 120–129.
[47] G.M. Torrie, J.P. Valleau, J. Comp. Phys. 23 (1977) 187–199.
[48] C. Bartels, M. Karplus, J. Comp. Phys. 18 (1997) 1450–1462.
[49] W. Kremer, B. Schuler, S. Harrieder, M. Geyer, W. Gronwald, C. Welker, R. Jaenicke, H.R. Kalbitzer, Eur. J. Biochem. 268 (2001) 2527–2539.
[50] D. Perl, C. Welker, T. Schindler, K. Schröder, M.A. Marahiel, R. Jaenicke, F.X. Schmid, New J. Phys. 5 (1998) 229–235.
[51] D. Wassenberg, C. Welkera, R. Jaenick, J. Mol. Biol. 289 (1999) 187–193.
[52] H. Kaya, H.S. Chan, J. Mol. Biol. 315 (2002) 899–909.
[53] P.C. Whitford, J.K. Noel, S. Gosavi, A. Schug, K.Y. Sanbonmatsu, J.N. Onuchic, Proteins: Struct. Funct. Bioinf. 75 (2009) 430–441.
[54] A.M. Ferrenberg, R.H. Swendsen, Phys. Rev. Lett. 61 (1988) 2635–2638.
[55] J.E. Shea, C.L. Brooks III, Annu. Rev. Phys. Chem. 52 (2001) 499–535.

Supplementary Information

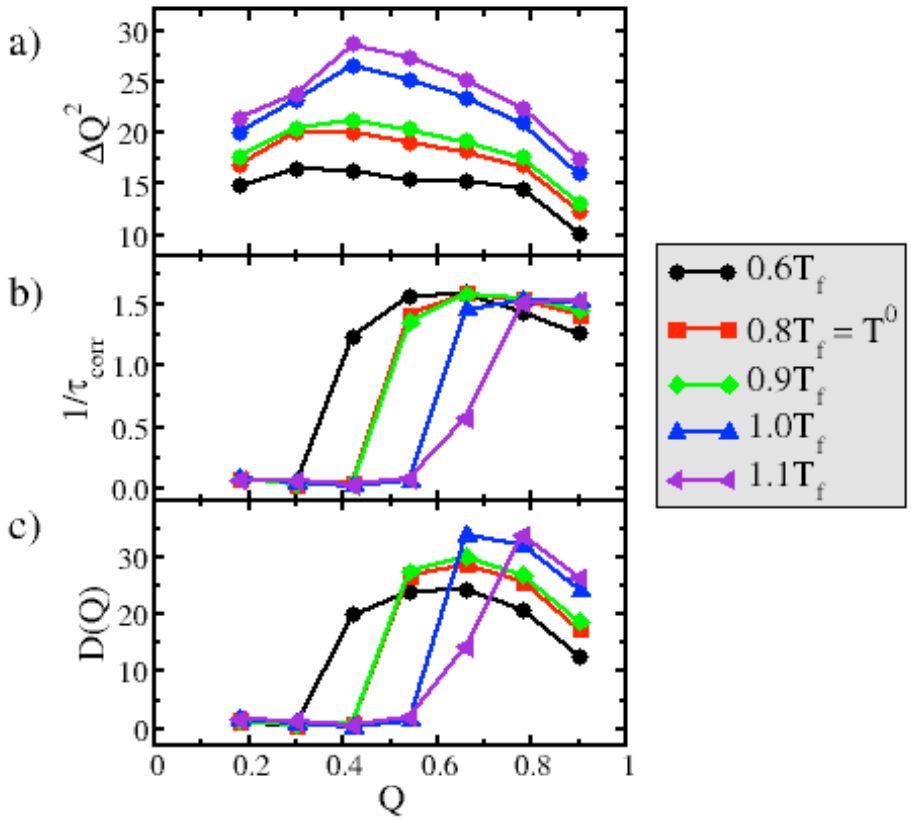
As seen in Figure S1a, the free energy profiles in Q with different temperatures show that at high temperatures, the $F(Q)$ shows the coexistence of unfolded and folded two stable states with high barrier height in between them. At low temperatures, the $F(Q)$ shows almost downhill folding with almost no free energy barrier. As shown in Figure S1b, the overall diffusion coefficient becomes larger as temperature increases.

As seen in Figure S2a, the fluctuations in Q is larger when temperature increases. The more and more downhill trend of underlying free energy profile in Q may lead to faster correlation times as temperature drops as shown in Figure S2b and faster diffusion coefficients before the transition state as shown in Figure S2c. After transition state, free energy profile is downhill for almost all temperature ranges leading to faster diffusion at higher temperatures as shown in Figure S2c (for similar free energy profile, higher temperature with higher thermal energy will lead to faster kinetics and therefore diffusion).

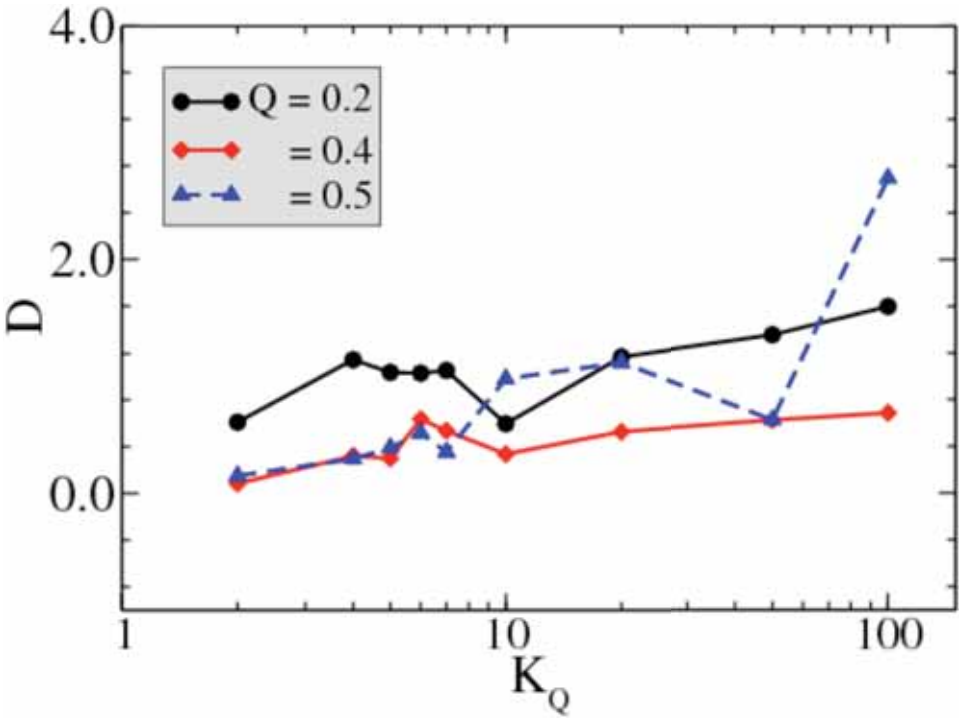
Figures



S1. a) Free energy profiles as a function of the reaction coordinate fraction of native contacts *F(Q)* for a set of different temperatures ranging from 0.6 to 1.1T_f. T_f is the protein folding temperature and T^0 is the optimal folding temperature. The free energy is expressed in multiples of *kT*. The orange area represents the transition state defined as the region between the extremes in *Q* at 1*kT* bellow the thermodynamic free energy barrier (free energy maximum) at T_f. b) Diffusion coefficients as a function of the normalized temperature *D(T)*, i.e., the absolute simulation temperature divided by T_f. The black circles are the diffusion coefficients estimated from the variance of *Q* over the autocorrelation time decay ($D(T) = \Delta Q^2(T)/\tau_{corr}(T)$, equation 4 in the main text) obtained in a long run for a fixed temperature. The solid line is a fit to the estimated data using the equation $D = D_0 \exp(-\beta^2 \Delta E^2)$ for the simulation temperature range $2T_g < T$ with β been $1/kT$. The parameters fit result in $D_0 = 34.8$ and $\Delta E^2 = 0.68$. *D* has unit of inverse of the simulation reduced time unit and T_g is characteristic glass transition temperature.



S2. Fraction of protein native contacts Q as reaction coordinate. a) Fluctuation of the reaction coordinate Q, ΔQ^2 b) inverse of the autocorrelation time decay, $1/\tau_{corr}$ c) position dependent diffusion coefficients for a range of temperatures from 0.6 to 1.1T_f. T_f is the transition folding temperature and T^0 is the optimal folding temperature. Each point in ΔQ^2 and $1/\tau_{corr}$ is extracted from a long simulation with the harmonic constraint in Q (position constrained) at the specific temperature and D is calculated simply by multiplying a) and b) (equation 4 in the main text). τ_{corr} is the characteristic decay of the correlation function define by equation 5 in the main text.



S3. The diffusion coefficient D as a function of K_Q for three values of Q . Each point corresponds to a particular simulation performed with the biased potential at a particular value of Q. All simulations are at the folding temperature T_f.

Apêndice C

Manuscrito *em preparação*

O manuscrito (em preparação), "*Topography of Funneled Landscape Determines the Thermodynamics and Kinetics of Protein Folding*", descreve além da teoria desenvolvida, os resultados acumulados durante o período de estágio no exterior com o grupo do Prof. Dr. Jin Wang[1]. O artigo também apresenta parte dos resultados obtidos durante o trabalho de doutorado no Brasil.

[1]Universidade e endereço para contato se encontram no artigo que segue nesse apêndice C.

Topography of Funneled Landscape Determines the Thermodynamics and Kinetics of Protein Folding

Ronaldo J. Oliveira [a], Paul Whitford [bc], Jorge Chahine [a], Vitor B.P. Leite [a], Jose N. Onuchic [d*], Jin Wang [ef*]

[a] Departamento de Física - Instituto de Biociências, Letras e Ciências Exatas
Universidade Estadual Paulista
São José do Rio Preto, SP 15054-000, Brazil

[b] Theoretical Biology and Biophysics Group, Theoretical Division
Los Alamos National Laboratory, Mail Stop K710, T-6
Los Alamos, NM 87545, USA

[c] International Institute for Complex Adaptive Matter
University of California at Davis
Davis, CA 95616, USA

[d] Center for Theoretical Biological Physics
University of California at San Diego
La Jolla, CA 92093, USA

[e] Department of Chemistry, Physics & Applied Mathematics
State University of New York at Stony Brook
Stony Brook, NY 11794, USA

[f] State Key Laboratory of Electroanalytical Chemistry
Changchun Institute of Applied Chemistry
Chinese Academy of Sciences
Changchun, Jilin 130021
People's Republic of China

* E-mail: jonuchic@ucsd.edu, jin.wang.1@stonybrook.edu

September 27, 2010

Abstract

We quantify the energy landscape of protein folding. This is realized through the exploration of the underlying density of states. By exploring the folding dynamics with lattice model, off lattice structure based model, and all atom force field models at different temperatures, we can obtain the thermodynamic and kinetic information. By converting the distribution of the energies from the results of the simulations in the canonical ensembles to the distribution of the energies in micro-canonical ensemble, we can obtain directly the underlying density of states of protein folding. Based on the density of states, we can identify three quantities essential for characterizing the folding landscape topography: the energy gap between the native state and the average non-native states $\delta E = |E_n - \bar{E}_{non-native}|$, the roughness or variance in the energies of the non-native states measured by the standard deviation of the energies in non-native states $\Delta E = \sqrt{< E^2 > - < E >^2}$, and the size of the funnel as measured by the entropy of the system $S = k_B ln\Omega$ (where Ω is the number of the states). We show that the dimensionless ratio between the gap, roughness and entropy of the system $\Lambda = \frac{\delta E}{\Delta E\sqrt{2S}}$ can quantify the degree of the funnel and determine the thermodynamics as well as kinetics of folding. We discovered that in the samples of proteins we investigated, all of them have funneled folding energy landscapes. The ratio Λ characterizing the topology of the underlying folding landscape provides a new way to classify different proteins. Further more the topology ratio Λ is also shown to be correlated strongly (monotonically) with the thermodynamic stability against traps characterized by the ratio of thermodynamic stability temperature versus trapping temperature. The landscape topology parameter Λ also monotonically correlates with the folding kinetic rates scaled by the sizes of the proteins. This study bridges the gap between the quantification of underlying folding energy landscape topography and the thermodynamics and kinetics of protein folding.

Introduction

The biomolecular functions are realized by the interactions among biomolecules. In order to study the interactions, the information of the underlying structures of the biomolecules is needed. The understanding of how the primary sequence information transforms to the three dimensional structures is the protein folding problem. Solving the protein folding problem is crucial for understanding the bio-molecular functions and structure based drug design.

In 1969, Levinthal presented a now called Levinthal paradox [1]. The possible configurational state space for a protein is huge. If the folding explores all the possible states, then it takes cosmological time to complete the folding process. In experiments and in nature, the folding typically happens in millisecond to seconds. The recently developed protein folding theory resolves the Levinthal paradox by assuming that the underlying energy landscape is funneled towards the native state [2–7]. There can be local traps on the way to folding native state. To guarantee the folding to be completed in a biological time scale in physiological conditions, the steepness of the protein folding funnel should be large compared with the roughness from the local traps. The new theory of protein folding successfully explained many folding experiments both qualitatively and quantitatively [2–8].

The theoretical and experimental community are converging to the new folding theory [3–7, 9–27]. There are still some concerns from the experimental community of how to quantify the landscape of the folding funnel and how that is related to the experimental measures such as melting temperature for thermodynamic stability, the rates of folding for kinetics etc [28, 29]. We will establish the quantitative connection between the underlying energy landscape topography and the thermodynamics and kinetics of protein folding, and address these concerns in the current study reported here.

It is obvious that the folding free energy landscape correlates with the stability and kinetics of the folding. But it varies and depends sensitively on the environmental conditions for a given sequence. The question is what determines the intrinsic properties of folding from just the sequence information which should depend less sensitively on the environments. Such information is embedded in the energies rather than free energies of the system. That is the sequence identity determines what the intrinsic interactions are while free energy is a synthesized entity (from energy, entropy and environments). So we should aim for energy instead of the free energy. It would be ideal to obtain such information. For folding, the energy we are interested in is in fact the effective energy (not just the pure interactions among atoms of the protein)

since the underlying energy landscape is typically solvent (water molecules) averaged [30]. So the energy funnel we see is in fact the funnel of the effective energy. The effective information is relatively easy to obtain compared to the real one since the real energy landscape requires the explicit solvent simulations which are still hard to carry for the folding. Furthermore, the folding mechanism is more likely to be determined by the effective energy from interactions among protein atoms averaged over water molecules rather than the pure interactions among protein atoms themselves. Because it is the effective interactions of hydrophobic and hydrophilic interactions which gives the driving force for protein folding.

We will quantify the effective energy landscape of protein folding. We will do so through the explorations of the underlying density of states. The density of states gives the intrinsic probabilistic distribution of energies, which does not depend on temperature explicitly, can reflect the true underlying energy landscapes. The information on density of state in statistical mechanics is usually obtained from the micro-canonical ensemble. The models and simulations of protein folding are usually performed at constant temperatures in canonical ensemble. We can transform from the results from the simulations of the canonical ensemble to the micro-canonical ensemble to obtain the density of states.

We will use lattice model, off lattice structure based model, and all atom force field models (with implicit solvent) to explore the folding dynamics of a sample of proteins at different temperatures. We will convert the distribution of the energies from the results of the simulations in the canonical ensembles to the distribution of the energies in micro-canonical ensemble. In this way, we can obtain directly the underlying density of states and therefore the intrinsic effective energy landscape of protein folding.

Based on the density of states, we can identify three quantities essential for characterizing the effective folding landscape topography: the energy gap between the native state and the average non-native states $\delta E = |E_n - \bar{E}_{non-native}|$, the roughness or variance in the energies of the non-native states measured by the standard deviation of the energies in non-native states $\Delta E = \sqrt{< E^2 > - < E >^2}$, the size of the funnel as quantitatively measured by the entropy of the system $S = k_B ln\Omega$ (where Ω is the number of the states). We show that the dimensionless ratio between the gap, roughness and entropy of the system $\Lambda = \frac{\delta E}{\Delta E\sqrt{2S}}$ can quantify the degree of the funnel and determine the thermodynamics as well as kinetics of folding [31–34].

We found the following results: (1) We discovered that in the samples of proteins we investigated, all of them have funneled folding energy landscapes. (2) The ratio Λ characterizing the topology of the underlying folding landscape provides a new way to classify different proteins.

(3) Further more the topology parameter Λ is shown to be correlated strongly (monotonically) with the thermodynamic stability against traps characterized by the ratio of thermodynamic stability temperature versus trapping temperature. (4) The landscape topology parameter Λ also monotonically correlates with the folding kinetic rates scaled by the sizes of the proteins. This study bridges the gap between the quantification of underlying folding energy landscape topography and the thermodynamics and kinetics of protein folding.

Results and Discussions

Density of States and Energy Landscape of Protein Folding

We will explore the effective energy landscape of quantify the topology to establish the link between them and thermodynamic and kinetic of protein folding. As mentioned in the method section, we first perform the Monte Carlo simulations on the lattice model, and molecular dynamics simulations on the off lattice structure based model and all atom force field model on a sample of proteins as shown in Figure 1 in different temperatures under canonical ensembles. We obtained the distribution of the energies (temperature dependent). Now we transform the distribution of the energies in canonical ensemble to the distribution of the energies in micro-canonical ensemble, the density of states. The density of states does not explicitly depend on temperature and therefore is intrinsic and reflect the underlying interactions.

To visualize the energy landscape of folding, we can project the energies to specific order parameters or reaction coordinates. The lowest order projection is the zero dimensional projection of the energies to themselves: distribution of the energies, which we can represent as distributions of spectrum lines. In Figure 2, we show the distribution or energy spectrum of the protein folding from the simulations with structure based model at the residue level. Each energy level of the distribution represents a bin of a sum of $exp[100]$ states except for the native energy level. The inset is a magnification of the energy levels distribution of albumin which shows how close the levels are at high energy end. The lowest (native) energy E_n is relocated to be at $E_n = 0$ for a better visualization purpose. The stability gap δE between the native energy and the average energy of non-native states is indicated in a vertical arrow for each protein. The standard deviation of the energy (roughness of the energy landscape) of non-native states ΔE is also indicated as a vertical arrow. Energy is in unit of kT. The energy gap represents on average how far away of non-native states relative to the native state, giving a quantitative measure of the bias or tendency towards the native state relative to other states.

On the other hand, the standard deviation of the energies measures the deviation from the homogeneous distribution giving the variance in energies. This gives a quantitative measures of the inhomogeneities in energy or roughness of the landscape.

We can see clearly there is a distinct gap between the native state and the rest of the non-native states as compared with the roughness of the energy landscape for all the proteins we have simulated. Therefore, all the proteins have significant gaps or bias towards the native states against roughness. However, the quantitative degrees of the biasing of native state relative to others are not uniform. As seen clearly in Figure 2, different proteins have different gaps and roughness. These differences will provide a way to classify proteins from the physical (interaction) perspectives. Furthermore, the differences in gap and roughness have direct consequences in thermodynamics and kinetics as we will show later.

Landscape Projection Along Order Parameters and Size, Steepness and Roughness of Protein Folding Funnel

In Figure 3a, we show a the energy distribution or density of states versus energy. Because of the large number of states, we explore the entropy of the configurational states (quantified as the logarithm of the density or number of states n(E)): $S = lnn(E)$ as function of energy E where n(E) is the number of states at given E. In Figure 3b, we do a one dimensional projection of the density of states or distribution of energies to a reaction coordinate or order parameter, Q which we define as fraction of native contacts (the contact is defined as the spatial contact between the two amino acid residues in the protein within certain cutoff distance). The entropy S(Q) versus Q is shown. The lowest (native) energy En is relocated to be at $E_n = 0$ for a better comparison purpose. Energy is in unit of kT. The energy E versus Q is plotted in Figure 3c.

From Figure 3a, we can see clearly the entropy or number (density) of states decreases monotonically with respect to energy which is physically intuitive. It reflects the fact that on average, the number of states goes down as we have lower energies. Or in other words, the states are less and less or sparely distributed as the energy becomes lower and lower. We notice that this is different from the canonical ensemble expectations $n(E,T)$ where T is the temperature and the low energies lead to higher populations of the states by Boltzmann law $n(E,T)$ $n(E)exp[-E/k_BT]$. Here we emphasize the distribution of energy states or density of state $n(E)$ we are interested in is intrinsic to the system. It can reflect the intrinsic underlying landscape of the protein folding. As we can see the quantitative degree of decreasing or the slope of the entropy towards the native states in energies are different for different proteins.

This also provides a physical way of classifying proteins from intrinsic energetic perspective.

From Figure 3b, we can see that the entropy or number (density) of states decreases monotonically with respect to fraction of native contacts Q. So as the systems gets closer and closer to the native state, the number of state is smaller and smaller. As we can see the quantitative degree of decreasing or the slope of the entropy towards the native states in Q are different for different proteins. This again provides a way of classifying proteins from structure perspective.

In Figure 3c, we can see the energy versus the fraction of native contact Q. We see that the energies monotonically goes down as the system is closer to the native state. As we can see the quantitative degree of decreasing or the slope of the energy towards the native states in Q are different for different proteins. This illustrates that the landscape of protein folding is biased or funneled towards the native state quantified by the slope of energy versus Q towards the native state.

In an analytical landscape of protein folding, we have proved that the energy gap is closely linked to the slope of the energy versus Q towards the native state. We also show this in Figure 3d and we see a strong correlation between the average slope of energy landscape (average slope of E vs Q) and energy gap. Therefore we clearly see all of our samples have funneled landscape biasing towards the native states, where the degrees of the biasing or slopes of the funnel are different from one protein to another.

In Figure 3e, we have shown that the roughness of the energy landscape ΔE measured by the standard deviation of energies versus fraction of native contacts. We see that the roughness of the landscape $\Delta E(Q)$ decreases as the system is closer to the native state. That is the landscape becomes less rough as it is closer to the native state.

In Figure 3f, we have shown the two dimensional projection of density of states (in logarithmic scale) in energy E and fraction of native contacts Q. So we can see a funnel or bias towards native state in both energy (less states with lower the energy towards native state) and structure (less states towards native state). The two dimensional projection gives a more complete picture of the protein folding funnel landscape. As the states moves towards the native structure, the states become less and less and entropy becomes lower and lower. This means that the size of the funnel is quantitatively measured by the entropy S(E,Q) available and decreases as E is lowered and Q is closer to native state value. So the protein folding funnel size is determined by two factors: one is the energy E which characterizes the interactions among residues in a physical way intrinsic to the system and the other is the fraction of native

contacts Q which characterizes the folding in a structural way which is also intrinsic to the system. As shown in Figure 3c, the energy decreases towards native state. So energy is biased towards the native state. So the size of the landscape shrinks towards the native state and the slope of the landscape and furthermore the roughness of the landscape also decreases towards the native state. This gives the complete picture of protein folding funnel.

Free Energy Profile of Protein Folding

In Figure 4, we see free energy profile or landscape as a function of the reaction coordinate fraction of similarity degree with the native protein structure F(Q) for the structure based C_α model a) only the one-state downhill folders and b) two-state folders. Each curve is a free energy at folding temperature $T = T_f$ and normalized by the respective kT_f. As we can see the free energy landscape for protein folding is not necessarily funneled. Some proteins are strong folder with one basin of attraction in free energy landscape. The slope of the free energy landscape for the one state folder measures how easy to fold connected with kinetics rates. For other proteins, two basins of attractions emerge leading to the two state behavior of the proteins often seen in protein folding experiments. The barrier height of the two state folder measures how difficult to fold connected directly to the kinetic experiments.

The shape of the free energy landscape depends on the temperatures explicitly. At folding temperatures T_f, for two state folder, the folding free energy is equal to the unfolding free energy. At lower temperatures than folding temperature T_f, the free energy is biased towards the native state as illustrated by Figure 4c. At higher temperatures than folding temperature T_f, the free energy is biased towards the non-native state as illustrated by Figure 4d. We use protein CI2 as an example to show the general phenomena. The relative free energy of native state and non-native states provides a measure of thermodynamics or equilibrium constants linking directly to the experiments.

The Temperature Dependence of Kinetic Rates of Protein Folding

In Figure 5, we see the kinetics of folding of different proteins with respect to temperatures. For the proteins in study, we see U-shape dependence of the mean first passage time (MFPT) τ_f as a function of a) absolute simulation temperature T b) normalized temperature $T = T_0$ and c) normalized temperature $T = T_f$. T_0 is the optimal temperature for folding, i.e., the temperature where the protein folding rate τ_f is minimal and T_f is the folding temperature.

The U shape dependence of fold kinetic rate versus temperature can be explained. At higher temperatures, from our folding thermodynamic studies as shown in Figure 4d, the non-native states are more preferred and the folded state becomes less stable. Therefore it is more difficult to proceed towards folding. On the other hand, when the temperatures are lower, although the native state are preferred, the traps of folding become more prominent. This slows down the kinetics. Therefore, there is an optimal temperature for each protein where the kinetic rate is minimal. Such kinetic behavior was investigated and seen experimentally [3–7, 9–27, 35, 36]. A key question is how to compare the kinetics of different proteins and at what temperature. We believe it is ideal to normalize the temperature according to the one where the minimum folding kinetics occurs for each protein and compare them. This would be an ideal comparison of the intrinsic kinetics of folding of different proteins.

Different proteins have different thermodynamic stabilities and therefore different folding temperatures T_f. The kinetic rates measured in experiments are often at different temperatures near their respective folding temperatures. We notice that the order of folding kinetics (fast to slow) is different if we compare the folding kinetic rates of different proteins at their respective folding temperatures and if we compare the folding kinetic rates of different proteins at their respective temperature of minimum kinetic rate of folding. The question is which comparison is the best one connected to the underlying intrinsic properties of folding landscape. We believe the kinetic comparison with respect to temperature of minimum folding rates links to the intrinsic properties of the folding landscape. This will be confirmed and illustrated in the later part of the paper.

The Influences of the Roughness of the Underlying Energy Landscape on Kinetic Rates of Protein Folding

In Figure 6, we illustrate how the roughness of protein folding landscape influences the kinetics and how is that related to the quantitative measure of the protein folding funnel Λ: gap to roughness ratio scaled with entropy. We show the folding of protein CspTm [37] results with different strengths for the energetic frustration term ϵ_{NC} in the interaction energies. We see U-shape dependence of the mean first passage time (MFPT) τ_f as a function of temperature T b) normalized temperature $T = T_0$. T_0 is the optimal temperature for folding, i.e., the temperature where the protein τ_f is minimal. We notice that that around folding temperature $T = T_f = 1$, τ_f is optimum at roughness parameter $\epsilon_{NC} = 0.2$. We see in Figure 6c that Λ is maximum at $\epsilon_{NC} = 0.2$. When we plot τ_f versus Λ, we see a monotonic relationship

between the two. It implies that when the underlying landscape is more funneled, measured by the dominant energy gap against roughness scaled with entropy, the kinetics of the underlying folding process is faster. In the table we see that the reason why the kinetic rate is faster at an optimal $\epsilon_{NC} = 0.2$ is due to the non-monotonic relationship between ϵ_{NC} and energy landscape roughness parameter ΔE. We see that the energy gap increases slightly with the increase of ϵ_{NC} due to the non-native interactions leads to the separation between native and non-native states. The roughness ΔE however first decreases with ϵ_{NC} and then increases with ϵ_{NC}. ΔE reaches minimum at $\epsilon_{NC} = 0.2$. The physical meaning of this is that a little bit shaking may help the protein folding while large shaking will certainly make proteins unstable.

The Underlying Energy Landscape Topology Determines the Thermodynamics and Kinetics of Protein Folding

In Figure 7, we illustrate how the protein folding landscape topological parameter Λ as the energy gap versus roughness scaled with entropy correlates with the folding thermodynamics and kinetics. In Figure 7a, we found that there is a good positive correlation between thermodynamic stability characterized by folding transition temperature relative to trapping temperature and the Λ. This implies that the more steep of the folding funnel or the less rugged the protein folding funnel or the less entropy (size) of the folding funnel, the more stable thermodynamically the protein folding against traps. Folding temperature can be measured from the experiments. So we can see the thermodynamical stability of proteins measured in experiments reflects the underlying topology of the protein folding energy landscape quantitatively (comparing different proteins). We can see the line is analytical prediction of the correlation between T_f/T_g and Λ. Although the qualitative trend is good, the quantitatively is not enough accurate. This is probably due to the analytical derivation used mean field approximations without explicitly taking into account of the surface effects for finite size of proteins (or capilarity of protein folding [38]).

We have calculated the kinetic rates of folding. Our simulation results of folding rate are in good agreements with the experiments consistent with previous studies [39]. We also invetigated size scaling which consistent with previous studies [16, 40]. However, the connection between the kinetics and underlying landscape topography was not established in previous studies.

In Figure 7b, we have shown that the logarithmic kinetic rates (at the temperature of minimum folding kinetics) scaled with the size of the protein $ln\tau_{min}/N^{2/3}$ is strongly negatively correlated with the protein folding landscape topological parameter Λ (the energy gap versus roughness scaled with entropy). This implies that the more steep of the folding funnel or the

less rugged the protein folding funnel or the less entropy (size) of the folding funnel, the faster the protein folding kinetics is. Folding kinetics can be measured from the experiments. So we can see the kinetics of proteins measured in experiments also reflects the underlying topology of the protein folding energy landscape in a quantitative way (by comparing different proteins).

So we can see the protein folding landscape can be quantified by the topological parameter Λ (the energy gap versus roughness scaled with entropy) which is strongly correlated with the thermodynamics and kinetics of folding. We can by comparing different proteins infer the underlying folding topology through the thermodynamic and kinetic experiments. On the other hand, we can also obtain the underlying landscape topography by computations. In this way, we can predict the folding thermodynamics and kinetics in a reliable way. These theoretical predications can be directly verified by the experiments.

We notice that previous studies have shown the folding kinetic rates is correlated to the contact order, an entropic measure from native structure of proteins [41, 42]. This correlation is good for fast folding proteins where the topology essentially determines the folding mechanisms. Our landscape shape measure Λ as the energy gap versus roughness scaled with entropy is more complete since it is suitable for fast folding proteins where topology of structures are dominant. Since Λ includes not only entropic factor but also energetic contributions through energy gap and roughness, it can also quantitatively describe the folding thermodynamics and kinetics when the protein folding is not necessarily fast and determined by the structure topology alone (the energetics of the underlying interactions must kick in for these type of proteins).

Conclusions

We quantify the energy landscape of protein folding. This is realized through the exploration of the underlying density of states. By exploring the folding dynamics with lattice model, off lattice structure based model, and all atom force field models at different temperatures, we can obtain the thermodynamic and kinetic information. By converting the distribution of the energies from the results of the simulations in the canonical ensembles to the distribution of the energies in micro-canonical ensemble, we can obtain directly the underlying density of states of protein folding.

Based on the density of states, we can identify three quantities essential for characterizing the folding landscape topography: the energy gap between the native state and the average non-native states $\delta E = |E_n - \bar{E_{non-native}}|$, the roughness or variance in the energies of the non-

native states measured by the standard deviation of the energies in non-native states $\Delta E = \sqrt{< E^2 > - < E >^2}$, and the size of the funnel as measured by the entropy of the system $S = k_B ln\Omega$ (where Ω is the number of the states). We show that the dimensionless ratio between the gap, roughness and entropy of the system $\Lambda = \frac{\delta E}{\Delta E \sqrt{2S}}$ can quantify the degree of the funnel and determine the thermodynamics as well as kinetics of folding.

We found the following results: (1) We discovered that in the samples of proteins we investigated, all of them have funneled folding energy landscapes. (2) The ratio Λ characterizing the topology of the underlying folding landscape provides a new way to classify different proteins. (3) Further more the topology parameter Λ is shown to be correlated strongly (monotonically) with the thermodynamic stability against traps characterized by the ratio of thermodynamic stability temperature versus trapping temperature. (4) The landscape topology parameter Λ also monotonically correlates with the folding kinetic rates scaled by the sizes of the proteins.

This study bridges the gap between the quantification of underlying folding energy landscape topography and the thermodynamics and kinetics of protein folding.

Acknowledgments

The authors thank Prof. Peter G. Wolynes and Prof. Ben Schuler for helpful discussions. J.C and R.J.O were supported by CAPES, Brazil. R.J.O, V.B.P.L and J.C were partially supported by the Brazilian agency CNPq. V.B.P.L and R.J.O were supported by FAPESP, Brazil. J.W was partially supported by NSF Career Award, and NSFC (China). P.C.W thanks US National Science Foundation I2CAM International Materials Institute Award, Grant DMR-0645461, for funding this international collaboration. P.C.W is funded by a LANL Director's Fellowship. This work was also supported by the Center for Theoretical Biological Physics sponsored by the NSF (Grant PHY-0822283) with additional support from NSF – MCB-0543906.

Tables

Table 1: Quantities extracted from the density of states of the lattice model for different hydrophobicity degree.

hydrophobicity	**low**	**high**
E_n	-84.0	-84.0
$\langle E \rangle$	1.84	-9.04
$\langle E^2 \rangle$	29.24	93.89
δE	-85.84	-74.96
ΔE	5.09	3.49
S_o	35.89	33.07
T_f	1.80	1.50
T_g	0.60	0.43
T_f/T_g	3.00	3.49
$\delta E/\Delta E$	16.88	21.45
Λ	1.99	2.64

[a] The sequence in study is the well known three letter code 0012 with 27 monomers.
[b] The characteristic glass temperature T_g is defined as $T_g = \sqrt{\Delta E^2/2S_o}$.

Table 2: Quantities extracted from the density of states.

Protein (PDB)	**trp-cage** (1L2Y)	**villin** (1YRF)	**FBP28** (1E0L)	**albumin** (1PRB)	**protein A** (1BDD)	**SH3** (1FMK)	**CI2** (1YPA)	**CspTm** (1G6P)	**α_3D** (2A3D)
Λ	1.58	2.61	2.99	4.13	5.03	5.89	5.31	5.53	6.28
E_n	3.0	-1.0	-18.0	-13.0	29.0	-55.0	-42.0	-129	-23.0
$\langle E \rangle$	85.34	165.49	163.36	224.97	249.22	286.37	287.67	207.53	303.86
$\langle E^2 \rangle$	7308.23	27411.54	26701.51	50624.00	62117.37	82013.73	82763.78	43076.47	92335.69
δE	82.34	166.49	181.36	237.97	220.22	341.37	329.67	336.53	326.86
ΔE	5.08	4.84	3.83	3.48	2.79	2.78	3.01	2.78	2.52
S_0	52.42	87.12	124.96	136.82	123.38	217.48	213.42	239.42	212.69
T_f	0.94	0.97	1.01	1.04	0.89	1.14	1.06	1.11	1.00
T_g	0.50	0.37	0.24	0.21	0.18	0.13	0.15	0.13	0.12
$\frac{T_f}{T_g}$	1.90	2.63	4.17	4.93	5.03	8.56	7.29	8.74	8.15
$\frac{\delta E}{\Delta E}$	16.22	34.39	47.35	68.39	78.98	122.88	109.68	121.11	129.5
T_0	0.50	0.45	0.60	0.75	0.70	0.95	0.90	0.90	0.80
$\frac{T_0}{T_g}$	1.01	1.23	2.48	3.57	3.94	7.13	6.19	6.69	6.54
# res.	20	35	36	53	60	61	64	66	73

[a]The C_α structure based (Gō) model was implemented in the simulations and the CSU software was used for each protein contact map.
[b] The characteristic glass temperature T_g is defined as $T_g = \sqrt{\Delta E^2/2S_o}$

Table 3: Quantities extracted from the density of states of the protein CspTm (pdb 1G6P) with increasing non-native energy interaction[a].

ϵ_{NC}	**0.0**	**0.1**	**0.2**[c]	**0.3**	**0.4**	**0.5**
Λ	5.53	5.69	6.14	5.81	5.78	5.83
E_n	-129.0	-132.0	-136.0	-140.0	-142.0	-145.0
$\langle E \rangle$	207.53	204.63	201.92	200.60	201.27	205.06
$\langle E^2 \rangle$	43076.47	41880.76	40779.76	40249.25	40518.52	42057.77
δE	336.53	336.63	337.92	340.60	343.27	350.06
ΔE	2.78	2.66	2.52	2.68	2.73	2.86
S_o	239.42	246.29	238.62	238.76	235.62	220.88
T_f	1.110	1.120	1.124	1.115	1.088	1.050
T_g	0.13	0.12	0.12	0.12	0.13	0.14
$\frac{T_f}{T_g}$	8.74	9.33	9.75	9.09	8.64	7.73
$\frac{\delta E}{\Delta E}$	121.11	126.35	134.16	127.01	125.54	122.59
T_0	0.90	0.92	0.98	0.98	0.98	1.00
$\frac{T_0}{T_g}$	7.09	7.66	8.5	7.99	7.78	7.36
# res.	66	66	66	66	66	66

[a] The C_α structure based (Gō) model was implemented in the simulations and the CSU software was used for each protein contact map.
[b] The characteristic temperature T_g is defined as $T_g = \sqrt{\Delta E^2 / 2S_o}$.
[c] This is the optimal non-native energy parameter.

Table 4: Quantities extracted from the density of states using force field[a].

Protein	**trp-cage**	**beta3s**

[a] The model implemented in the simulations was the Replica-Exchange Molecular Dynamics (REMD) with force field ff99SB in AMBER 10.
[b] The characteristic temperature T_g is defined as $T_g = \sqrt{\Delta E^2 / 2S_o}$.
[c] This is the optimal non-native energy parameter.

Figures

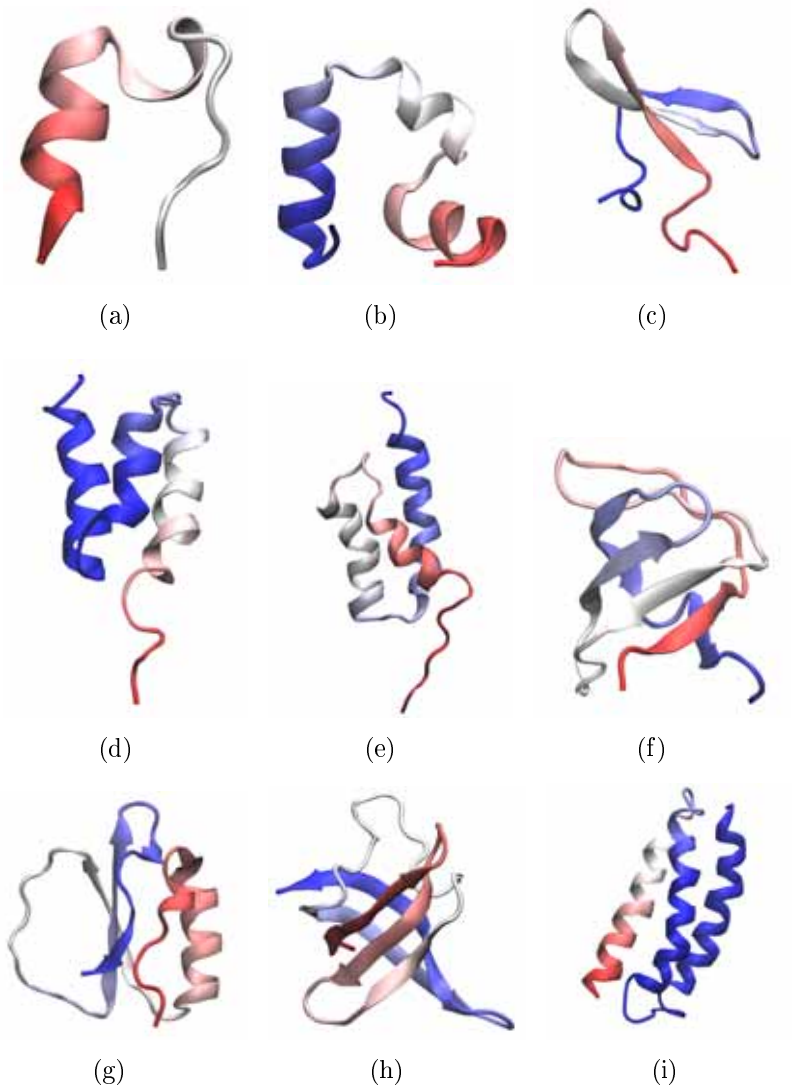


Figure 1: The proteins structure used in this study for the folding simulations are shown in cartoon representation with its Protein Data Base code (PDB) a) tryptophan cage (1L2Y) b) villin headpiece subdomain (1YRF) c) ww domain FBP28 (1E0L) d) albumin binding domain (1PRB) e) protein A (1BDD) f) src homology 3 domain SH3 (1FMK) g) chymotrypsin inhibitor 2 CI2 (1YPA) h) cold shock protein CspTm (1G6P) i) α_3D (2A3D). The structures were created using the package Visual Molecular Dynamics (VMD) [43] and are colored by an index along the chain from red (N-terminus) to blue (C-terminus).

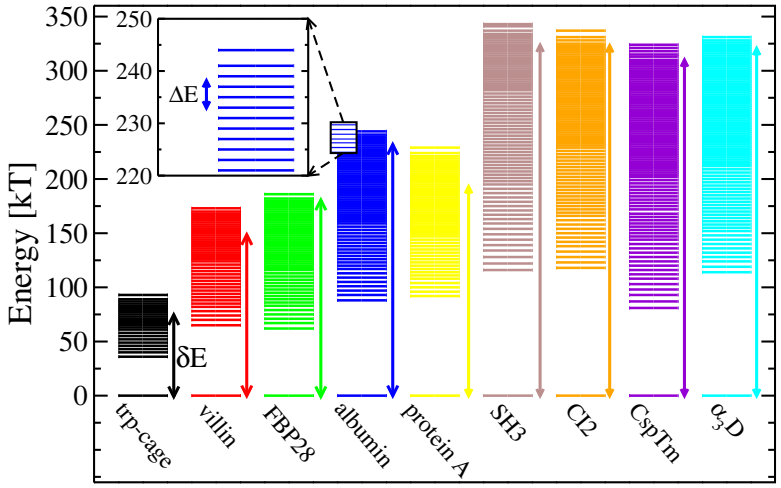
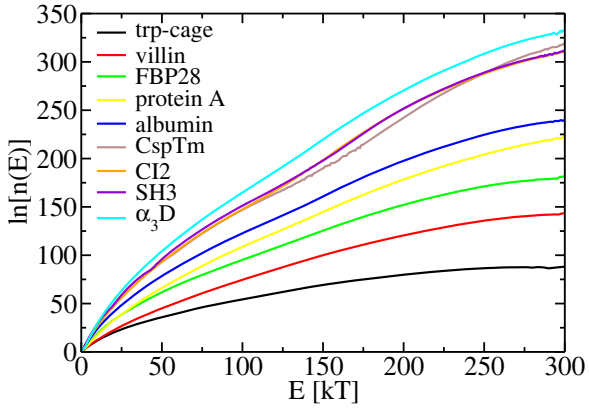
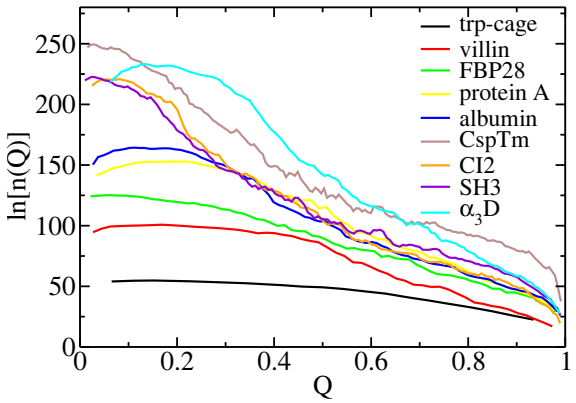


Figure 2: The distribution of the energy levels (energy spectrum) which resembles the energy landscape funnel in zero dimension (0d-funnel) for all proteins. The lowest (native) energy E_n is relocated to be at $E_n = 0$ for a better visualization purpose. The stability gab δE between the native energy and the average energy mode is indicated in a vertical arrow for each protein. Each energy level of the distribution represents a bin of a sum of e^{100} states except for the native energy level. The inset is a magnification of the latest energy levels distribution of albumin which shows how close the levels are at high energy. The standard deviation of the energy (roughness of the energy landscape) ΔE is also indicated as a vertical arrow. Energy is in unit of kT.





(b)

Figure 3: Logarithm of the density of states, the energy landscape funnel in one dimension (1d-funnel), as a function of a) the total protein energy and b) fraction of native contacts for each protein. The lowest (native) energy E_n is relocated to be at $E_n = 0$ for a better comparison purpose. Energy is in unit of kT.

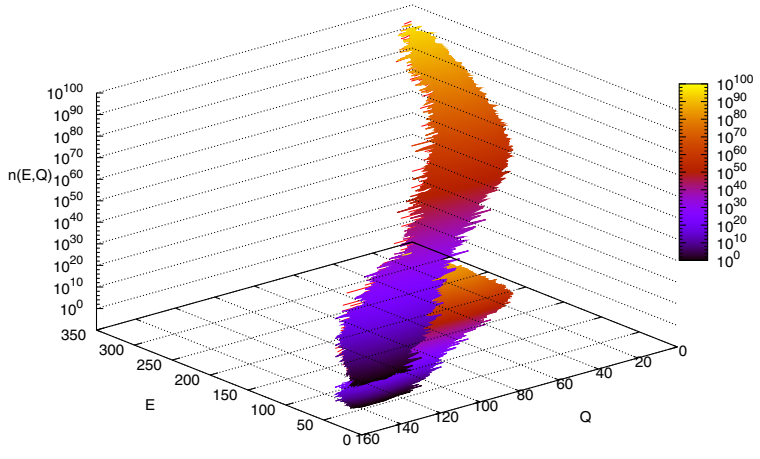


Figure 4: The distribution of density of states in logarithmic scale, the energy landscape funnel in two dimension (2d-funnel), as a function of the protein total energy E and the number of native contacts Q, $n(E,Q)$, for CI2. Energy is in unit of kT and the colorbar grows exponentially from 1 (purple) to 10^{100} (yellow).

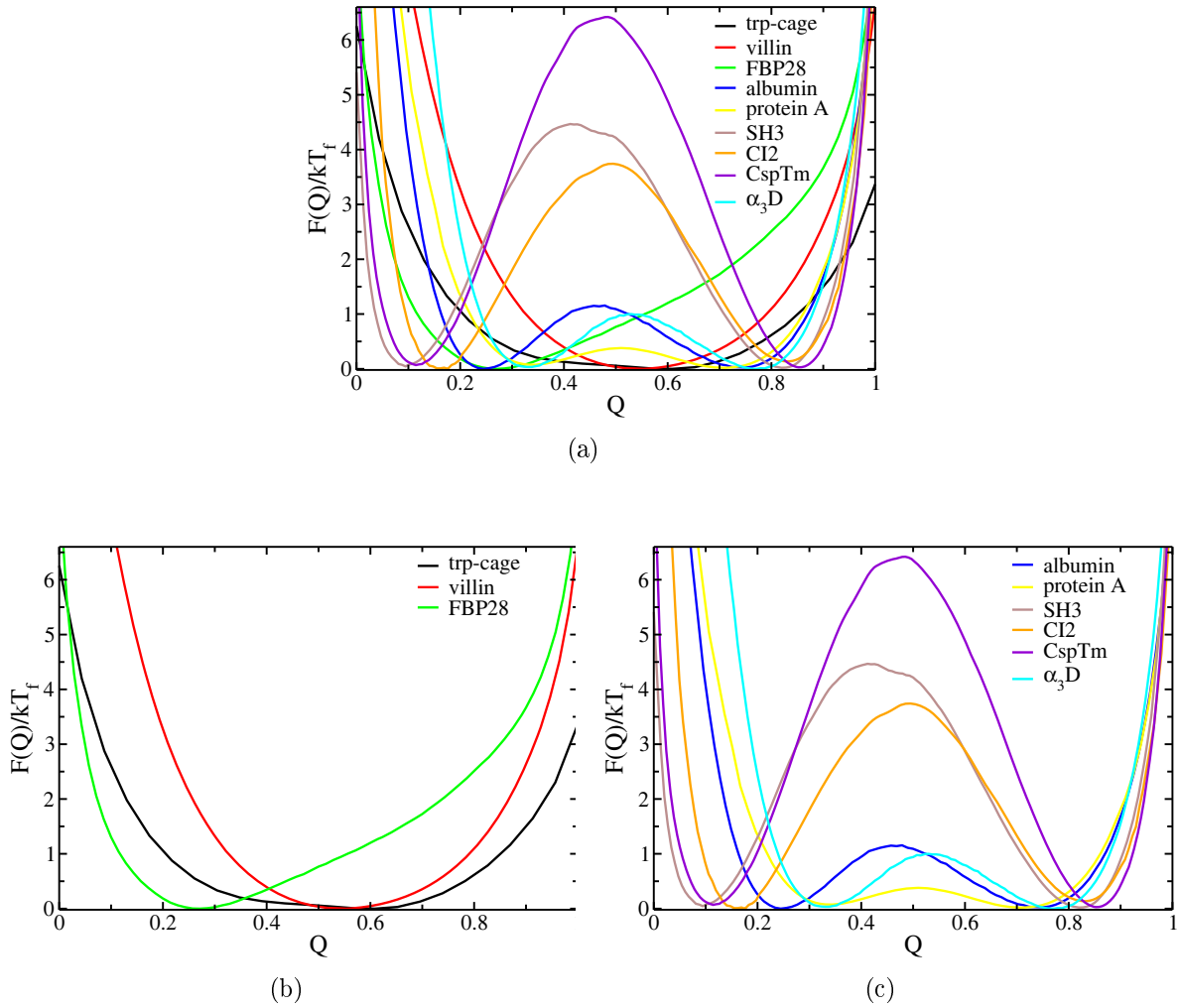


Figure 5: Free energy profile as a function of the reaction coordinate fraction of similarity degree with the native protein structure $F(Q)$ for the structure based C_α model for a) all proteins in study b) only the one-state downhill protein folding and c) two-state folder. Each curve is a free energy at $T = T_f$ and normalized by the respective kT_f.

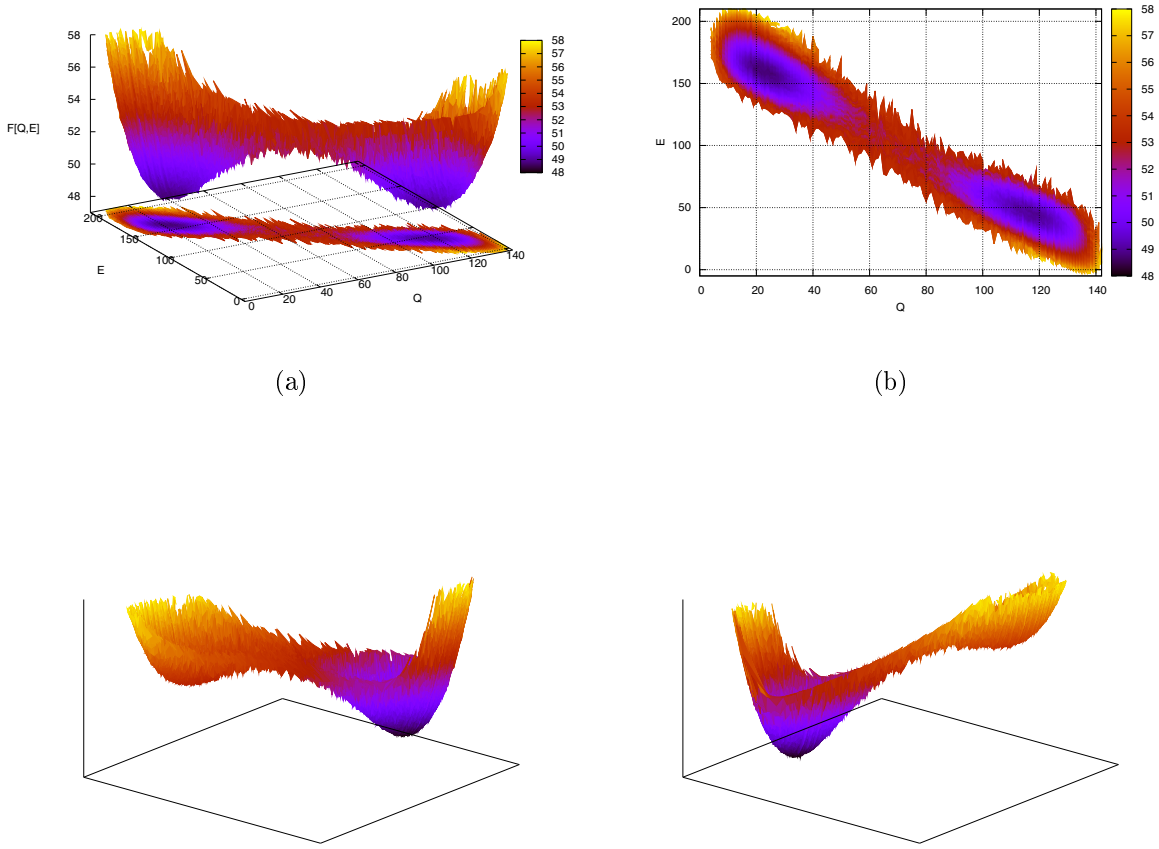
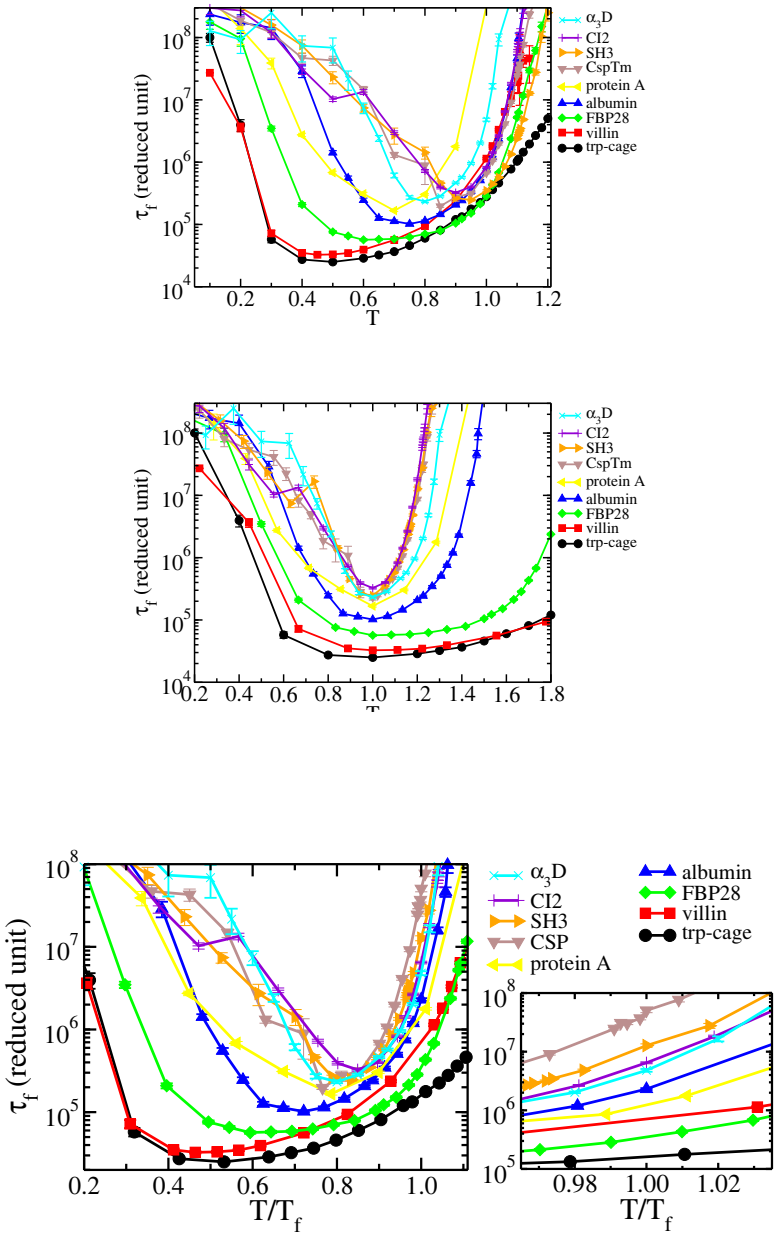


Figure 6: Free energy as a function of the total protein energy E and the reaction coordinate number of native protein Q, $F(E,Q)$, for CI2. a) The free energy surface with the color map at the bottom and b) the F surface with the view from the top as the same as the color map from a). a) and b) the temperature is at $T = T_f$. c) The free energy surface for $T < T_f$ and d) $T > T_f$. Free energy is in unit of kT and the colorbar increases from purple (free energy minima) to yellow (highest free energy) with red been at the transition state region. The white region is not probed by the protein.



(c)

Figure 7: For the proteins in study, it is shown the U-shape dependence of the mean first passage time (MFPT) τ_f as a function of a) absolute simulation temperature T b) normalized temperature T/T_0 and c) normalized temperature T/T_f. T_0 is the optimal temperature for folding, *i.e.*, the temperature where the protein τ_f is minimal and T_f is the folding temperature. The error bar is the standard deviation of the mean and the maximum simulation time is $\tau_{max} = 5 \times 10^8$ in reduced unit.

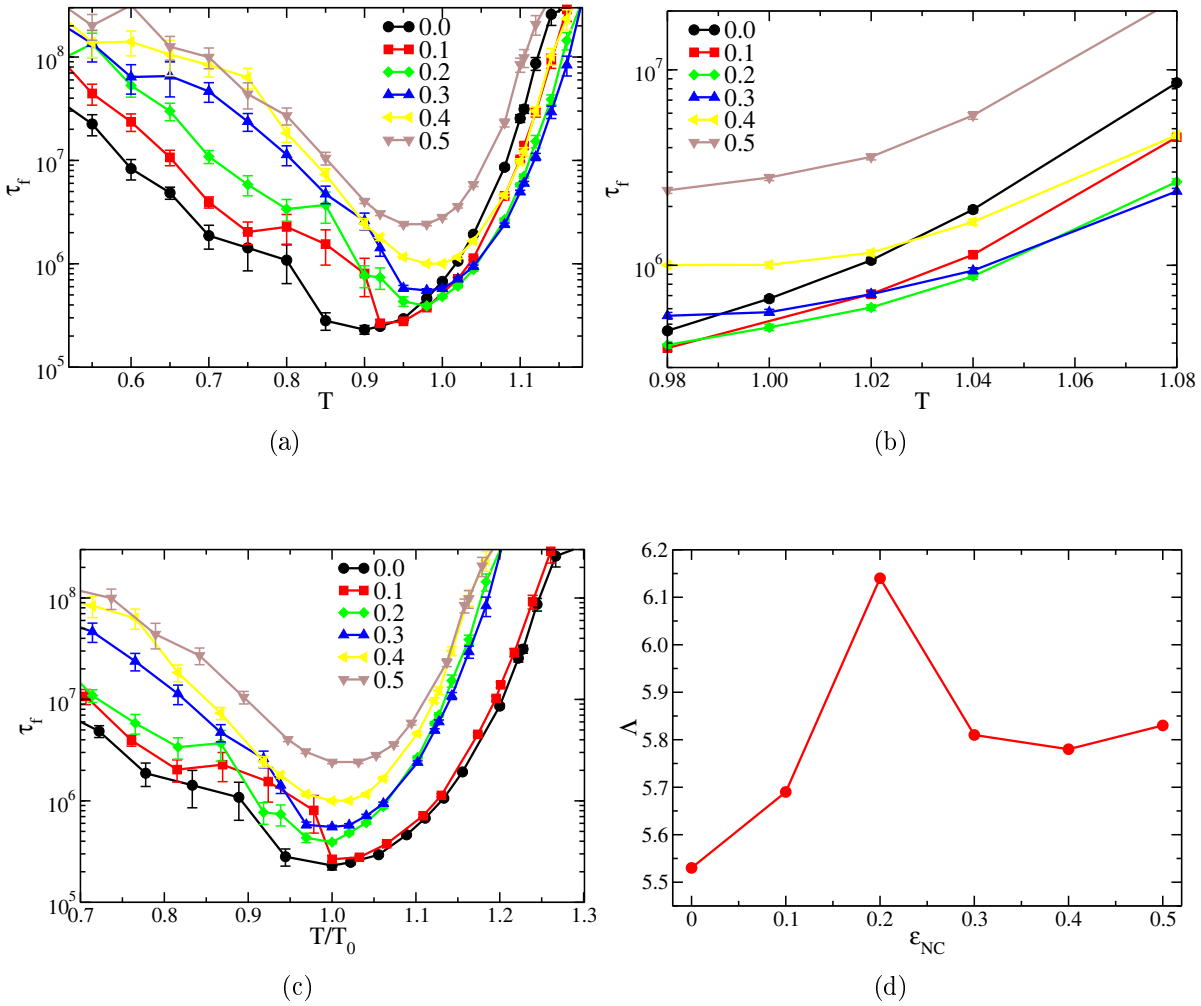
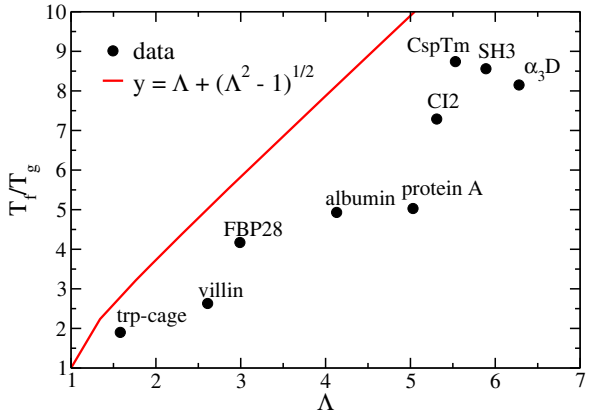
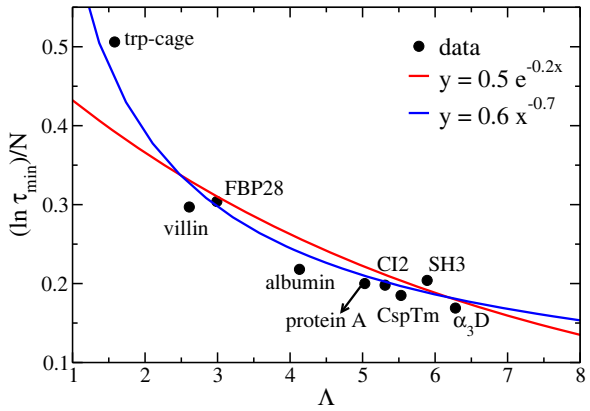


Figure 8: The CspTm results with different strength for the energetic frustration term ϵ_{NC}. It is shown the U-shape dependence of the mean first passage time (MFPT) τ_f as a function of a) and b) absolute simulation temperature T c) normalized temperature T/T_0. T_0 is the optimal temperature for folding, *i.e.*, the temperature where the protein τ_f is minimal and T_f is the folding temperature. a) is magnificatied in b) where it is possible to notice that around T_f, τ_f is optimum at $\epsilon_{NC} = 0.2$ which is an opposite trend than in d). For a a) b) and c), the error bar is the standard deviation of the mean and the maximum simulation time is $\tau_{max} = 5 \times 10^8$ in reduced unit.



(a)



(b)

Figure 9: a) The ratio between folding phase transition and glassy trapping temperatures T_f/T_g as a function of Λ, the ratio of gap to roughness modulated by entropy for all proteins. b) The logarithm of mininum mean first passage time τ_{min} scaled by the number of residues N versus Λ for each protein. Two fitted functions to the data are also shown in b). The data are from the structure based C_α simulations.

References

[1] C. Levinthal. Mossbauer spectroscopy in biological systems. In P. Debrunner, J. Tsibris, and E. Munch, editors, *Proceedings of a meeting held at Allerton house, Monticello, Illinois*, pages 22–24. University of Illinois Press, Urbana, 1969.

[2] J. D. Bryngelson and P. G. Wolynes. Spin-glasses and the statistical mechanics of protein folding. *Proc. Natl. Acad. Sci. USA*, 84:7524–7528, 1987.

[3] J. D. Bryngelson and P. G. Wolynes. Intermediates and barrier crossing in a random energy-model (with applications to protein folding). *J. Phys. Chem.*, 93(19):6902–6915, 1989.

[4] Hue Sun Chan and Ken A. Dill. Transition states and folding dynamics of proteins and heteropolymers. *The Journal of Chemical Physics*, 100(12):9238, 1994.

[5] V. I. Abkevich, A. M. Gutin, and E. I. Shakhnovich. Free energy landscape for protein folding kinetics: Intermediates, traps, and multiple pathways in theory and lattice model simulations. *J. Chem. Phys.*, 101(3):6052–6062, 1994.

[6] 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(3):167–195, March 1995. PMID: 7784423.

[7] Jin Wang, Jose Onuchic, and Peter Wolynes. Statistics of kinetic pathways on biased rough energy landscapes with applications to protein folding. *Physical Review Letters*, 76(25):4861, June 1996.

[8] Cecilia Clementi, Hugh Nymeyer, and José Nelson 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. *Journal of Molecular Biology*, 298(5):937–953, May 2000.

[9] H Frauenfelder, F Parak, and R D Young. Conformational substates in proteins. *Annual Review of Biophysics and Biophysical Chemistry*, 17(1):451–479, 1988.

[10] H Frauenfelder, SG Sligar, and PG Wolynes. The energy landscapes and motions of proteins. *Science*, 254(5038):1598–1603, December 1991.

[11] Laura S. Itzhaki, Daniel E. Otzen, and Alan R. Fersht. The structure of the transition state for folding of chymotrypsin inhibitor 2 analysed by protein engineering methods: Evidence

for a nucleation-condensation mechanism for protein folding. *Journal of Molecular Biology*, 254(2):260–288, November 1995.

[12] A. M. Gutin, V. I. Abkevich, and E. I. Shakhnovich. Chain length scaling of protein folding time. *Physical Review Letters*, 77(27):5433, December 1996.

[13] Flavio Seno, Cristian Micheletti, Amos Maritan, and Jayanth R. Banavar. Variational approach to protein design and extraction of interaction potentials. *Physical Review Letters*, 81(10):2172, 1998.

[14] D. K. Klimov and D. Thirumalai. Linking rates of folding in lattice models of proteins with underlying thermodynamic characteristics. *The Journal of Chemical Physics*, 109(10):4119, 1998.

[15] J. Sabelko, J. Ervin, and M. Gruebele. Observation of strange kinetics in protein folding. *Proceedings of the National Academy of Sciences of the United States of America*, 96(11):6031 –6036, May 1999.

[16] Marek Cieplak, Trinh Xuan Hoang, and Mai Suan Li. Scaling of folding properties in simple models of proteins. *Physical Review Letters*, 83(8):1684, 1999.

[17] Hüseyin Kaya and Hue Sun Chan. Energetic components of cooperative protein folding. *Physical Review Letters*, 85(22):4823, November 2000.

[18] Hüseyin Kaya and Hue Sun Chan. Towards a consistent modeling of protein thermodynamic and kinetic cooperativity: how applicable is the transition state picture to folding and unfolding? *Journal of Molecular Biology*, 315(4):899–909, January 2002.

[19] Benjamin Schuler, Everett A. Lipman, and William A. Eaton. Probing the free-energy surface for protein folding with single-molecule fluorescence spectroscopy. *Nature*, 419(6908):743–747, October 2002.

[20] Everett A. Lipman, Benjamin Schuler, Olgica Bakajin, and William A. Eaton. Single-Molecule measurement of protein folding kinetics. *Science*, 301(5637):1233–1235, August 2003.

[21] Houbi Nguyen, Marcus Jäger, Alessandro Moretto, Martin Gruebele, and Jeffery W. Kelly. Tuning the free-energy landscape of a WW domain by temperature, mutation, and truncation. *Proceedings of the National Academy of Sciences of the United States of America*, 100(7):3948 –3953, April 2003.

[22] Haw Yang and X. Sunney Xie. Probing single-molecule dynamics photon by photon. *The Journal of Chemical Physics*, 117(24):10965, 2002.

[23] Chi-Lun Lee, Chien-Ting Lin, George Stell, and Jin Wang. Diffusion dynamics, moments, and distribution of first-passage time on the protein-folding energy landscape, with applications to single molecules. *Physical Review E*, 67(4):041905, April 2003.

[24] Chi-Lun Lee, George Stell, and Jin Wang. First-passage time distribution and non-Markovian diffusion dynamics of protein folding. *The Journal of Chemical Physics*, 118(2):959, 2003.

[25] Yaoqi Zhou, Chi Zhang, George Stell, and Jin Wang. Temperature dependence of the distribution of the first passage time: Results from discontinuous molecular dynamics simulations of an All-Atom model of the second beta-Hairpin fragment of protein G. *Journal of the American Chemical Society*, 125(20):6300–6305, May 2003.

[26] J Wang. The complex kinetics of protein folding in wide temperature ranges. *Biophysical Journal*, 87(4):2164–2171, 2004.

[27] Vitor B. P. Leite, José N. Onuchic, George Stell, and Jin Wang. Probing the kinetics of single molecule protein folding. *Biophysical Journal*, 87(6):3633–3641, December 2004. PMID: 15465871 PMCID: 1304877.

[28] S. Walter Englander. PROTEIN FOLDING INTERMEDIATES AND PATHWAYS STUDIED BY HYDROGEN EXCHANGE. *Annual Review of Biophysics and Biomolecular Structure*, 29(1):213–238, 2000.

[29] Jon Rumbley, Linh Hoang, Leland Mayne, and S. Walter Englander. An amino acid code for protein folding. *Proceedings of the National Academy of Sciences of the United States of America*, 98(1):105 –112, January 2001.

[30] P. Wolynes, J. Onuchic, and D Thirumalai. Navigating the folding routes. *Science*, 267(5204):1619–1620, 1995.

[31] Jin Wang and Gennady M. Verkhivker. Energy landscape theory, funnels, specificity, and optimal criterion of biomolecular binding. *Physical Review Letters*, 90(18):188101, May 2003.

[32] Jin Wang, Weimin Huang, Hongyang Lu, and Erkang Wang. Downhill kinetics of biomolecular interface binding: Globally connected scenario. *Biophysical Journal*, 87(4):2187–2194, October 2004. PMID: 15454421 PMCID: 1304644.

[33] Jin Wang, Chilun Lee, and George Stell. The cooperative nature of hydrophobic forces and protein folding kinetics. *Chemical Physics*, 316(1-3):53–60, September 2005.

[34] Jin Wang, Li Xu, and Erkwang Wang. Optimal specificity and function for flexible biomolecular recognition. *Biophysical Journal*, 92(12):L109–L111, June 2007.

[35] Nicholas D. Socci and José Nelson Onuchic. Kinetic and thermodynamic analysis of proteinlike heteropolymers: Monte carlo histogram technique. *The Journal of Chemical Physics*, 103(11):4732, 1995.

[36] Corey Hardin, Michael P. Eastwood, Michael C. Prentiss, Zadia Luthey-Schulten, and Peter G. Wolynes. Associative memory hamiltonians for structure prediction without homology: α/β proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 100(4):1679 –1684, February 2003.

[37] Daniel Nettels, Irina V. Gopich, Armin Hoffmann, and Benjamin Schuler. Ultrafast dynamics of protein collapse from single-molecule photon statistics. *Proceedings of the National Academy of Sciences*, 104(8):2655 –2660, February 2007.

[38] Peter G. Wolynes. Folding funnels and energy landscapes of larger proteins within the capillarity approximation. *Proceedings of the National Academy of Sciences of the United States of America*, 94(12):6170 –6175, June 1997.

[39] Leslie L. Chavez, José N. Onuchic, and Cecilia Clementi. Quantifying the roughness on the free energy landscape: Entropic bottlenecks and protein folding rates. *Journal of the American Chemical Society*, 126(27):8426–8432, July 2004.

[40] Nobuyasu Koga and Shoji Takada. Roles of native topology and chain-length scaling in protein folding: A simulation study with a Gō-like model. *Journal of Molecular Biology*, 313(1):171–180, October 2001.

[41] Steven S. Plotkin, Jin Wang, and Peter G. Wolynes. Statistical mechanics of a correlated energy landscape model for protein folding funnels. *The Journal of Chemical Physics*, 106(7):2932, 1997.

[42] Carol A. Rohl, Charlie E.M. Strauss, Kira M.S. Misura, David Baker, Ludwig Brand, and Michael L. Johnson. Protein structure prediction using rosetta. In *Numerical Computer Methods, Part D*, volume Volume 383, pages 66–93. Academic Press, 2004.

[43] W. Humphrey, A. Dalke, and K. Schulten. VMD: Visual Molecular Dynamics. *J. Mol. Graphics*, 14(1):33–38, 1996.

Autorizo a reprodução xenográfica para fins de pesquisa.

São José do Rio Preto, 1 de Agosto de 2011.