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

Paulo Ricardo Mouro

Estudo da interação não-nativa no enovelamento de proteínas

São José do Rio Preto
2014

Paulo Ricardo Mouro

Estudo da interação não-nativa no enovelamento de proteínas

Dissertação apresentada como parte dos requisitos para obtenção do título de Mestre em Biofísica Molecular, junto ao Programa de Pós-Graduação em Biofísica Molecular, do Instituto de Biociências, Letras e Ciências Exatas da Universidade Estadual Paulista "Júlio de Mesquita Filho", Campus de São José do Rio Preto.

Orientador: Prof. Dr. Vitor Barbanti Pereira Leite

Co-orientador: Prof. Dr. Ronaldo Junio de Oliveira

São José do Rio Preto
2014

Mouro, Paulo Ricardo.
Estudo da interação não-nativa no enovelamento de proteínas / Paulo Ricardo Mouro. -- São José do Rio Preto, 2014
76 f. : il., gráfs., tabs.

Orientador: Vitor Barbanti Pereira Leite
Coorientador: Ronaldo Junio de Oliveira
Dissertação (mestrado) – Universidade Estadual Paulista “Júlio de Mesquita Filho”, Instituto de Biociências, Letras e Ciências Exatas

1. Biologia molecular. 2. Biofísica. 3. Proteínas globulares. 4. Dobramento de proteína. 5. Dinâmica molecular. I. Leite, Vitor Barbanti Pereira. II. Oliveira, Ronaldo Junio de. III. Universidade Estadual Paulista "Júlio de Mesquita Filho". Instituto de Biociências, Letras e Ciências Exatas. IV. Título.

CDU – 577.112

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

Paulo Ricardo Mouro

Estudo da interação não-nativa no enovelamento de proteínas

Dissertação apresentada como parte dos requisitos para obtenção do título de Mestre em Biofísica Molecular, junto ao Programa de Pós-Graduação em Biofísica Molecular do Instituto de Biociências, Letras e Ciências Exatas da Universidade Estadual Paulista "Júlio de Mesquita Filho", Campus de São José do Rio Preto.

Comissão Examinadora

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

Prof. Dr. Antonio Francisco Pereira de Araújo
UNB – Brasília

Prof. Dr. Jorge Chahine
UNESP – São José do Rio Preto

São José do Rio Preto
04 de abril de 2014

Agradecimentos

Agradeço a Deus, por ter me dado saúde, força e perseverança durante todos esses anos.

Aos professores Vitor e Ronaldo pela orientação e ensinamentos sem os quais este trabalho não poderia ser realizado.

Aos professores Jorge Chahine e Leandro pelas contribuições decisivas na formação deste trabalho.

Aos funcionários Paulinho, Barbosa, Bruno, Marcelino e Ilva pelo apoio técnico.

Aos meus familiares, principalmente ao meu pai, a minha mãe e a minha irmã, por não hesitarem em me dar apoio, carinho e conselhos em todos os momentos nos quais precisei.

A Rafaela, por estar sempre perto e enfrentar comigo todos os momentos deste trabalho.

Aos companheiros de grupo e aos amigos da pós-graduação.

Aos companheiros de moradia: Lucas, Mateus e Sérgio.

A CAPES pelo apoio financeiro.

Resumo

O estudo dos princípios físico-químicos que regulam o processo de enovelamento tornou-se central a fim de fornecer respostas sobre os mecanismos de enovelamento das proteínas. Neste contexto, a teoria do relevo da superfície de energia surge para apoiar os avanços teóricos e experimentais na compreensão destes mecanismos. O panorama energético de proteínas globulares se assemelha a um funil de estruturas que são progressivamente enoveladas para o estado nativo, estado minimamente frustrado. É bem estabelecido que a adição de uma pequena quantidade de frustração energética aumenta a velocidade de enovelamento para certas proteínas. Neste trabalho, aplicamos o modelo baseado em estrutura Cα para simular um grupo de proteínas utilizando a ordem de contato (CO) como coordenada de reação, e descobrimos que CO e barreira de energia livre no estado de transição (ΔF) correlacionam bem com a variação na quantidade de contatos não-nativos (ΔA) no regime de frustração ideal. Descobrimos também que, ΔF e ΔA separam as proteínas simuladas por seus "motifs". Estes resultados computacionais são corroborados por um modelo analítico. Como consequência, o regime de frustração ótima para o enovelamento de proteínas pode ser previsto analiticamente.

Palavras-chave: modelo baseado em estrutura, superfície de energia, dinâmica molecular.

Abstract

The study of physicochemical principles which governs the folding process became central in order to provide answers for the protein folding mechanism. In this context, the energy landscape theory has been supporting theoretical and experimental advances in the understanding this mechanism. The energy landscape of globular proteins resembles a funnel of structures progressively folded en route to the native state, minimally frustrated state. It is well established that an addition of small amount of energetic frustration enhances folding speed for certain proteins. We applied the Cα structure-based model to simulate a group of proteins with the contact order (CO) as the reaction coordinate and we found that CO and free energy barrier at the transition state (ΔF) correlates with nonnative contacts variation (ΔA) at the optimum frustration regime. We also found that ΔF and ΔA cluster the simulated proteins by their fold motifs. These computational findings are corroborated by analytical model. As a consequence, optimum frustration regime for protein folding can be predicted analytically.

keywords: structure-based model, energy landscape, molecular dynamics

Lista de Símbolos e Abreviações

A	Fração de contatos não-nativos;
ACO	Ordem de contato absoluta;
C_α	Carbono alfa;
CO	Ordem de contato parcial;
CSU	Contact of Structural Units;
E(T,Q,A)	Energia dependente da temperatura e das coordenadas de reação;
E_n	Energia do estado nativo;
F(T,Q,A)	Energia livre dependente da temperatura e das coordenadas Q e A;
GROMACS	GROningen MAchine for Chemical Simulations
k_β	Constante de Boltzmann;
L	Número de aminoácidos da proteína;
M	Número total de contatos;
M_Q	Soma de todos os contatos nativos que podem aparecer em Q;
N	Número de aminoácidos da proteína;
PDB	Protein Data Bank;
P-TS	Região de pré-transição;
Q	Fração de contatos nativos;
RCO	Ordem de Contato Relativa;
S_c	Entropia conformacional;
T_F^0	Temperatura de enovelamento;
TS	Região de transição;
Unf.	Região Desenovelada (unfolded);

WHAM	Weighted Histogram Analysis Method;
z	Contatos por resíduo;
$b^{\mathbf{2}}$	Variância energética não-nativa;
ϵ	Energia de Interação;
$\Delta S_{i,j}$	Distância entre os aminoácidos i e j que formam um contato nativo;
Γ	Estado conformacional de uma estrutura;
n(E,Q,A)	Número de estados com energia E de acordo com Q e A;
$\neq$	Indica que a grandeza está no estado de transição;
$\langle Q_i \rangle$	Probabilidade do contato i se formar;

Sumário

Capítulo 1

Introdução

As proteínas são polímeros de aminoácidos unidos por ligações peptídicas. Esta ligação é feita entre um grupo amino livre de um aminoácido com um grupo carboxilato de outro, formando uma ligação CO-NH com a liberação de uma molécula de água [1]. A ligação peptídica possui um caráter parcial de dupla ligação devido a estabilização por ressonância. Este caráter confere uma organização planar rígida e favorece a conformação trans da ligação peptídica. Estas característas atuam de forma importante na construção da conformação tridimensional das cadeias polipeptídicas.

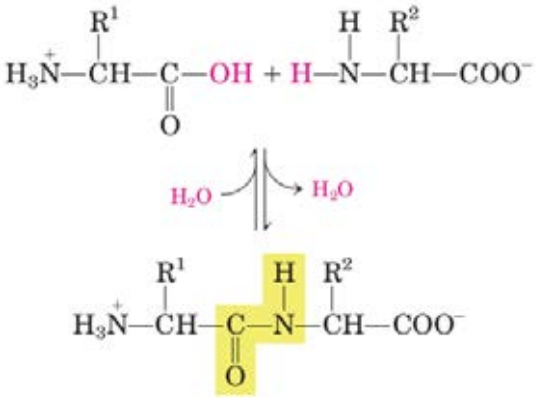

Figura 1.1: **Formação da ligação peptídica**. Reação de condensação entre dois aminoácidos, no qual o aminoácido com cadeia lateral R^1 libera um grupo OH simultaneamente com a liberação de um hidrogênio do aminoácido com cadeia lateral R^2. Ambos, OH e H que serão liberados, estão representados com uma coloração rosa na figura. Após a liberação, forma-se uma molécula de água e a ligação entre os dois aminoácidos é permitida (representada em amarelo).[2]

As proteínas podem conter uma ou mais cadeias polipeptídicas. As variações no comprimento e na sequência de aminoácidos contribuem para a diversidade da forma e

das funções biológicas desempenhadas por estas. Praticamente todas as transformações moleculares que definem o metabolismo celular são mediadas pela catálise proteica. As proteínas também exercem funções regulatórias e estruturais nas células [1].

Uma peça chave para decifrar a função de uma dada proteína é a compreensão da sua estrutura [1]. Os polímeros de aminoácidos são classificadas dentro de quatro níveis de complexidade estrutural, sendo eles: a estrutura primária, secundária, terciária e quaternária.

A estrutura primária de uma proteína consiste na sequência de aminoácidos da sua cadeia polipeptídica. Já na estrutura secundária inclui-se os padrões regulares de dobramentos de polipeptídeos, como as hélices, as folhas pregueadas e as voltas. A estrutura terciária de uma proteína descreve o dobramento dos elementos estruturais secundários e especifica as posições de cada átomo na proteína, incluindo os das cadeias laterais. As proteínas, particularmente as com massa molecular a cima de 100kD, são constituídas por mais de uma cadeia proteica. Essas subunidades polipeptídicas se arranjam em uma geometria específica, esse arranjo é conhecido como estrutura quaternária [1].

O processo que leva a proteína a sair de sua estrutura primária até atingir a estrutura terciária ou quaternária é chamado de enovelamento.

O enovelamento de proteínas é um mecanismo importante para os sistemas vivos [3]. Ultimamente, o foco sobre essa reação não está estritamente direcionado para o papel fundamental desempenhado pelo enovelamento [5], mas aponta para o conhecimento dos fatores físico-químicos que regulam o processo de enovelamento. Tais fatores podem ajudar no fornecimento de respostas para alguns dos problemas ainda não resolvidos em genômica funcional e biotecnologia, como por exemplo no desenho consistente de fármacos e enzimas, e no controle de doenças genéticas [8] e neurodegenerativas, incluindo as doenças de Alzheimer e Parkinson [9].

A problemática do enovelamento de proteínas está fundamentada em responder o porquê e como uma proteína consegue encontrar uma estrutura funcional única em um curto período de tempo [10]. Na década de 60, Chris Anfinsen propôs uma Hipótese Termodinâmica para o mecanismo do enovelamento de proteínas. Nesta hipótese foi apresentada a capacidade de um polipeptídeo desenovelado se enovelar espontaneamente para a estrutura funcional. Tal fato evidenciou que a informação suficiente para o correto enovelamento da proteína estava embutida na sequência de aminoácidos [11].

Em 1968, Cyrus Levinthal questionou sobre o tempo necessário para uma proteína efetuar o processo de enovelamento considerando a acessibilidade de todas as conformações possíveis para chegarmos a uma estrutura única de acordo com a sequência

de aminoácidos. Tal contexto ficou conhecido como o Paradoxo de Levinthal. O paradoxo se centralizava na exemplificação de que se este processo fosse feito por caminhos aleatórios, o tempo necessário para o enovelamento de uma proteína com cem aminoácidos e duas conformações para cada ligação peptídica era da ordem de 10^{18} segundos, o que é comparável com a idade do universo. Neste caso, Levinthal concluiu que era necessária a presença de rotas de enovelamento, ou seja, o processo não seria mais feito com base em uma busca aleatória, mas sim por um processo dirigido, o que excluiria a dependência temporal astronômica [12].

O até então paradoxo levantado por Levinthal foi resolvido com o surgimento de uma nova teoria para o processo de enovelamento de proteínas. Tal teoria é baseada na superfície do relevo de energia (energy landscape theory)[13, 14, 15, 16]. Conceitualmente, esta superfície apresenta uma forma "afunilada" e por meio de uma abordagem termodinâmica, considera-se que no topo deste funil há uma alta entropia dada a grande quantidade de estruturas desenoveladas possíveis para uma mesma proteína. A proteína é então direcionada para o estado nativo passando por estados parcialmente enovelados [18]. Uma representação esquemática da superfície de relevo de energia pode ser vista na figura 1.2.

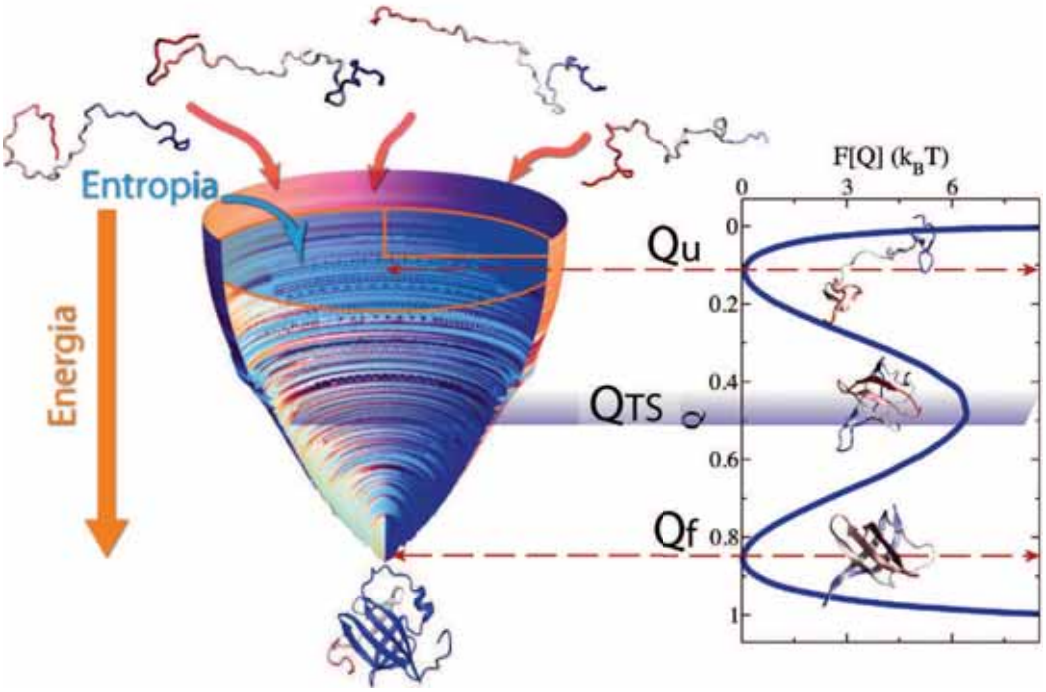

Figura 1.2: **Diagrama esquemático do relevo da superfície de energia.** Direita: Perfil de energia livre em função da fração de contatos nativos formados. Q_u = Conformação desenovelada. Q_{TS} = Conformação no estado de transição. Q_f = Conformação enovelada. Figura adaptada de [4].

O processo de enovelamento começa com a formação dos contatos entre os aminoácidos, e a medida que os contatos são feitos a proteína começa a ficar cada vez mais parecida estruturalmente com a proteína nativa, e a quantidade de estruturas possíveis diminuem em cada estágio que se avança no enovelamento. A medida que o processo flui de forma progressiva, a entropia e a energia do sistema também diminuem. Neste novo modelo (energy landscape) é permitida a existência de algumas regiões com mínimos locais de energia, e que podem "armadilhar" o enovelamento. Sendo assim, a curvatura do funil deve ser bastante íngreme a fim de superar estes pontos de mínimos locais e fazer com que a proteína atinja o estado de mínimo global de energia, o estado nativo [14, 18].

O comportamento cinético e termodinâmico de sistemas com uma superfície de relevo de energia irregular (ou rugoso) é bem distinto daqueles com uma superfície regular (ou lisa). Em casos no qual a superfície de energia é irregular há uma grande competição entre as interações, afetando a energia. Esta competição é chamada de frustração [19, 20]. De acordo com a teoria de "spin-glas" um sistema é dito frustrado quando não é capaz de satisfazer todas as interações de forma simultânea. No caso das proteínas naturais, elas são ditas minimamente frustradas. Esta denominação foi dada por Bryngelson e Wolynes na tentativa de diferenciar uma proteína de um heteropolímero aleatório, na qual diferentemente deste, as proteínas minimizam a frustração ao longo do enovelamento [17].

As fontes de frustração de uma proteína podem ser amplamente classificadas como energética e topológica [6]. A frustração energética está associada com a sequência de aminoácidos que compõe a proteína [6] e com a energia de interação entre os aminoácidos, na qual todas as interações favoráveis não podem ser acomodadas ao mesmo tempo [19]. Já a frustração geométrica está relacionada com a interconversão entre duas estruturas, visto que para ocorrer a interconversão há a necessidade de quebrar vários contatos favoráveis [19]. A frustração geométrica é, em parte, um problema de volume excluso [6]. Ambas as origens de frustração estão relacionadas com proteínas com barreira energética alta entre o estado desenovelado e enovelado [19].

Alguns trabalhos colocam em pauta um limiar sobre qual o efeito da inserção de frustração energética sobre o enovelamento, e procuram discutir como, ou por que, em alguns casos a frustração energética favorece e em outros ela é desfavorável ao processo de enovelamento [7, 8]. Tal favorecimento dado pelo acréscimo de frustração está baseado principalmente em dois resultados: o aumento na acessibilidade do estado nativo e a diminuição do tempo de enovelamento. Contessoto e colaboradores [7], por meio de dinâmica molecular utilizando um potencial acrescido de um termo referente à frustração energética, encontraram uma excelente correlação entre frustração ótima

e ordem de contato nativa, motivando este trabalho. Esta frustração ótima foi determinada para cada proteína estudada, e é calculada de acordo com o grau máximo de frustração que se pode acrescentar no modelo para que o tempo de enovelamento diminua. Mais detalhes sobre esta correlação podem ser vistos no Apêndice B.

Para acompanhar o processo de enovelamento por meio de dinâmica molecular há a necessidade de se utilizar uma coordenada de reação, que pode ser o raio de giro, a fração de contatos nativos (Q) e etc.

Neste trabalho iremos abordar o efeito da frustração energética de acordo com a variação de contatos não-nativos entre o estado de transição e o estado desenovelado para 19 proteínas, utilizando como parâmetro de ordem a fração de contatos nativos (Q) e a ordem de contato parcial (CO). Tais informações serão a base para comparar e verificar os dados obtidos por simulação computacional com a teoria descrita por Plotkin e Cecilia [8].

Capítulo 2

Objetivo

Este trabalho teve como principais objetivos:

• Verificar o comportamento da ordem de contato como coordenada de reação a fim de acompanhar o processo de enovelamento, com o intuito de introduzí-la no modelo analítico para o estudo da frustração no enovelamento de proteínas;

• Verificar a correlação entre o modelo teórico desenvolvido por Cecilia Clementi e Steven Plotkin [8] com os dados obtidos por meio de dinâmica molecular para 19 proteínas, correlacionando os resultados com a presença de uma frustração ótima determinada por Contessoto et all [7];

• Determinar a correlação entre a variação na quantidade de contatos não-nativos, altura da barreira de energia livre e ordem contato,determinando se o regime de frustração ótima para o enovelamento de proteínas pode ser previsto analiticamente.

Capítulo 3

Teoria

3.1 Modelo analítico para o enovelamento com frustração

Neste tópico será abordado o modelo analítico desenvolvido por Cecilia Clementi e Steven Plotkin [8]. Tal modelo foi criado para investigar o efeito da frustração energética sobre o enovelamento de acordo com o comportamento das interações não-nativa. As interações não-nativas são introduzidas como contatos adicionais entre pares que não estão em contato no estado nativo. Desta forma, é verificado o efeito destas interações sobre a taxa de enovelamento e sobre a barreira de energia livre, como será detalhado nos próximos parágrafos.

Dois parâmetros de ordem são utilizados para a descrição do processo de enovelamento no modelo analítico: a fração de contatos nativos (Q) e a fração de contatos não-nativos (A).

Estes dois parâmetros de ordem possuem valores entre 0 e 1. Para Q=0 a proteína encontra-se desenovelada enquanto que para Q=1, a proteína está em sua forma nativa. A fração de contatos não-nativos não é independente de Q, quanto mais contatos nativos formados, menos interações não-nativas são permitidas, e eventualmente, no modelo teórico, não é permitida a presença de contatos não-nativos quando Q=1.

Seja a energia do estado nativo dado por E_N e o número total de contatos no estado nativo dado por M, temos que:

$$E_N = M\epsilon = zN\epsilon, \tag{3.1}$$

no qual ϵ é definida como a energia média de atração nativa ($\epsilon < 0$). N e z são, respectivamente, o número de aminoácidos da proteína e a quantidade de contatos por resíduo. A variância da energia de interação nativa é desprezada ($\delta\epsilon^2 = 0$).

Já para a análise da contribuição não-nativa é considerado o efeito de duas energias que governam estas interações: a energia média de interação não-nativa ϵ_{NN} e a variância energética da interação não-nativa b^2. Porém, as interações não-nativas são designadas para serem mais fracas quando comparadas com as interações nativas.

Para uma configuração com MA contatos não-nativos, a energia não-nativa total é tomada para ser distribuída Gaussianamente com média em $MA\epsilon_{NN}$ e variância de MAb^2.

Seja a entropia conformacional S_c descrita em função de Q e A, as energias das configurações para um ensemble de estados caracterizados por (Q,A) também devem ser representadas por uma distribuição Gaussiana com média em $QM\epsilon + AM\epsilon_{NN}$ e variância de AMb^2. Neste caso, a parte extensiva do logaritmo do número de estados (n) com energia E e parâmetros de ordem (Q,A) é dado por:

$$ln[n(E,Q,A)] = S_c(Q,A) - \frac{[E-(QM\epsilon + AM\epsilon_{NN})]^2}{2AMb^2} \tag{3.2}$$

Sabendo que T^{-1}=$\delta S/\delta E$, e que a entropia S é dada pela equação 3.2, podemos encontrar a energia térmica, a energia livre e a entropia em função de Q, A e T:

$$T^{-1} = \frac{\delta S}{\delta E} = \frac{\delta ln(n)}{\delta E} = -\frac{2[E-(QM\epsilon + AM\epsilon_{NN})]}{2AMb^2} \tag{3.3}$$

logo,

$$\frac{1}{T} = \frac{-E}{AMb^2} + \frac{QM\epsilon}{AMb^2} + \frac{AM\epsilon_{NN}}{AMb^2} \Rightarrow \frac{E}{M} = \epsilon Q + \epsilon_{NN}A - \frac{AMb^2}{T} \tag{3.4}$$

Organizando os termos da equação 3.4, chegamos na expressão da energia térmica por contato:

$$\frac{E}{M} = \epsilon Q + A\left(\epsilon_{NN} - \frac{b^2}{T}\right) \tag{3.5}$$

Para encontrarmos a expressão da entropia em termos de Q, A e T, basta substituir o valor de E encontrado em 3.5 na expressão 3.2:

$$S(Q,A,T) = S_c(Q,A) - \left(\frac{b^2}{2T^2}\right) AM = M\left(\frac{s_c(Q,A)}{z}\right) - \left(\frac{b^2}{2T^2}\right) AM \tag{3.6}$$

sendo $s_c(Q,A)$ a entropia conformacional por resíduo. Neste caso, a entropia $S(Q,A,T)$ por contato é dada por:

$$\frac{S(Q,A,T)}{M} = \frac{s_c(Q,A)}{z} - \left(\frac{b^2}{2T^2}\right) A \tag{3.7}$$

Sabendo que a energia livre é dada pela diferença da energia térmica pela multiplicação da entropia pela temperatura,

$$\frac{F}{M} = \frac{E}{M} - \frac{TS}{M} \tag{3.8}$$

podemos substituir as expressões de E(Q,A,T) e S(Q,A,T) em 3.8 para encontrarmos a equação da energia livre para o modelo em questão:

$$\frac{F(Q,A,T)}{M} = \epsilon Q + \left(\epsilon_{NN} - \frac{b^2}{2T}\right) A - \frac{T s_c(Q,A)}{z} \tag{3.9}$$

A entropia conformacional $S_c(Q,A)$, de acordo com Clementi e Plotkin, é calculada utilizando a teoria de campo médio e é feita em termos da fração de empacotamento não-nativa η. Quando η=1 temos o número máximo de contatos não-nativos presentes $MA_{max} = M\eta(1-Q)$. Neste caso, a entropia conformacional de um polímero com Q contatos nativos e fração de empacotamento η é dada por:

$$S_c(Q,\eta) = N(1-Q)\left\{ ln\frac{\nu}{e} - \left(\frac{1-\eta}{\eta}\right) ln\,(1-\eta) - \frac{1}{6}\left[\left(\frac{\bar{\eta}(Q)}{\eta}\right)^{2/3} - 1\right]^2\right\} \tag{3.10}$$

$$S_c(Q,\eta) \equiv N(1-Q)s_{NN}(Q,\eta), \tag{3.11}$$

no qual $ln\left(\frac{\nu}{e}\right)$ é a entropia máxima por resíduos no estado colapsado e $\bar{\eta}(Q) = \bar{l}(Q)^{-1/2}$, no qual $\bar{l}$ é o comprimento médio do looping formado por contatos nativos em um determinado Q.

3.1.1 Efeito da interação não-nativa na barreira de energia

De acordo com o modelo analítico, no momento no qual o estado enovelado e desenovelado possuem a mesma probabilidade, dizemos que a proteína está na sua temperatura de enovelamento $T_F{}^{\mathrm{o}}$. Este estado é dado pela equação 3.9 quando $F(0,A) \approx F(1,0)$ e $\epsilon_{NN} = b^2 = 0$.

Considerando que $Q \approx 0$ no estado desenovelado e $A = 0$ no estado enovelado, a temperatura de enovelamento é dada por:

$$T_F{}^{\mathrm{o}} = \frac{z|\epsilon|}{s_c(0, A^*(0))} \tag{3.12}$$

no qual $A^*(Q)$ é o valor mais provável de A num determinado Q.

Utilizando as seguintes definições:

$$\Delta A^*(Q) \equiv A^*(Q) - A^*(0) \tag{3.13}$$

$$\Delta s_{nn}(Q) \equiv s_{nn}(Q, A^*(Q)) - s_{nn}(0, A^*(0)) \tag{3.14}$$

pode-se calcular a variação da energia livre entre o estado desenovelado e um estado com Q e A qualquer,

$$\Delta F(Q,T) \equiv F(Q, A^*(Q), T) - F(0, A^*(0), T) \tag{3.15}$$

No qual,

$$\frac{F(Q, A^*(Q), T)}{M} = \epsilon Q - \frac{T(1-Q)s_{nn}(Q, A^*(Q))}{z} + \left(\epsilon_{NN} - \frac{b^2}{2T}\right) A^*(Q) \tag{3.16}$$

e

$$\frac{F(0, A^*(0), T)}{M} = -\frac{T s_{nn}(0, A^*(0))}{z} + \left(\epsilon_{NN} - \frac{b^2}{2T}\right) A^*(0) \tag{3.17}$$

Sendo assim:

$$\frac{\Delta F(Q,T)}{M} = Q\left(\epsilon + \frac{T s_{nn}(0, A^*(0))}{z}\right) + \left(\epsilon_{NN} - \frac{b^2}{2T}\right) \Delta A^*(Q) - \frac{T(1-Q)\Delta s_{nn}(Q)}{z} \tag{3.18}$$

Calculando a variação da energia livre na temperatura de enovelamento $T_F{}^{\mathrm{o}}$, o primeiro termo da equação 3.18 se anula, e se considerarmos que $\epsilon_{NN} = b^2 = 0$, o último termo da expressão 3.18 se torna a variação de energia livre na ausência de forças não-nativas ΔF^{o}. Se ao invés de calcularmos ΔF sobre um Q qualquer, mas restringirmos para um Q na barreira de energia ($Q^{\neq}$), temos a seguinte relação:

$$\frac{\Delta F^{\neq}}{T_F{}^{\circ}} = \frac{\Delta F^{\circ\neq}}{T_F{}^{\circ}} + M\left(\frac{\epsilon_{NN}}{T_F{}^{\circ}} - \frac{b^2}{2T_F{}^{\circ 2}}\right)\Delta A^*(Q^{\neq}), \tag{3.19}$$

ou seja, a variação da energia livre entre a barreira e o estado desenovelado é igual a variação da energia livre entre os mesmos estados na ausência de forças não-nativas acrescida de um termo referente a frustração. O fato de $\epsilon_{NN} < 0$, faz com que o termo entre parênteses que multiplica a variação de contatos não-nativos na expressão acima, seja sempre negativo. Quando $\Delta A^*(Q^{\neq})$ for maior do que zero, temos uma diminuição da barreira de energia livre quando comparada com a barreira na ausência de interações não-nativas, assim sendo, as interações não-nativas favoreceriam o enovelamento. O contrário acontece quando $\Delta A^*(Q^{\neq}) < 0$, ou seja, em proteínas nas quais ocorre este último caso, as interações não-nativas fazem com que a barreira aumente, desfavorecendo o enovelamento. Além disso, a equação 3.19 pode ser usada para encontrar o valor das energias de interação não-nativa sabendo a alteração da barreira de energia livre.

3.2 Ordem de Contato

O conceito de Ordem de Contato foi introduzido por Plaxco, Simons e Baker [21] como um parâmetro que demonstrasse a importância de contatos locais ou não-locais para a proteína no estado nativo. A ordem de contato nada mais é do que uma média entre as distâncias dos pares que efetuam um contato. Apesar desta simplicidade matemática, existem grandes evidências de que as taxas de enovelamento para proteínas pequenas e com dois estados é bem correlacionada com a ordem de contato associadas com a estrutura nativa [22]. Se esta ordem de contato, chamada de absoluta, for dividida pelo número de aminoácidos presentes na proteína, o parâmetro agora passa a ser chamado de ordem de contato relativa.

Ordem de Contato Absoluta:

$$ACO = \frac{1}{M}\sum^{M}\Delta S_{i,j} \tag{3.20}$$

Ordem de Contato Relativa:

$$RCO = \frac{1}{M.L}\sum^{M}\Delta S_{i,j}, \tag{3.21}$$

M é a quantidade de contatos presentes no estado nativo, $\Delta S_{i,j}$ é a distância em resíduos que separa um contato formado pelos aminoácidos i e j, e L é a quantidade

de aminoácidos que compõe a proteína.

A ordem de contato também pode ser uma maneira de quantificar o enovelamento quando calculada em função do parâmetro de ordem Q. Neste caso, temos a ordem de contato parcial [22], dada pela seguinte expressão:

$$CO(Q) = \frac{\sum_{i=1}^{M_Q} L_i \langle Q_i \rangle}{M_Q}, \tag{3.22}$$

no qual M_Q é o número total de contatos nativos possíveis quando a proteína possui um parâmetro de ordem Q. L_i é a distância, em resíduos, ao longo da cadeia entre os aminoácidos que compõem o contato i. $\langle Q_i \rangle$ é a probabilidade que o contato i tem de se formar quando temos uma configuração com fração Q de contatos nativos formados.

Quando a ordem de contato parcial é calculada para o estado nativo, Q=1, esta se transforma na ordem de contato absoluta descrita por Plaxco e colaboradores.

Capítulo 4

Metodologia

4.1 Modelo Baseado em Estrutura

Para representar de forma aceitável alguns dos fenômenos desencadeados durante o enovelamento das proteínas utilizando simulação computacional, tem-se usado com grande sucesso modelos minimalistas, que apesar de não abrangerem todas as interações envolvidas no processo, são capazes de explicar os fundamentos básicos que regem o enovelamento. Dentre estes modelos minimalistas, destacamos o modelo Cα, que foi utilizado neste trabalho.

O modelo Cα é dito um modelo baseado em estrutura visto que utiliza valores obtidos da resolução estrutural de proteínas depositadas em bancos de dados, como o PDB [27, 28], a fim de construir o potencial energético que define as conformações de uma dada proteína. Durante a dinâmica no qual o modelo Cα é submetido, o potencial que define a energia das conformações é conhecido como modelo G$\bar{o}$ [25] uma vez que a ideia principal é dar importância para as interações entre aminoácidos que residem nos contatos nativos, e em seguida, escolher a energia dos contatos que minimizam a energia total do estado nativo [26].

A simplicidade do modelo Cα está baseada na referência à cadeia de aminoácidos como uma cadeia de esferas simples centralizadas nas posições dos carbonos alfas. Tais esferas são mantidas juntas por meio de ligações e ângulos de interação, e a geometria nativa fica contida no potencial diedro e no termo de interação não-local [26].

A expressão que define a energia de uma configuração Γ baseada na conformação nativa Γ^0 para o modelo Cα é dada por:

$$
\begin{aligned}
V(\Gamma,\Gamma^0) &= \sum_{bonds}\epsilon_r(r-r_0)^2+\sum_{angles}\epsilon_\theta(\theta-\theta_0)^2 \\
&+ \sum_{backbone}\epsilon_\phi\left\{\left[1-\cos_(\,\phi-\phi_0)\right]+\frac{1}{2}\left[1-\cos\left(3\left(\phi-\phi_0\right)\right)\right]\right\} \qquad (4.1)\\
&+ \sum_{contacts}\epsilon_c\left[5\left(\frac{d_{ij}}{r_{ij}}\right)^{12}-6\left(\frac{d_{ij}}{r_{ij}}\right)^{10}\right]+\sum_{non-cont}\epsilon_{NC}\left(\frac{\sigma_{NC}}{r_{ij}}\right)^{12}
\end{aligned}
$$

De acordo com a expressão 4.1, o potencial que rege este modelo baseado em estrutura é calculado via a contabilização de 5 termos referentes a algumas das interações que atuam na estrutura nativa. O primeiro termo remete a um potencial harmônico representando a ligação entre dois carbonos α adjacentes, no qual r_0 é a distância entre os dois carbonos α ligados entre si na estrutura nativa. A segunda somatória também forma um potencial harmônico, mas desta vez, um potencial harmônico angular formado por três carbonos α em sequência na cadeia polipeptídica, no qual θ_0 é o ângulo formado pelos três resíduos na conformação nativa. Já o terceiro termo da expressão 4.1, é uma função de três mínimos, que leva em consideração a torção realizada pela cadeia, no qual, ϕ_0 é o angulo diédro formado por quatro carbonos α em sequência. O quarto termo contabiliza a interação entre o carbono α i e j que formam um contato na estrutura nativa, para isso é utilizado um potencial 10-12, no qual d_{ij} é o valor da distância entre estes carbonos que realizam um contato nativo. Por fim, o último termo remete a todos os carbonos α que não realizam um contato nativo. Este termo é utilizado para manter a distância máxima de aproximação entre os carbonos α, no qual σ_{NC} possui valor de $4\mathring{A}$ e está correlacionado com o volume ocupado por um carbono no modelo. As constantes: ϵ_r, ϵ_θ, ϵ_ϕ e ϵ_{NC} são todas dadas em unidades de ϵ_c e valem, respectivamente, 100, 20, 1 e 1. [23, 25]

4.2 Proteínas estudadas

Para testar a correlação entre a diferença na quantidade de contatos não-nativos no estado de transição para com o estado desenovelado, a presença de uma frustração ótima, e a ordem de contato, utilizamos como objeto de estudo um grupo de 19 proteínas com características distintas, como por exemplo, o número de aminoácidos que compõem a cadeia, a quantidade de contatos na estrutura nativa e a ordem de contato absoluta. Todas as proteínas estudadas tiveram sua frustração ótima determinada por Contessoto e colaboradores [7] e podem ser vistas com mais detalhes no Quadro 1.

Quadro 1 Proteínas Estudadas

NOME	PDB	# a.a.	M	Cartoon	NOME	PDB	# a.a.	M	Cartoon
ACR	1ARR	53	58		PtG	2K0P	56	139	
HP36	1VII	36	55		ADA2h	1PBA	81	175	
PSBD	2PDD	43	64		CI2	1CIS	66	152	
α3D	2A3D	73	136		SH3	1FMK	61	152	
PtABD	1BDC	60	102		Ubiquitin	1UBQ	76	188	
EnHD	1ENH	104	111		CSPTm	1G6P	66	180	
IM9	1IMP	86	178		αAIT	2AIT	74	196	
ACBD	2ABD	86	182		HPr	1HDN	85	222	
HHCC	1HRC	104	246		TWIg	1WIU	93	253	
PtL	2PTL	60	136						

#a.a. = Número de aminoácidos na cadeia. M = Contatos Nativos

4.3 Detalhes da Simulação

Antes de iniciar a simulação com o modelo Cα é preciso extrair as informações estruturais da proteína nativa por meio dos dados depositados no PDB *(protein data bank)* [27, 28]. Com essas informações, conseguimos extrair a posição de cada aminoácido e quais resíduos estão presentes na formação de contatos nativos.

Baseado na estrutura nativa, a etapa que segue, é a utilização do pacote *CSU* [29] para a elaboração do primeiro mapa de contato. Esta ferramenta utiliza dados extraídos do PDB para verificar quais contatos estão sendo formados na proteína nativa, utilizando todos os átomos e não apenas os carbonos alfas, levando em consideração

fatores como: acessibilidade de solvente, interação entre aminoácidos hidrofóbicos e hidrofílicos, a capacidade de certos aminoácidos efetuarem ligações de hidrogênio e etc. Com este mapa finalizado, dá-se o início da dinâmica molecular utilizando o arquivo de topologia gerado pelo Structure-based Models in Gromacs (SMOG) disponível online [30, 31]. A simulação foi elaborada com os recursos computacionais do Shiva Cluster do departamento de Física do Instituto de Biociências, Letras e Ciências exatas da UNESP e realizada com o pacote de dinâmica molecular *GROMACS versão 4.5-5* [32] utilizando integrador estocástico e o acoplamento térmico de Berendsen. As proteínas foram simuladas sobre 10^9 passos tendo as informações guardadas a cada 5000 passos, obtendo no total 200000 frames. A coordenada de reação utilizada para acompanhar o enovelamento foi a fração de contatos nativos Q, sendo que de acordo com a dinâmica, um contato nativo era aceito em uma conformação Γ se a distância entre os aminoácidos nesta conformação era menor do que $1.2d_{ij}$, no qual d_{ij} é a distância entre os resíduos na estrutura nativa.

Os perfis de grandezas termodinâmicas, tais como, Energia térmica, Energia Livre, Entropia e calor específico foram calculados por meio do método dos múltiplos histogramas (WHAM - *Weighted Histograms Analysis Method [33])* detalhado no apêndice A, sendo que a temperatura de enovelamento T_F^0 foi definida como a temperatura do maior valor referente ao calor específico.

A segunda etapa que estruturou este trabalho concerne no cálculo dos contatos não-nativos. Neste caso, foi definido como um contato não-nativo, todo e qualquer contato entre dois aminoácidos separados por no mínimo três resíduos e que estejam distantes a no máximo $6\mathring{A}$. Um fator limitante é que estes contatos não podem estar presentes na primeira lista de contatos gerada pelo *CSU*. Outro fator preponderante para que um contato possa ser chamado de não-nativo é que ele não pode ter uma probabilidade superior a 30% de estar presente nas estruturas com $Q > 0.9$. Sendo assim, caso algum contato apresente uma probabilidade maior do que 30% de aparecer nas configurações com $Q > 0.9$ ele é retirado da lista de contatos não-nativos e é acrescentado na lista de contatos nativos. Com um novo mapa de contato nativo, repete-se os passos referentes a dinâmica citados acima. Não utilizamos o pacote *CSU* para determinar o mapa de contato não-nativo pois estamos interessados em ver aspectos gerais, e o CSU utilizado depois da simulação computacional (feita com o modelo Cα) restringe bastante a lista de contatos, o que não é de nosso interesse.

Com o mapa de contato nativo e não-nativo finalizados, é possível calcular a ordem de contato parcial e torná-la uma coordenada de reação que possa servir para acompanhar o processo de enovelamento. O mesmo procedimento feito com o WHAM utilizando Q foi feito utilizando a ordem de contato parcial, CO, como coordenada.

4.4 Cálculo da variação de contatos não-nativos

A diferença na quantidade de contatos não-nativos entre o estado de transição para com o estado desenovelado (ΔA) foi realizada tomando a média da quantidade de contatos não-nativos ($\bar{A}$) formados dentro de cada região.

$$\Delta A = \bar{A}(TS) - \bar{A}(Unf) \tag{4.2}$$

A região de transição (TS) e a região desenovelada (Unf) foram delimitadas utilizando o perfil de energia livre para cada proteína. A região desenovelada envolve todos os estados presentes desde a configuração inicial (CO = 0 ou Q=0) até a configuração no qual a proteína atinge o valor de 15% de ΔF após passar pelo primeiro mínimo de energia livre. Já a região de transição, envolve todas as configurações localizadas dentro da barreira de energia com valores de energia livre superiores a 85% de ΔF. Sendo que ΔF foi calculado utilizando como base o pico da barreira de energia livre ($F^{\neq}$) e o primeiro mínimo (F_{unf}), como exemplificado na figura 4.1. Os valores de 15% e 85% que delimitam as regiões desenoveladas e de transição foram escolhidos pois foram os valores que conseguiram separar melhor as duas regiões em todas as proteínas, mesmo aquelas que apresentavam baixíssima barreira de energia livre.

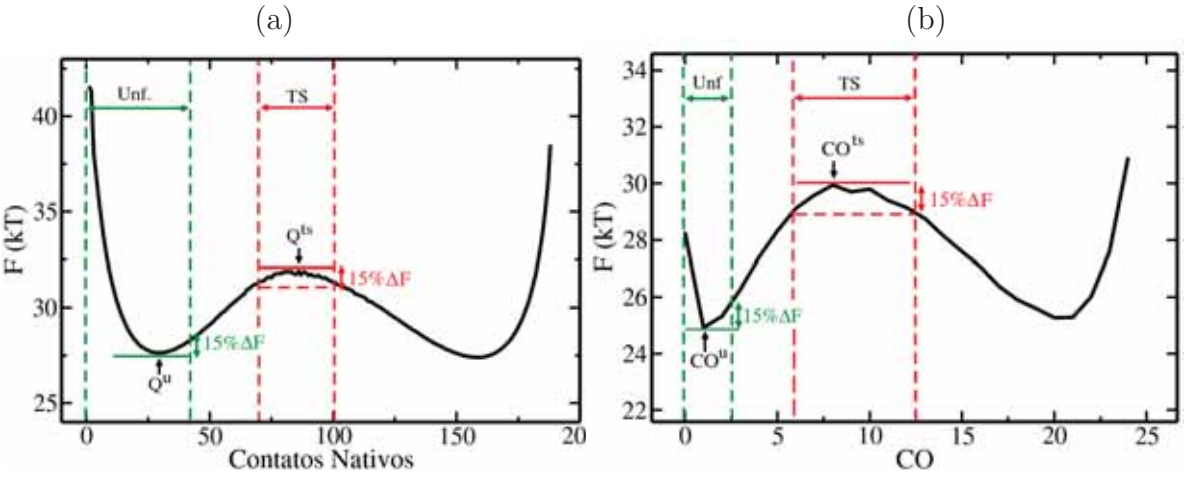

Figura 4.1: **Localização da região desenovelada e de transição** (a) Perfil de energia livre (F) em função do número de contatos nativos para a proteína Ubiquitin. (b) Perfil de energia livre (F) em função da ordem de contato (CO) para a proteína Ubiquitin. Q^u e CO^u remetem respectivamente o número de contatos nativos e a ordem de contato do primeiro mínimo de energia livre. Q^{ts} e CO^{ts} remetem ao número de contatos nativos e a ordem de contato do pico da barreira de energia livre. Em (a) $\Delta F = F(Q^{ts}) - F(Q^u)$, em (b) $\Delta F = F(CO^{ts}) - F(CO^u)$. A região entre as duas linhas verticais tracejadas na cor verde é a região chamada de desenovelada (Unf) e a região delimitada pelas linhas verticais tracejadas na cor vermelha é a região de transição (TS).

Capítulo 5

Resultados

5.1 Ordem de Contato Parcial como coordenada de reação

A ordem de contato calculada em função da fração de contatos nativos se mostrou bem comportada e capaz de descrever o processo de enovelamento, apresentando para todas as proteínas estudadas neste trabalho um crescimento praticamente linear com a fração de contatos nativos, como mostrado na figura 5.1.

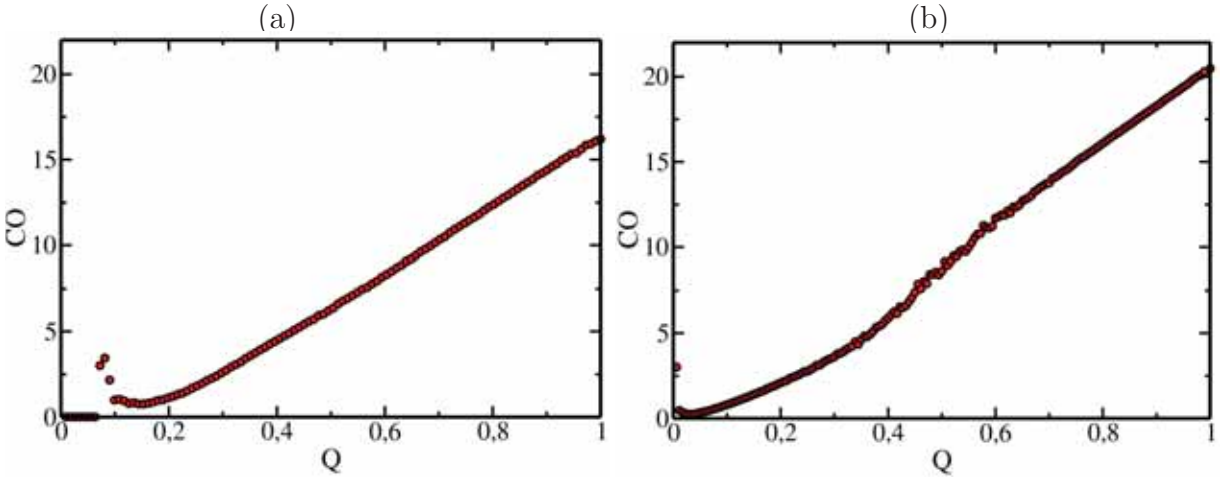

Figura 5.1: **Ordem de contato (CO) em função da fração de contatos nativos (Q)** (a) Dados obtidos para a proteína EnHD. (b) Dados obtidos para a proteína CSPTm

Para calcular as grandezas termodinâmicas utilizando o WHAM, tivemos que transformar os valores reais de CO em valores inteiros, permitindo a existência de estados degenerados, ou seja, uma mesma ordem de contato poderia estar associada a dois ou mais estados conformacionais. Esta degenerescência implica na diminuição

de pontos que formam os perfis de energia livre, entropia e energia térmica. Outro fator que caracteriza CO quando utilizada como coordenada de reação é que esta tem a capacidade de nos dizer que tipo de contato está sendo feito em um determinado estado configuracional.

Para todas as proteínas estudadas houve um deslocamento para a esquerda da posição do pico da barreira de energia livre quando calculada utilizando CO em relação a posição da barreira quando calculada utilizando Q (vide Tabela I). Este deslocamento nos mostra que, utilizando Q como coordenada de reação, sabemos quantos contatos são feitos em um determinado estado mas não sabemos que tipo de contato são esses, mas quando utilizamos a ordem de contato como coordenada e o pico se desloca para valores pequenos de CO, isso nos mostra que para a proteína se enovelar há a necessidade de formar inicialmente contatos de curto alcance. A presença de contatos de curto alcance nas configurações localizadas no estado de pré-transição não foi somente evidente nas proteínas formadas exclusivamente por α-hélice, mas também é bastante visível nas proteínas que apresentam grande quantidade de contatos de longo alcance na estrutura nativa, como nas proteínas formadas exclusivamente por folhas-β, além disso, os contatos com maiores probabilidades no estado de pré-transição estão localizados principalmente em regiões de "loopings"curtos da estrutura nativa. Já os contatos de longo alcance são formados com grande probabilidade na região de transição, ajudando a proteína vencer a barreira energética, como exemplificado na figura 5.2

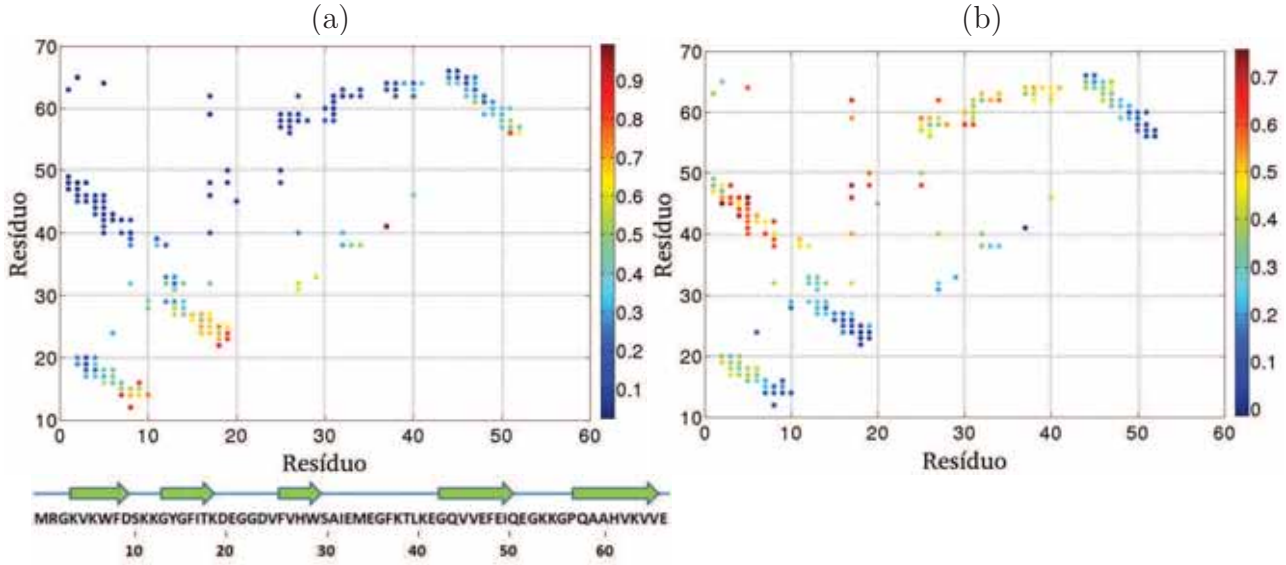

Figura 5.2: **Probabilidade dos contatos nativos da proteína CSPTm estarem formados** (a) Na região de pré-trasição. Embaixo; Arranjo das estruturas secundárias de acordo com a sequência de aminoácidos.As flechas preenchidas com a cor verde representam a formação de folhas-β. (b) Diferença na probabilidade entre a região de pós transição e P-TS.

5.2 Variação de contatos não-nativos

De acordo com o modelo teórico desenvolvido por Clementi e Plotkin [8] o sinal positivo ou negativo da variação na quantidade de contatos não-nativos é o que vai definir se a frustração favorece ou não o enovelamento, sendo assim, temos na Figura 5.3 uma representação típica dos dois casos.

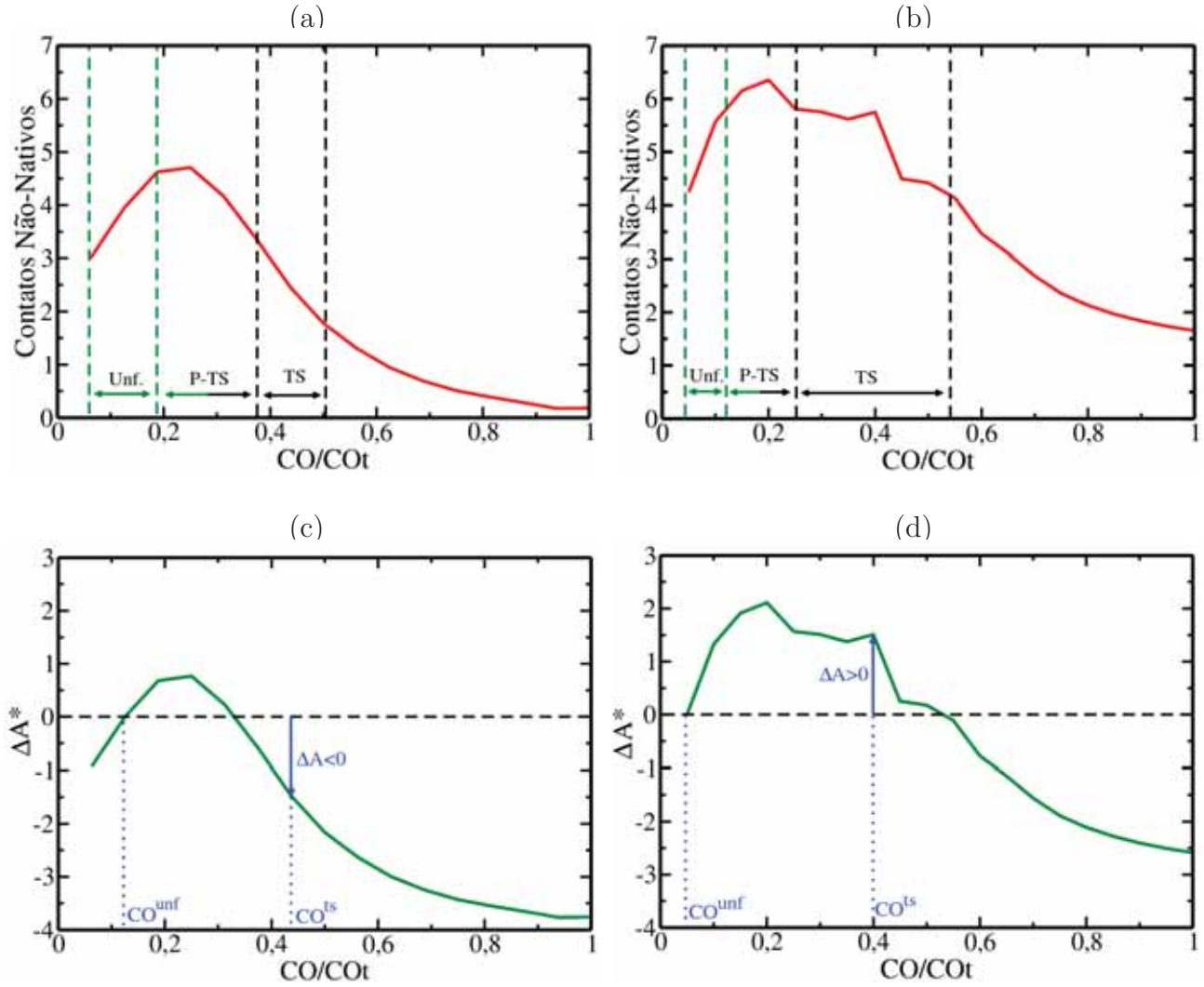

Figura 5.3: **Contatos não nativos em função da ordem de contato normalizada**. (a) Número médio de contatos não-nativos para a proteína EnHD. (b) Número médio de contatos não-nativos para a proteína CSPTm. (c) Variação de contatos não-nativos para a proteína EnHD (d) Variação de contatos não-nativos para a proteína CSPTm. As linhas tracejadas nas figuras (a) e (b) indicam onde foram delimitadas as regiões: desenovelada (Unf.), de transição (TS) e de pré transição (P-TS). $\Delta A^* = \bar{A}(CO^{ts}) - \bar{A}(CO^{unf})$. CO^{ts} é um estado conformacional localizado dentro da região de transição e CO^{unf} é o estado onde está localizado o primeiro mínimo de energia livre.

De acordo com a Figura 5.3, podemos notar que para as proteínas nas quais a frustração dificilmente contribui de forma favorável ao enovelamento (representados

pela proteína EnHD), em nenhum, ou em poucos estados acessado pela proteína durante todo o processo, o valor médio de contatos não-nativos supera o valor referente ao ponto localizado dentro do estado desenovelado. Por outro lado, vemos que para as proteínas as quais a frustração tem um grande potencial de ajudar o enovelamento, a quantidade de contatos não-nativos cresce consideravelmente após o ponto escolhido para representar o estado desenovelado (CO^{unf}). E para todos os casos, o maior valor de contatos não-nativos aparece na região de pré-transição, assim como descrito no modelo teórico.

Todos os resultados obtidos no cálculo da variação de contatos não-nativos podem ser vistos na Tabela I.

Tabela I: Dados obtidos para as 19 proteínas estudadas.

Nome	M	COt	$Q_M^{\neq} \div CO_{COt}^{\neq}$ [a]	$\Delta F(Q^{\neq})$	$\Delta F(CO^{\neq})$	$\Delta A(Q)^b$	$\Delta A(CO)^b$
ACR*	58	6	–	0.00	0.00	–	–
HP36*	55	8	–	0.00	0.00	–	–
PSBD*	64	11	–	0.00	0.00	–	–
$\alpha_3 D$	136	15	1.38	0.21	0.52	-5.44	-11.6
PtABD	102	15	1.52	0.37	0.75	-0.98	-5.80
EnHD	111	16	1.11	0.75	0.74	-9.81	-13.3
IM9	178	23	1.37	1.80	2.64	-4.44	-5.88
ACBD	182	26	1.08	1.32	2.03	-0.52	-6.80
HHCC	246	25	1.42	1.59	2.45	-7.10	-8.70
PtL	136	19	1.19	2.34	3.21	2.05	5.60
PtG	139	21	1.24	2.94	3.78	4.17	7.50
ADA2h	175	29	1.05	2.52	3.09	3.60	4.00
CI2	152	21	1.21	3.17	4.01	-1.25	-2.60
SH3	152	22	1.02	4.03	4.38	3.94	3.40
Ubiquitin	188	24	1.39	4.27	5.01	1.27	1.46
CSPTm	180	20	1.20	5.95	6.66	5.00	3.78
HPr	222	31	1.23	5.69	6.77	1.30	1.85
αAIT	196	27	1.30	6.65	7.00	11.68	9.56
TWIg	253	33	1.79	5.98	6.93	10.19	11.32

[a] Relação entre a posição do pico de energia livre quando calculado em função de Q e quando calculado em função de CO.

[b] Variação na fração de contatos não-nativos $\times 10^{-3}$.

* Proteínas que não apresentaram barreira de energia livre

Dentre as 19 proteínas testadas, três não apresentaram barreira de energia livre, seja utilizando Q ou CO como coordenada de reação. Neste caso, se a proteína não apresenta barreira de energia livre, não há porque tentar diminuí-la ainda mais, portanto, para estas proteínas o acréscimo de frustração só atrapalharia o enovelamento. Somente quatro proteínas conflitam a teoria com o que mostra os dados obtidos computacionalmente por Contessoto ET al, no qual as proteínas: IM9, ACBD, HHCC e CI2 possuem uma frustração ótima diferente de zero, e era de se esperar que o ΔA destas proteínas fossem positivos. Porém, exceto para a CI2, a frustração ótima para as outras três proteínas é o menor valor que fora testado por Contessoto et al [7].

Talvez, um maior refino na forma de calcular a frustração ótima para cada proteína testada por Contessoto e colaboradores eliminasse os resultados que entraram em discordância com o que foi proposto no modelo analítico, ou seja, a frustração ótima para as proteínas HHCC, ACBD, IM9 e CI2 poderiam ter o valor de ϵ_f^{opt} alterado para $\epsilon_f^{opt} = 0$. Por outro lado, definir que a frustração energética favorece o enovelamento olhando somente para o sinal da variação na quantidade de contatos não-nativos soa como um fator demasiadamente abrupto para separar os dois casos. Na próxima seção, será introduzida a idéia de uma região intermediária entre os dois extremos, separadas principalmente pela altura da barreira de energia e pela estrutura nativa adquirida.

5.3 Correlação entre frustração, ordem de contato e fração de contatos não-nativos

De acordo com o modelo analítico desenvolvido por Clementi e Plotkin, o favorecimento no processo de dobragem está diretamente correlacionado com a quantidade de contatos não-nativos que são formados dentro do estado de transição quando comparados com a quantidade de contatos não-nativos que aparecem no estado desenovelado, e de acordo com esta quantidade, a frustração energética pode aumentar ou diminuir a barreira de energia livre. A figura 5.4 nos traz a exata correlação entre a altura da barreira de energia livre e a variação na quantidade de contatos não-nativos de acordo com os dados das 19 proteínas estudadas.

A correlação linear de 0.81 entre a variação na quantidade de contatos não-nativos e a altura da barreira de energia livre, nos mostra que há uma forte correlação entre estes dois parâmetros, o que de certa forma corrobora o modelo analítico. Podemos verificar na figura 5.4 que as proteínas foram suavemente separadas em três grupos, separados da ordem de 1kT de diferença na altura da barreira;

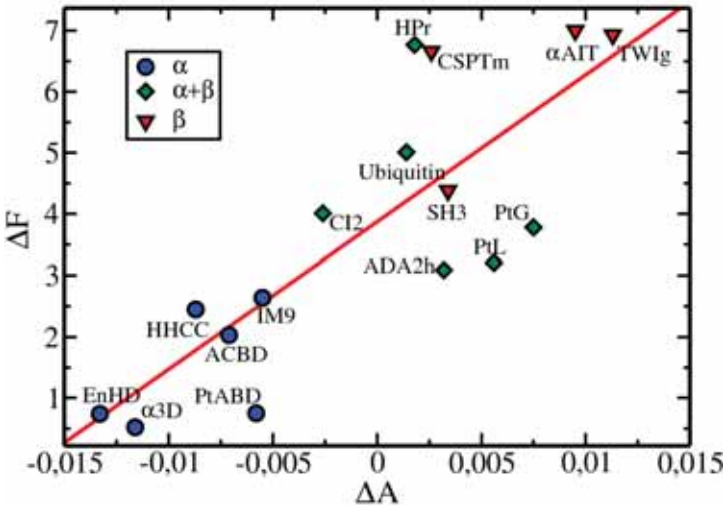

Figura 5.4: **Correlação entre altura da barreira e variação na quantidade de contatos não-nativos** Dados extraídos da Tabela 1 utilizando a ordem de contato como coordenada de reação. As proteínas estão coloridas de acordo com a estrutura nativa depositada no PDB. ΔF é a altura da barreira de energia livre, ΔA é a variação na fração de contatos não-nativos. A Correlação linear entre ΔF e (ΔA) é de 0.81

Grupo A = ACR, HP36,PSBD, $\alpha 3D$, PtABD e EnHD;

Grupo B = IM9, ACBD, HHCC, PtL, PtG, ADA2h, CI2, SH3 e Ubiquitin;

Grupo C = CSPTm, HPr, αAIT, TWIg;

De posse da figura 5.4, podemos pensar na formação dos três grupos como um ponto de convergência entre os resultados propostos pelo modelo analítico e os resultados obtidos na determinação da frustração ótima calculado por Contessoto e colaboradores, pois a formação destes três grupos nos remete a um possível grupo intermediário (Grupo B) que estaria entre o grupo no qual as proteínas são incontestavelmente desfavorecidas (Grupo A) pelo acréscimo de frustração energética e o grupo nos quais as proteínas são favorecidas (Grupo C), sendo que aquelas quatro proteínas (IM9, HHCC, ACBD e CI2) estariam localizadas neste grupo intermediário.

Separando as proteínas nestes três grupos, podemos notar que o grupo cuja frustração desfavorece o enovelamento seria povoado principalmente por proteínas com estrutura em α-hélice. Já na região intermediária estariam principalmente as proteínas formadas pela associação de α-hélice e folhas-β, e na região na qual as proteínas seriam favorecidas pelo acréscimo de frustração estariam aquelas formadas principalmente por folhas-β. Dada a relação entre os grupos com a estrutura nativa, é de se esperar uma boa correlação entre ΔA e ordem de contato, o que pode ser visto na figura 5.5.

As boas correlações entre ΔA com ΔF e RCO torna a associação entre a ordem

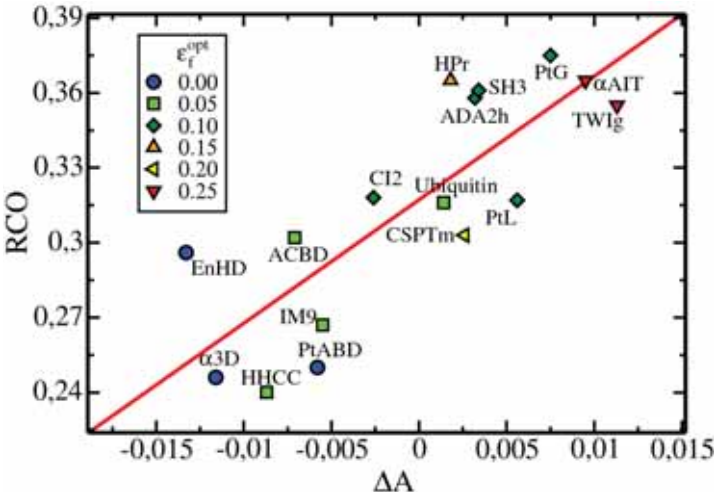

Figura 5.5: **Correlação entre ordem de contato relativa (RCO) e variação na fração de contatos não-nativos (ΔA)** Dados extraídos utilizando a ordem de contato como coordenada de reação. As proteínas estão coloridas de acordo com a frustração ótima determinada por Contessoto ET al [7]. A Correlação linear entre RCO e (ΔA) é de 0.81

de contato e a altura da barreira de energia livre bons candidatos para prever o efeito da frustração energética nas proteínas, assim como foi proposto por Contessoto e colaboradores. A correlação linear de 0.84 entre $RCO \times \Delta F$ com ΔA ilustrado na figura 5.6 aproxima mais uma vez o modelo analítico dos resultados computacionais, confirmando a boa correlação entre a quantidade de contatos não-nativos com a presença de uma frustração ótima.

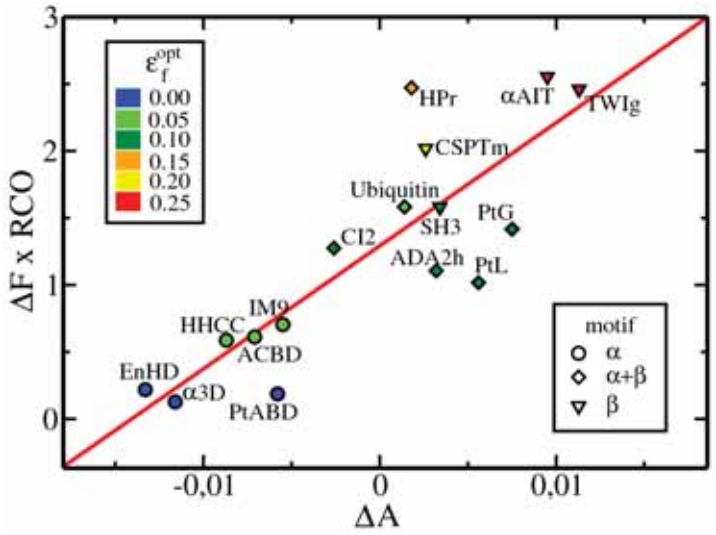

Figura 5.6: **Correlação entre RCO$\times\Delta F$ e ΔA.** Dados extraídos utilizando a ordem de contato como coordenada de reação. As proteínas estão coloridas de acordo com a frustração ótima determinada por Contessoto ET al [7] e estão representadas de acordo com a estrutura nativa no PDB. A Correlação linear entre RCO$\times\Delta F$ e (ΔA) é de 0.84

Capítulo 6

Conclusões e Perspectivas Futuras

A ordem de contato utilizada como coordenada para acompanhar o processo de enovelamento se mostrou capaz de substituir a já bem estabelecida fração de contatos nativos no modelo analítico, visto que todos os resultados obtidos utilizando a fração de contatos nativos (Q) concordaram qualitativamente com os resultados obtidos utilizando a ordem de contato parcial (CO), e uma vez que estudos correlacionam frustração com ordem de contato, se mostra oportuno utilizar a ordem de contato como coordenada de reação em um modelo analítico que estuda o efeito da frustração no enovelamento. Além disso, a utilização da ordem de contato para descrever o processo de enovelamento nos traz informações diferentes das que são obtidas quando realizadas utilizando a fração de contatos nativos, aumentando nosso nível de conhecimento sobre o assunto. Tais informações explicitaram a necessidade das 19 proteínas estudadas em iniciar o processo de enovelamento formando contatos nativos entre aminoácidos próximos, sendo que os contatos entre aminoácidos distantes são feitos durante e pós a região de transição, auxiliando a proteína ultrapassar a barreira de energia livre.

De acordo com os resultados, quanto maior a presença de contatos de longo alcance na estrutura nativa maior é a quantidade de contatos não-nativos no estado de transição e de pré-transição. Sendo assim, a inserção de uma energia de interação entre os aminoácidos que efetuam um contato não-nativo está diretamente relacionado com a capacidade da frustração em diminuir o tempo de enovelamento, visto que estas proteínas com grande presença de contatos não-nativos são também as que apresentam um valor maior de frustração ótima.

A boa correlação entre altura da barreira sem frustração, variação na fração de contatos não-nativos e ordem de contato, aproxima o modelo teórico desenvolvido por Clementi e Plotkin dos resultados obtidos para proteínas reais por meio de simulação

computacional utilizando um potencial com frustração [7]. Além disso, temos mais um indício de que estes parâmetros podem servir para prever o efeito do acréscimo da frustração no enovelamento, sem precisar inserir termos específicos de frustração na simulação, assim como foi feito neste trabalho. Desta forma, de acordo com os resultados obtidos, vemos que o acréscimo de frustração dificilmente favorecerá as proteínas que apresentam baixa barreira de energia livre e baixa ordem de contato, formados estritamente por α-hélice. Já para proteínas com alta barreira de energia e alta ordem de contato, representadas tipicamente por proteínas com grande presença de folhas-β, o acréscimo de frustração energética provavelmente favorecerá o processo de enovelamento.

Em trabalhos futuros deverá ser calculado a altura da barreira de energia livre acrescentando um pouco de frustração nestas mesmas proteínas, mantendo o mesmo mapa de contato, para que seja possível então verificar a diferença entre a altura da barreira de energia livre sem frustração e com frustração energética, verificando se esta diferença correlaciona bem com os outros termos presentes na Eq.3.19. Desta forma, além de dizer em quais proteínas o acréscimo de frutração favorece o enovelamento, será possível prever qual o valor da frustração energética, e determiná-la tanto computacionalmente quanto teoricamente.

Referências Bibliográficas

[1] VOET, D.; VOET, J. G.; PRATT, C. W. **Fundamentos de bioquímica: a vida em nível molecular**,2. Ed. Porto Alegre: Artmed, 2008.

[2] NELSON, D. L; COX, M. M. **Lehninger: Principles of Biochemistry**, 4. Ed.

[3] PACI, E.; VENDRUSCOLO, M.; KARPLUS, M. Validity of G$\bar{o}$ Models: Comparison with a Solvent-Shielded Empirical Energy Decomposition. **Biophysical Journal**,v. 83, p. 3032-3038, Dec. 2002.

[4] WANG, J.; OLIVEIRA, R. J.; CHU, X.; WHITFORD, P. C et al. Topography of funneled landscape determines the thermodynamics and kinects of protein folding. **Proc. Natl. Acad. Sci. USA**, v 109, no. 39, p. 15763-15768, Set. 2012.

[5] FERSHT, A. R. **Structure and mechanism in protein science: a guide to enzyme catalysis and protein folding**, New York: W.H. Freeman, 1999.

[6] SHEA, J-E.; ONUCHIC, J. N.; BROOKS, C. L. Exploring the origins of topological frustration: Design of a minimally frustrated model of fragment B of protein A. **Proc. Natl. Acad. Sci. USA**, v. 96, p. 12512–12517, Out. 1999.

[7] CONTESSOTO, V. G.; LIMA, D. T.; OLIVEIRA, R. J.; BRUNI, A. T.; CHAHINE, J.; LEITE, V. B. Analyzing the effect of homogeneous frustration in protein folding.**Proteins**, v. 81, p. 1727-1737, Out. 2013.

[8] CLEMENTI, C.; PLOTKIN, S. S.; The effects of nonnative interactions on protein folding rates: Theory and simulation. **Protein Science**, v. 13, p 1750-1766, Jul. 2004.

[9] DOBSON, C. M.; Protein folding and misfolding. **Nature** 423, p 884-890, Dez. 2003.

[10] CHEN, C. **Investigating Nonnative Contacts in Protein Folding.** Thesis (Master of Science) - The Graduated Faculty of the University of Akron. 2009.

[11] ANFINSEN, C. B.; HABER, E.; SELA, M.; WHITE, F. H. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. **Proc. Natl. Acad. Sci. USA**, v. 47, p. 1309-1314, Set. 1961.

[12] ZWANZIG, R.; BAGCHI, B. Levinthal's paradox. **Proc. Natl. Acad. Sci. USA**. v. 89, p. 20-22, Jan. 1992.

[13] LEOPOLD, P. E.; MONTAL, M.; ONUCHIC, J. N. Protein folding funnels: A Kinetic approach to the sequence-structure relationship. **Proc. Natl. Acad. Sci. USA**, v 89, p. 8721-8725, Set. 1992.

[14] BRYNGELSON, J. D.; ONUCHIC, J. N.; SOCCI, N. D.; WOLYNES, P. G. Funnels, pathways and the energy landscape of protein-folding: A synthesis. **Proteins**, v. 21, p. 167-195, Mar. 1995.

[15] WOLYNES, P.; ONUCHIC, J. N.; THIRUMALAI, D. Navigating the folding routes. **Science**, v. 267, p. 1619-1620, Mar. 1995.

[16] DILL, K. A.; CHAN, H. S. From Levinthal to pathways to funnels. **Nature Structural Biology.**, v. 4, p. 10-19, 1997.

[17] BRYNGELSON, J. D.; WOLYNES, P. G. Spin-glasses and the statistical mechanics of protein folding. **Proc. Natl. Acad. Sci. USA**, v. 84, p. 7524-7528, Nov. 1987.

[18] ONUCHIC, J. N.; NYMEYER, A. E.; GARCIA, A. E.; CHAHINE, J.; SOCCI, N. D. The energy landscape theory of protein folding: Insights into folding mechanisms and scenarios. **Adv. Protein Chem.**, v. 53, p. 87-152, 2000.

[19] OAKLEY, M. T.; WALES, D. J; JOHNSTON, R. L. The effect of nonnative interactions on the energy landscape of frustrated model proteins. **Journal of Atomic and Molecular Physics**, v. 2012.

[20] LUBCHENKO, V. Competing interactions create functionality through frustration. **Proc. Natl. Acad. Sci. USA**, v. 105, p. 10635-10636, Ago. 2008.

[21] PLAXCO, K. W.; SIMONS, K. T.; BAKER, D. Contact Order, tansition state placement and the refolding rates of single domain proteins. **J. Mol. Biol.**, v. 277, p. 985-994, 1998.

[22] CLEMENTI, C.; GARCIA, A. E.; ONUCHIC, J. N. Interplay among Tertiary contacts, secundary structure formation and side-chain packing in the protein folding mechanism: All-atom representation study of protein L. **J. Mol. Biol.** v. 326, p. 933-953, 2003.

[23] CONTESSOTO, V. G. **Estudo do efeito da adição de frustração no enovelamento de proetínas utilizando modelos baseados em estrutura**. Tese (Mestrado em Biofísica Molecular)- UNESP, São José do Rio Preto, 2012.

[24] LIMA, D. T.; **Regimes de frustação ótima em enovelamento de proteínas com modelos baseado em estrutura**. Tese (Mestrado em Biofísica Molecular) - UNESP, São José do Rio Preto, 2012.

[25] CLEMENTI, C.; NYMEYER, H.; ONUCHIC, J. N. 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**, v. 298, p. 937-953, 2000.

[26] HOANG,T. X.; CIEPLAK, M. Sequencing of folding events in Go-like proteins. **J. Chem. Phys.**, v. 113, Ago. 2000.

[27] BERMAN, H. M.; WESTBROOK, J.; FENG, Z.; GILLILAND, G.; BATH, T. N.; WEISSIG, H.; SHINDYALOV, I. N.; BOURNE, P. E. The Protein Data Bank. **Nucleic Acid Research**, v. 28, p. 235-242, Jan. 2000.

[28] RCBS PDB. Disponível em: http://www.rcbs.org/pdb/home/home.do . Acessado em 1/02/2013.

[29] SOBOLEV, V.; SOROKINE, A.; PRILUSKY, J.; ABOLA, E.E.; EDELMAN, A. Automated analysis of interatomic contacts in proteins. **Bioinformatics**, v. 15, p. 327 -332, Abr. 1999.

[30] NOEL, J. K.; WHITFORD, P. C.; SANBONMATSU, K. Y.; ONUCHIC, J. N. SMOG@ctbp: simplified deployment of structure-based models in GROMACS. **Nucleic Acids Research**, v. 38, p. W657-W661, JUN. 2010.

[31] SMOG@ctbp: Structure-based MOdels in Gromacs. Disponível em: http://http://smog-server.org. Acessado em 01/02/2013.

[32] VAN DER SPEL, D.; LINDAHL, E.; HESS, B.; GROENHOF, G.; MARK, A. E.; BERENDSEN, H. J. C. GROMACS: fast, flexible, and free. **J. Comp. Chem.**, v. 26, p. 1701-1718, 2005.

[33] FERRENBERG, A. M., SWENDSEN, R. H. New monte carlo technique for studying phase transitions. **Phys Rev Lett**, v. 61, p. 2635-2638, 1988.

[34] KUMAR, S.; ROSENBERG, J. M.; BOUZIDA, D.; SWENDSEN, R. H.; KOLLMAN, P. A. The weighted histogram analysis method for free-energy calculations

on biomolecules. I. The method. **Journal of Computational Chemistry**, v. 13, p. 1011-1021, Out. 1992.

Apêndice A

WHAM – *Weighted Histograms Analysis Method*

O método dos múltiplos histogramas tem como objetivo aumentar a informação sobre o sistema. Utilizando valores oriundos da simulação em uma gama de temperaturas, fazemos com que os dados sejam melhores amostrados para que se possa ter uma boa estatistica sobre as conformações adquiridas pelo sistema. Combinando as saídas das simulações para uma única proteína com o WHAM, somos capazes de gerar grandezas termodinâmicas, principalmente, no que se diz respeito aos perfis do calor específico de acordo com a temperatura e a energia livre do sistema com relação a coordenada de reação, por exemplo Q [25, 23, 34].

O método é baseado no fato de que o logaritmo da distribuição de probabilidade P(Q) dos valores correlacionados com um certo Q em uma temperatura fixa pode servir como uma estimativa do perfil da energia livre para aquela temperatura [25].

Utilizando dados da mecânica estatística, para um conjunto de N simulações com C_n configurações para a temperatura $T_n = \frac{1}{k_B \beta_n}$, temos que a densidade de estado é dada por:

$$\Omega_n(Q) = H_n(Q) \exp\left(\beta_n E - f_n\right) \tag{A.1}$$

no qual $H_n(Q)$ é o histograma da coordenada Q e f_n é a energia livre adimensional. Os valores para f_n são dados arbitrariamente, e ao longo do método são realizados interações que testam os valores até que se atinja a convergência esperada.

A probabilidade de o sistema adquirir a conformação Q numa dada temperatura T, é calculada da seguinte forma:

$$P(Q) = \frac{\sum_{n=1}^{N} H_n(Q) \exp(-\beta_n E)}{\sum_{n=1}^{N} C_n(Q) \exp(-\beta_n E + f_n)} \tag{A.2}$$

De posse da probabilidade calculada de acordo com a equação A.2, podemos calcular a energia livre do sistema e as demais grandezas termodinâmicas:

$$\exp(-f) = \sum P(Q) \tag{A.3}$$

Apêndice B

Correlação entre frustração ótima e ordem de contato absoluta

Contessoto et all [7, 24, 23], por meio de dinâmica molecular acrescentando um novo termo associado à frustração ao potencial do modelo C_α, conseguiram separar em dois grupos as mesmas 19 proteínas utilizadas neste trabalho. A distinção foi feita de acordo com ordem de contato absoluta e altura da barreira de energia livre. O potencial acrescido ao modelo C_α está representado na equação B.1

$$V_f(r) = -\epsilon_f \exp\left[-\frac{(r_{ij} - \bar{d})^2}{{\sigma_f}^2}\right] \tag{B.1}$$

$\bar{d}$ é a média das distâncias dos contatos nativos, $\sigma_f = 1\ \mathring{A}$ e ϵ_f é o parâmetro de frustração em unidades de ϵ_c.

Em um dos grupos, foi detectado que a presença de frustração sempre atrapalha o enovelamento. Já no outro grupo de proteínas, foi determinado uma frustração ótima diferente de zero, que favorecia o processo de enovelamento. Os dois grupos podem ser visualizados na Figura B.1. A correlação encontrada pelo produto entre a ordem de contato absoluta com a altura da barreira de energia de acordo com a frustração ótima, pode ser vista na Figura B.2.

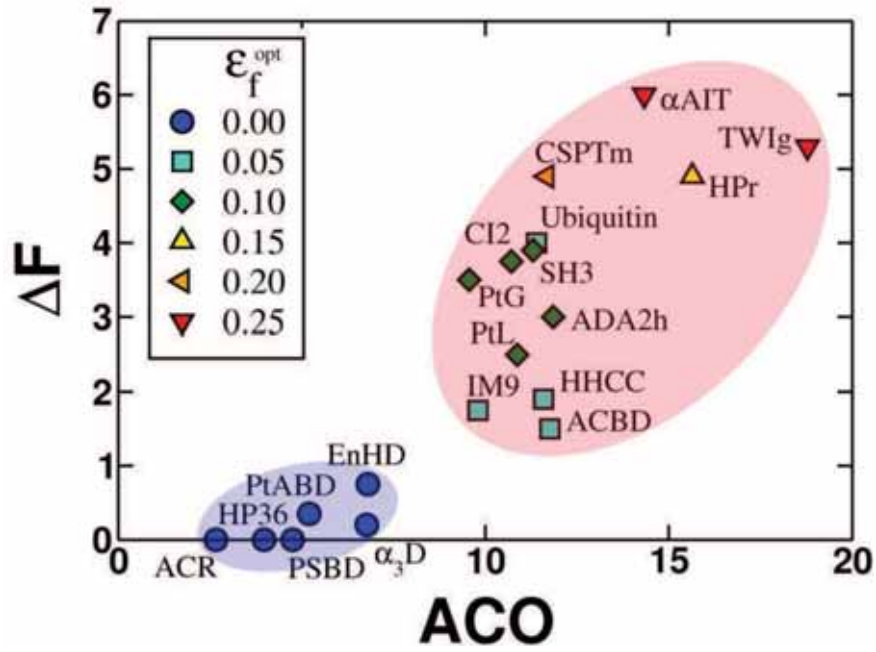

Figura B.1: **Representação da relação entre barreira de energia livre ΔF e a ordem de contato absoluta ACO para todas as proteínas.** O valor da frustração ótima de cada proteína é apresentado pelas diferentes cores e formatos geométricos. Figura extraída de [23].

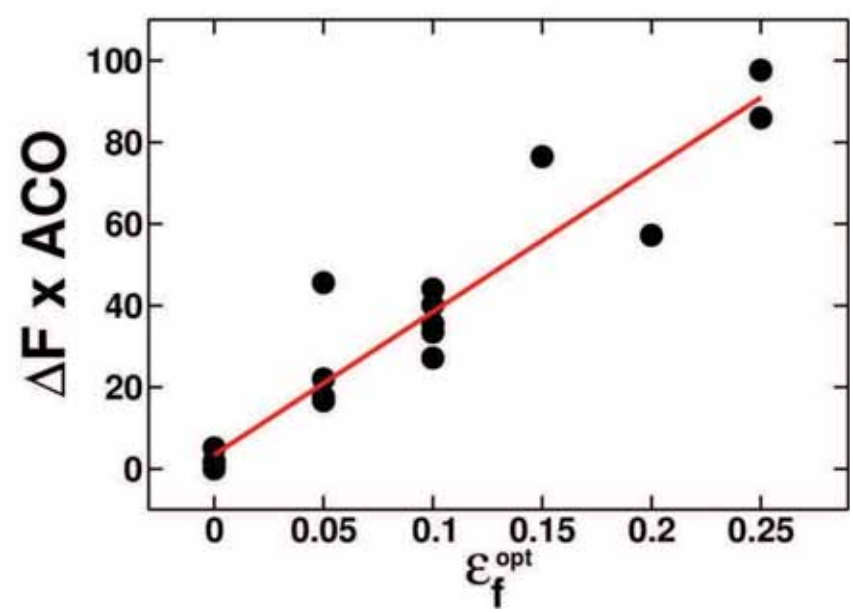

Figura B.2: **Representação do produto ΔF x ACO em função da frutração ótima ϵ_f^{opt}.** Foi encontrado um valor de correlação positiva de 0.95. Figura extraída de [23].

Apêndice C

Manuscrito em produção

Manuscrito em elaboração “Quantifying contact order and free energy barrier correlation with nonnative interactions for protein folding speed limit”. Esta versão é a primeira que foi elaborada e está em fase de criação. O manuscrito discute, além da teoria e metodologia utilizada, os resultados encontrados durante o mestrado.

Quantifying contact order and free energy barrier correlation with nonnative interactions for protein folding speed limit

Mouro P.R.[a], Contessoto V.G.[a], Chahine J.[a],
Oliveira R.J.[b], Leite V.B.P.[a*]

[a] Departamento de Física
Instituto de Biociências, Letras e Ciências Exatas
Universidade Estadual Paulista
São José do Rio Preto, SP, Brazil, 15054-000
[b] Departamento de Física
Instituto de Ciências Exatas, Naturais e Educação
Universidade Federal do Triângulo Mineiro
Uberaba, MG, Brazil, 38064-200

[*] Corresponding author
E-mail:vleite@sjrp.unesp.br
Telephone: +55 (17) 3221-2240
Fax: +55 (17) 3221-2247

Short Title: Contact order and energetic frustration for protein folding
Key words: structure based model, energy landscape
molecular dynamics, C-alpha model

April 16, 2014

Abstract

Protein folding is essential for the maintenance of living systems and the physico-chemical principles, which governs this process, became a central problem for the human life longevity. In this context, the energy landscape theory has been supporting theoretical and experimental advances in the understanding of protein folding mechanisms. The energy landscape of globular proteins resembles a funnel of structures progressively folded en route to the native state, minimally frustrated state. It is well stablished that an addition of small amount of energetic frustration enhances folding speed for certain proteins. We applied the C_α structure-based model to simulate a group of proteins with the contact order (CO) as the reaction coordinate and we found that CO and free energy barrier at the transition state (ΔF) correlates with nonnative contacts variation (ΔA) at the optimum frustration regime. We also found that ΔF and ΔA cluster the simulated proteins by their fold motifs. These computational findings are corroborated by analytical model. As a consequence, optimum frustration regime for protein folding can be predicted analytically.

Introduction

The underlying folding mechanism of a polypeptide chain to its compact three-dimensional structure is one of the grand challenges of modern science. Failure in the process to achieve the correct folded native state can cause a series of pathological conditions like neurodegenerative disorders, including Alzheimer's and Parkinson's diseases [1]. Understanding the thermodynamics and kinetics of protein folding can shed some light on the prevention of these neurodegenerative disorders. Over the last decades, the energy landscape theory has been a consistent framework in revelling protein folding mechanisms [1–5]. This theory states that the energy landscape of globular proteins resembles a funnel of struc-

tures progressively folded en-route to the native state with its bottleneck narrowed at the transition state in between unfolded and folded ensembles [6–9]. The energy landscape theory is successful in explaining qualitatively and quantitatively folding studies in theoretical [10–13] as well as experimental [14–17] investigations. Computational models has been developed based in the energy surface idea in order to predict folding mechanisms, rates and stability parameters, which correlates with wet-lab experiments [9, 18–22].

The funnel theory delineates protein folding as an ensemble of conformations gradually diffusing from the unfolded state, high entropic and energetic conformations at the top of the surface funnel, to the native state, lowest entropic and energetic state at the bottom [23–27]. Proteins are naturally designed so that the folding pathways down the funnel are not dominated by bumps due to local energetic traps, roughness in the energy surface [28–33]. Well-designed sequences have the ability to fold completely in biological time scale and the funnel slope must be steep enough to overcome these bumps and minimize local energetic trapping, or energetic frustration [21]. Frustration occurs due to the impossibility of satisfying all favorable energetic interactions simultaneously during folding events [34–37]. Kinetically foldable proteins tend to maximize propitious interactions, thus biological proteins where naturally selected throughout the evolution process so that the native state is minimally frustrated [34–36]. Apart from energetic frustration, protein topology, given by the chain connectivity (backbone), shapes folding reaction as the protein surpasses the energetic barrier located at transition state and becomes more natively-like. Protein native topology and minimally-frustrated landscape are the key ingredients for the structure-based (Gō) model employed here to carry molecular dynamics simulations [38–40]. The native structure-base model is absent of nonnative energetic frustration and topology effects and yet, the model captures important characteristics of protein folding mechanisms when applyed with computer simulations [39, 41].

In the previous work [42], it was quantified the optimum energetic frustration that can be added to the structure-based model in order to enhance folding rates. It was found that the optimum frustration regime correlates with the free energy barrier and the Baker contact order [43]. The energy landscape theory for the folding speed limit upon the effect of nonnative energetic frustration is well establish [44, 45]. In this manuscript, we connect the missing link between theory [45] and simulation [44]. Using contact order as the reaction coordinate, we found that at the optimum frustration regime, the free energy barrier correlates with the nonnative contacts variation, a parameter which can be obtained by the analytical model. Our main goal was to determine, for a given protein, the optimum nonnative energy for the folding speed limit by an analytical model and with low computation effort.

Methods

To test the correlation between the difference in the amount of non-native contacts in the transition state toward the unfolded state, with the presence of a optimal frustration and the contact order, we used as an object of study a group of 19 proteins [42] . Proteins are coarse-grained in C_α atom level of simplification [12,39]. The details of how the simulation was taken can be seen in the Support Information.

Contact order

The concept of contact order was introduced by Plaxco, Simons and Baker [43] as a parameter that would demonstrate the importance of local or non-local contacts for a native state protein. The contact order is an average between the pair distances that make a contact, and there are huge evidences that the folding rates for small proteins

with two states is related to the contact order associated with the native structure [41]. If this contact order, called absolute (ACO), is divided by the number of amino acids in a protein, the parameter is now called relative contact order.

Relative contact order

$$RCO = \frac{1}{M.L}\sum^{M} \Delta S_{i,j}, \tag{1}$$

Being $\Delta S_{i,j}$ the distance in residues that separate two amino acids that make a native contact, M the total amount of native contacts and L the amount of amino acids in a protein. The contact order can also be used to monitor the process of folding when calculated in along the parameter Q (fraction of native contacts). In this case, we have the partial contact order [41], given by the following expression:

$$CO(Q) = \frac{\sum_{i=1}^{M_Q} L_i \langle Q_i \rangle}{M_Q}, \tag{2}$$

in which M_Q is the total amount of possible contacts when a protein possess a order parameter Q. L_i is the distance, in residues, along the chain between amino acids that compose the contact i. $\langle Q_i \rangle$ is the probability that the contact i has on forming when there is a configuration with Q fraction of formed native contacts.

Nonnative contacts variation

Any contact between two amino acids separated by, at least four residues, that are far away from each other at maximum $6\mathring{A}$ has been defined as a non-native contact. A limiting factor is that these contacts cannot be present at the first list generated by *CSU* [46]. Another important factor so the contact can be called non-native is that it cannot have a possibility over 30% of being present on structures with Q $>$ 0.9. Thus, in case any contact have a probability over 30% it is taken off the list of non-native contacts and it

is added to a native contacts list.

Having finished the map of native and non-native contacts, it is possible to calculate a partial contact order and make it a reaction coordinate to accompany the folding process. The same procedure to outline the thermodynamic quantities done with WHAM [47, 48] using Q as the order parameter was done using the partial contact order, CO, as coordinate.

The difference in quantity on non-native contacts between the transition state and the unfolded state (ΔA) was made using the average of quantity of non-native contacts ($\bar{A}$) created inside each region.

$$\Delta A = \bar{A}(TS) - \bar{A}(Unf) \tag{3}$$

The place of transition (TS) and the unfolded region (Unf) were boarded using a profile of free energy to each protein. The unfolded region involves all states since the initial configuration (CO = 0) until the configuration in which the protein reaches a value of 15% of ΔF after going through the first minimum in free energy. The transition region involves all configurations located inside the energy barrier with values of free energy above 85% of ΔF. ΔF was calculated based on the peak of the free energy barrier ($F^{\neq}$) and on the first minimum (F_{unf}), as shown in Figure 1.

Effect of nonnative interactions on the free energy barrier

Two order parameters are used to describe the process of folding at analytic model developed by Clementi and Plotkin [45]: the fraction of native contacts (Q) and the fraction of non-native contacts (A). These two order parameters have values between 0 and 1. When Q = 0, the protein is found unfolded, whereas when Q = 1 it is in its native form. The

fraction of non-native contacts depends of Q, so the more native contacts formed, less non-native interactions are allowed, which, eventually, in a theoretical model, does not allowed the presence of non-native contacts when Q = 1. Being the energy of the native state given as E_N and the total amount of contacts in a native state given as M, it is known that:

$$E_N = M\epsilon = zN\epsilon, \tag{4}$$

ϵ is defined as the mean native attraction energy ($\epsilon < 0$). N e z are the number of amino acids of the protein and the amount of contacts per residue, in that order. The variance in the native interaction energies is not considered ($\delta\epsilon^2 = 0$).

Two energy scales are used to analyze the non-native contribution, which are: the mean energy of a nonnative interaction ϵ_{NN} and the energetic variance of nonnative interaction b^2. However, non-native interactions are designed to be weaker when compared to native interactions.

Being the conformational entropy S_c described along with Q and A, the energies of configurations for an ensemble of states characterized by (Q, A) is assumed Gaussianly distributed with mean of $QM\epsilon + AM\epsilon_{NN}$ and variation of AMb^2. Using tools provided by statistical mechanics, thermal energy expressions, free energy and entropy in function with Q, A and T can be found:

$$\frac{E}{M} = \epsilon Q + A\left(\epsilon_{NN} - \frac{b^2}{T}\right) \tag{5}$$

$$\frac{S(Q,A,T)}{M} = \frac{s_c(Q,A)}{z} - \left(\frac{b^2}{2T^2}\right)A \tag{6}$$

in which $s_c(Q, A)$ is the conformational entropy per residue.

$$\frac{F(Q,A,T)}{M} = \epsilon Q + \left(\epsilon_{NN} - \frac{b^2}{2T}\right) A - \frac{T s_c(Q,A)}{z} \tag{7}$$

According to Cecilia and Plotkin, the conformational entropy $S_c(Q,A)$ is calculated using the mean field theory and it is made in terms of non-native packing fraction η. When η=1, the number of non-native contacts reach its maximum $MA_{max} = M\eta(1-Q)$. In this case, the conformational entropy of a polymer with Q native contacts and packing fraction η is given as:

$$S_c(Q,\eta) = N(1-Q)\left\{ ln\frac{\nu}{e} - \left(\frac{1-\eta}{\eta}\right) ln\,(1-\eta) - \frac{1}{6}\left[\left(\frac{\bar{\eta}(Q)}{\eta}\right)^{2/3} - 1\right]^2\right\} \tag{8}$$

$$S_c(Q,\eta) \equiv N(1-Q)s_{NN}(Q,\eta), \tag{9}$$

in which $ln\left(\frac{\nu}{e}\right)$ is the maximum entropy per residue at collapsed state, and $\bar{\eta}(Q) = \bar{l}(Q)^{-1/2}$, in which $\bar{l}$ is the average length of looping formed by native contacts in a certain Q.

According to the analytic model, at the time in which the state of folding and unfolding have the same probability it is said that the protein is in its folding temperature ${T_F}^{\circ}$. This state is given by the Equation 7 when $F(0,A) \approx F(1,0)$ and $\epsilon_{NN} = b^2 = 0$.

Considering $Q \approx 0$ at unfolded state and $A = 0$ at folded state, the folding temperature is as it follows:

$${T_F}^{\circ} = \frac{z|\epsilon|}{s_c(0, A^*(0))} \tag{10}$$

in which $A^*(Q)$ is the most probable value of A in a certain Q.

Using the following definitions:

$$\Delta A^*(Q) \equiv A^*(Q) - A^*(0) \tag{11}$$

$$\Delta s_{nn}(Q) \equiv s_{nn}(Q, A^*(Q)) - s_{nn}(0, A^*(0)) \tag{12}$$

it is possible to calculate the discrepancy of free energy between the unfolded state and a state with any Q and A,

$$\Delta F(Q,T) \equiv F(Q, A^*(Q), T) - F(0, A^*(0), T) \tag{13}$$

Calculating the changing of free energy between a unfolded state and the transition state at folding temperature ${T_F}^{\mathrm{o}}$ is found that: the height of the free energy barrier ($\Delta F^{\neq}$) corresponds to the height of the free energy barrier between the same states at the absence of non-native forces ($\Delta F^{\mathrm{o}\neq}$) added from a term referred to a frustration, as represented in the Equation 14 below:

$$\frac{\Delta F^{\neq}}{{T_F}^{\mathrm{o}}} = \frac{\Delta F^{\mathrm{o}\neq}}{{T_F}^{\mathrm{o}}} + M\left(\frac{\epsilon_{NN}}{{T_F}^{\mathrm{o}}} - \frac{b^2}{2{T_F}^{\mathrm{o}2}}\right)\Delta A^*(Q^{\neq}), \tag{14}$$

Since ϵ_{NN} has always a negative value, the Equation 14 must be analyzed in two circumstances. First, if $\Delta A^*(Q^{\neq}) > 0$, there's a decrease in the free-energy barrier when compared with the absence of non-native interactions ($\Delta F^{\neq} < \Delta F^{\mathrm{o}\neq}$), thus the non-native interactions would be responsible for the decrease of the energy barrier. However, the opposite happens when $\Delta A^*(Q^{\neq}) < 0$, or, in other words, proteins in which this last case happens, the non-native interactions are responsible for the increase of the barrier, decreasing the folding rates.

Results

Contact Order as the reaction coordinate

The contact order calculated in function of the fraction of native contacts proved to be well behaved and able to describe the folding process, reporting for all proteins studied in this work an almost linear growth with the fraction of native contacts, as shown in Figure 2.

Transforming the real values of CO in integer values were necessary to calculate the thermodynamic quantities using WHAM, allowing the existence of degenerate states, i.e., the same contact order could be associated with two or more conformational states. This implies in a decrease of points that form the profiles of free energy, entropy and heat energy when compared to profiles using the fraction of native contacts Q as the order parameter. Another factor that differs the contact order when used as the reaction coordinate, is that it has the ability to tell us what kind of contact is being made in a particular conformational state.

The peak of free energy barrier was displaced to the left in all profiles studied when calculated using the contact order, compared with the position when calculated using Q. This shift shows that, using Q as order parameter, it is known how many contacts are required for the protein overpass the energy barrier, but it isn't known what type of contact is being made. When the contact order is used as coordinated and the peak is displaced to small values of CO, this shows that for the protein to overpass the free energy barrier and get in the folded state, short-range contacts are initially needed. The presence of short-range contacts in settings located in the pre-transition state was not only evident in proteins formed exclusively of α-helix, but it is also quite visible in proteins displaying large amount of long-range contacts, such as those formed solely by β-sheets.

Nonnative contacts variation

According to the theoretical model developed by Clementi and Plotkin, positive or negative sign of the variation in the amount of non-native contacts will determine if frustration favors, or not, the folding process, so we have in Figure 3 a typical representation of the two cases.

According to the Figure 3, it is noted that for proteins in which the frustration hardly contributes favorably to the folding process (represented by ENHD protein) in none, or in a few states accessed by the protein throughout the process, the average number of non-native contact exceeds the value of the point located inside the unfolded state. On the other hand, it is seen that for proteins which frustration has great potential to assist the folding process, the non-native contacts quantity increases considerably after the point chosen to represent the unfolded state. And all cases, the largest numbers of non-native contacts appear on the pre-transition region, as described in the theoretical model.

All results obtained in the calculation of the nonnative contacts variation can be seen in Table I.

Among the 19 proteins tested, three didn't show free energy barrier using either Q or CO as the reaction coordinate. In this case, if the protein has no free energy barrier, there is no reason to reduce it even more, so for these proteins, the addition of frustration hinders the folding.

Only four proteins opposed the analytical theory with the data obtained computationally by Contessoto ET al, in which proteins: IM9, ACBD, HHCC and CI2 have an optimal frustration different of zero, and it was expected that the ΔA of these proteins would be positive. However, except for CI2, the optimal frustration for the other three proteins is the lowest value that was tested by Contessoto et al.

Perhaps, further refining the calculation of the optimal frustration for each protein

tested by Contessoto et al, would eliminate the results that came into disagreement with what was proposed in the analytical model, ie the optimal frustration for HHCC, ACBD, IM9 and CI2 could be changed for $\epsilon_f^{opt} = 0$. On the other hand, it sounds too abrupt to define that the energetic frustration favors the folding looking at only the sign of the variation in the amount of non-native contacts. In the next section, will be introduced the idea of an intermediate region between the two extremes, separated mainly by the energy barrier height and the native structure acquired.

Correlation between free energy barrier, contact order and non-native contacts

According to the analytical model, the assistance in the folding process is directly correlated with the amount of non-native contacts which are formed in the transition state compared to the amount of non-native contacts that appear in the unfolded state, and according to this quantity, energetic frustration can increase or decrease the free energy barrier. The Figure 4 brings us the exact correlation between the free energy barrier height and the variation in the amount of non-native contacts according to the data of 19 proteins studied.

The linear correlation value between the variation in the amount of non-native contacts and free energy barrier height shows that there is a strong correlation between these two parameters, which in a way confirms the analytical model. In Figure 4 (a), we can check that the proteins were gently separated into three groups, separated in the order of 1kT difference in height of the barrier;

Group A = ACR, HP36,PSBD, $\alpha 3D$, PtABD e EnHD;

Group B = IM9, ACBD, HHCC, PtL, PtG, ADA2h, CI2, SH3 e Ubiquitin;

Group C = CSPTm, HPr, αAIT, TWIg;

The formation of the three groups is thought as a convergence point between the results proposed by the analytical model and the results obtained in the determination of optimal frustration calculated by Contessoto et al, since the formation of these three groups leads us to a possible intermediate group (group B) that would be among the group in which the proteins are unquestionably favored (group C) by the addition of frustration and the group in which proteins aren't favored (group A), and those four proteins (IM9, HHCC, ACBD and CI2) are located in this middle group.

Separating the proteins in these three groups, it is noted that the group whose frustration impair the folding process would be populated mainly by proteins with helix-α structure. Already the intermediate group would be formed mainly by proteins with the association of helix-α and β-sheets, and at last, in the group in which proteins would be favored by the addition of frustration would be formed primarily by proteins with β-sheets. Given the relationship between the groups treated with the protein's native structure, it is to be expected a good correlation between ΔA and contact order, which can be seen in Figure 4 (b).

The good correlation between ΔA, ΔF and RCO makes the association between the contact order and the free energy barrier height good candidates to predict the effect of energetic frustration on the proteins, as well as proposed by Contessoto et al. The linear correlation value of 0.84 between RCO$\times\Delta F$ with ΔA illustrated in Figure 5 closer again the analytical model and the computational results, confirming the good correlation between the amount of non-native contacts with the presence of an optimal frustration.

Conclusions

The contact order used as a coordinated to accompany the folding process showed itself able to replace the well-established fraction of native contacts in the analytical model, since all the results obtained, using the fraction of native contacts, agreed qualitatively with the results obtained via partial contact order. Furthermore, the use of the contact order to describe the folding process brings us different information from those ones obtained when performing the fraction of native contacts. This fact increases our level of knowledge on the matter. Such information made explicit the necessity of the protein to form short-range native contacts to starts de folding process, and to form long-range contacts to overcome the free-energy barrier. This dependence must be correlated with the entropic cost of forming long-range contacts in relation to the entropic cost of forming short-range contacts.

Although it appears to be controversial that the presence of non-native contacts may favor the folding process, this must be explained mainly because, when inserting an energetic frustration, we add an attractive interaction energy in the midst of the contacts, that, in some ways, may favor the approach of amino acids that will make contacts in the native structure, accelerating the kinetics of the process, especially in proteins with β-sheet structures, due to the a larger presence of long-range contacts and, as discussed in the paragraph above, long-range contacts are formed with higher probability after the proteins win the energy barrier. The energetic frustration has a limit. In order to this limit act favorable on the folding process, it must be related to a threshold value in which the native state is still more stable than unfolded states.

The good correlation among barrier height, variation in the fraction of non-native contacts and the contact order brings closer the theoretical model developed by Clementi

and Plotkin to the results for real proteins by means of computer simulation. In addition, we have further evidence that these parameters can be used to predict the addition of frustration effects in folding, without insert specific terms of frustration in the computational simulation. Thus, according to the results, we see that the addition of frustration hardly favors the proteins that have low free-energy barrier and low contact order, formed strictly by α-helix. So, for proteins with high energy barriers and high contact order, typically represented by proteins with large presence of β-sheets, the increase of energetic frustration probably favors the folding process, decreasing the folding time.

Acknowledgments

PRM and VGC were supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brazil. RJO is funded by Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG). VBPL is funded by Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP). JC and VBPL were supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). We also thank GridUnesp for the computational resources.

References

1. Dobson Christopher M.. Protein folding and misfolding. Nature 2003;426:884–890. URL http://dx.doi.org/10.1038/nature02261.

2. Baldwin R.L.. The nature of protein folding pathways: The classical versus the new view. J Biomol NMR 1995;5:103–109.

3. Dill K. A., Chan H. S.. From Levinthal to pathways to funnels. New Journal of Physics 1997;4:10–19.

4. Pande V. S., Grosberg A. Y., Tanaka T.. On the theory of folding kinetics for short proteins. Folding and Des 1997;2:109–114.

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

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

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

8. Wolynes P. G., Onuchic J. N., Thirumalai D.. Navigating the folding routes. Science 1995;267:1619–1620.

9. Wang J., Oliveira R. J., Chu X., Whitford P. C., Chahine J., Han W., Wang E., Leite V. B. P.. The topography of funneled landscapes determines the thermody-

namics and kinetics of protein folding. Proc Natl Acad Sci USA 2012;109:15763–15768.

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

11. Shoemaker B. A., Wang J., Wolynes P. G.. Structural correlations in protein folding funnels. Proc Natl Acad Sci USA 1997;94:777–782.

12. Whitford P. C., Noel J. K., Gosavi S., Schug A., Sanbonmatsu K., Onuchic J. N.. An all-atom structure-based potential for proteins: Bridging minimal models with all-atom empirical forcefields. Proteins: Struct Funct Bioinf 2009;75:430–441.

13. Oliveira R. J., Whitford P. C., Chahine J., Leite V.B.P., Wang J. Coordinate and time-dependent diffusion dynamics in protein folding. Methods 2010;52:91–98.

14. Fersht A. R.. Characterizing transition states in protein folding: an essential step in the puzzle. Curr Opin Struct Biol 1995;5:79–84.

15. Garcia-Mira M. M., Sadqi M., Fischer N., Sanchez-Ruiz J. M., Muñoz V.. Experimental identification of downhill protein folding. Science 2002;298:2191 –2195.

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

17. Chung H. S., Louis J. M., Eaton W. A.. Experimental determination of upper bound for transition path times in protein folding from single-molecule photon-by-photon trajectories. Proc Natl Acad Sci USA 2009;106:11837 –11844.

18. Koga N., Takada S.. Roles of native topology and chain-length scaling in protein folding: A simulation study with a gō-like model. J Mol Biol 2001;313:171–180.

19. Chavez L. L., Onuchic J. N., Clementi C.. Quantifying the roughness on the free energy landscape: Entropic bottlenecks and protein folding rates. J Am Chem Soc 2004;126:8426–8432.

20. Snow C. D., Sorin E. J., Rhee Y. M., Pande V. S.. How well can simulation predict protein folding kinetics and thermodynamics? Annu Rev Biophys Biomol Structure 2005;34:43–69.

21. Gosavi S., Chavez L. L., Jennings P. A., Onuchic J. N.. Topological frustration and the folding of interleukin-1 beta. J Mol Biol 2006;357:986–996.

22. Chu Xiakun, Gan Linfeng, Wang Erkang, Wang Jin. Quantifying the topography of the intrinsic energy landscape of flexible biomolecular recognition. Proceedings of the National Academy of Sciences 2013;201220699. URL `http://www.pnas.org/content/early/2013/06/06/1220699110`.

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

24. Fersht Alan R. Nucleation mechanisms in protein folding. Curr Opin Struct Biol 1997;7:3–9. URL `http://www.sciencedirect.com/science/article/pii/S0959440X97800024`.

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

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

27. Xu Weixin, Lai Zaizhi, Oliveira Ronaldo J., Leite Vitor B. P., Wang Jin. Configuration-dependent diffusion dynamics of downhill and two-state protein folding. J Phys Chem B 2012;116:5152–5159. URL http://dx.doi.org/10.1021/jp212132v.

28. Onuchic J. N., Wolynes P. G., Luthey-Schulten Z., Socci N. D.. Toward an outline of the topography of a realistic protein-folding funnel. Proc Natl Acad Sci USA 1995;92:3626–3630.

29. Onuchic J. N., Luthey-Schulten Z., Wolynes P. G.. Theory of protein folding: The energy landscape perspective. Annu Rev Phys Chem 1997;48:545–600.

30. Onuchic J. N., Nymeyer H., Garcia A. E., Chahine J., Socci N. D.. The energy landscape theory of protein folding: Insights into folding mechanisms and scenarios. In Adv. Protein Chem.. Elsevier. volume 53. 2000; 87–152.

31. Eaton William A, Thompson Peggy A, Chan Chi-Kin, Hage Stephen J, Hofrichter James. Fast events in protein folding. Structure 1996;4:1133–1139. URL http://www.sciencedirect.com/science/article/pii/S0969212696001219.

32. Eaton William A, Muñoz Victor, Thompson Peggy A, Chan Chi-Kin, Hofrichter James. Submillisecond kinetics of protein folding. Curr Opin Struct Biol 1997;7:10–14. URL http://www.sciencedirect.com/science/article/pii/S0959440X97800036.

33. Ozkan S. Banu, Bahar Ivet, Dill Ken A.. Transition states and the meaning of phi-values in protein folding kinetics. Nature Structural & Molecular Biology 2001;8:765–769. URL http://www.nature.com/nsmb/journal/v8/n9/full/nsb0901-765.html.

34. Shakhnovich E. I., Gutin A. M.. Formation of unique structure in polypeptide chains. theoretical investigation with the aid of a replica approach. Biophys Chem 1989;34:187–199.

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

36. Goldstein R. A., Luthey-Schulten Z. A., Wolynes P. G.. Optimal protein-folding codes from spin-glass theory. Proc Natl Acad Sci USA 1992;89:4918 –4922.

37. Abkevich V. I., Gutin A. M., Shakhnovich E. I.. Free energy landscape for protein folding kinetics: Intermediates, traps, and multiple pathways in theory and lattice model simulations. J Chem Phys 1994;101:6052–6062. URL http://scitation.aip.org/content/aip/journal/jcp/101/7/10.1063/1.467320.

38. Gō Nobuhiro. Theoretical studies of protein folding. Ann Rev Biophys Bioeng 1983;12:183–210.

39. Clementi C., Nymeyer H., Onuchic J. N.. 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 2000;298:937–953.

40. Papoian Garegin A., Wolynes Peter G.. The physics and bioinformatics of binding and folding – an energy landscape perspective. Biopolymers 2003;68:333–349. URL http://onlinelibrary.wiley.com/doi/10.1002/bip.10286/abstract.

41. Clementi C., Garcia A. E., Onuchic J. N.. Interplay among tertiary contacts, secondary structure formation and side-chain packing in the protein folding mechanism: All-atom representation study of protein L. J Mol Biol 2003;326:933–954.

42. Contessoto V. G., Lima D. T., Oliveira R. J., Bruni A. T., Chahine J., Leite V. B. P.. Analyzing the effect of homogeneous frustration in protein folding. Proteins: Struct Funct Bioinf 2013;81:1727–1737.

43. Plaxco K. W., Simons K. T., Baker D.. Contact order, transition state placement and the refolding rates of single domain proteins. J Mol Biol 1998;277:985–994.

44. Plotkin S. S.. Speeding protein folding beyond the Gō model: How a little frustration sometimes helps. Proteins: Struct Funct Genet 2001;45:337–345.

45. Clementi C., Plotkin S. S.. The effects of nonnative interactions on protein folding rates: Theory and simulation. Prot Sci 2004;13:1750–1766.

46. Sobolev V., Wade R., Vried G., Edelman M.. Molecular docking using surface complementarity. Proteins: Struct Funct Genet 1996;25:120–129.

47. Ferrenberg A. M., Swendsen R. H.. New monte carlo technique for studying phase transitions. Phys Rev Lett 1988;61:2635–2638.

48. Ferrenberg A. M, Swendsen R. H.. Optimized monte-carlo data analysis. Phys Rev Lett 1989;63:1195–1198.

Table

Table I. Data obtained for the 19 proteins studied.

Protein	M	COt	$Q_M^{\neq} \div CO_{COt}^{\neq}$ [a]	$\Delta F(Q^{\neq})$	$\Delta F(CO^{\neq})$	$\Delta A(Q)^b$	$\Delta A(CO)^b$
ACR*	58	6	–	0.00	0.00	–	–
HP36*	55	8	–	0.00	0.00	–	–
PSBD*	64	11	–	0.00	0.00	–	–
$\alpha_3 D$	136	15	1.38	0.21	0.52	-5.44	-11.6
PtABD	102	15	1.52	0.37	0.75	-0.98	-5.80
EnHD	111	16	1.11	0.75	0.74	-9.81	-13.3
IM9	178	23	1.37	1.80	2.64	-4.44	-5.88
ACBD	182	26	1.08	1.32	2.03	-0.52	-6.80
HHCC	246	25	1.42	1.59	2.45	-7.10	-8.70
PtL	136	19	1.19	2.34	3.21	2.05	5.60
PtG	139	21	1.24	2.94	3.78	4.17	7.50
ADA2h	175	29	1.05	2.52	3.09	3.60	4.00
CI2	152	21	1.21	3.17	4.01	-1.25	-2.60
SH3	152	22	1.02	4.03	4.38	3.94	3.40
Ubiquitin	188	24	1.39	4.27	5.01	1.27	1.46
CSPTm	180	20	1.20	5.95	6.66	5.00	3.78
HPr	222	31	1.23	5.69	6.77	1.30	1.85
αAIT	196	27	1.30	6.65	7.00	11.68	9.56
TWIg	253	33	1.79	5.98	6.93	10.19	11.32

[a] Relationship between the free energy peak position when calculated in terms of Q and when calculated in terms of CO.

[b] Variation in the fraction of nonnative contacts $\times 10^{-3}$.

* Proteins that showed no free energy barrier

Figure legends

Figure 1. Free energy profile (F) in function with the contact order (CO). CO^u and CO^{ts} are, as it follows, the contact order of the first minimum of free energy, and the contact order of the peak of a free energy barrier. $\Delta F = F(CO^{ts}) - F(CO^u)$. The region between two first lines traced is called unfolded region (Unf) and the region limited by the third and forth line, which envolves the free energy barrier peak, is called transition region (TS).

Figure 2. Contact order (CO) in function of native contacts fraction (Q). (a) Data obtained for the EnHD protein. (b) Data obtained for the CSPTm (protein).

Figure 3. Non-native contacts in function of the normalized contact order. (a) Average number of non-native contacts for the EnHD protein. (b) Average number of non-native contacts for the CSPTm protein. (c) Variation of non-native contacts for the EnHD protein. (d) Variation of non-native contacts for the CSPTm protein. The traced lines in Figures (a) and (b) indicate where the regions were limited: unfolded (Unf), transition (TS) and pre-transition (P-TS). $\Delta A^* = \bar{A}(CO^{ts}) - \bar{A}(CO^{unf})$. CO^{ts} is a conformational state localized inside the transition region and CO^{unf} is the state where the firs minimum of free energy is located.

Figure 4. (a)Correlation between the free-energy barrier height and the variation of the amount of non-native contacts. Data obtained from Table 1 using the contact order as coordinate of reaction. The proteins are colored according to the native structure in the PDB. ΔF is the height of free-energy barrier while ΔA is the variation in non-native contacts fraction. The linear correlation between ΔF and (ΔA) is of 0.81.

(b)Correlation between the relative contact order and the variation in quantity of non-native contacts. Data obtained from Table 1 using the contact order as coordinate of reaction. The proteins are colored according to the optimal frustration determined by Contessoto ET al [42]. RCO is the relative contact order, ΔA is the variation of non-native contacts fraction. The correlation between RCO and (ΔA) is of 0.81.

Figure 5. Correlation between RCO $\times\Delta F$ e ΔA. Data obtained using the contact order as coordinate reaction. The proteins are colored according to the good frustration determined by Contessoto ET al [42] and are represented according to the native structure in PDB. The linear correlation between RCO $\times\Delta F$ and (ΔA) is of 0.84.

Figures

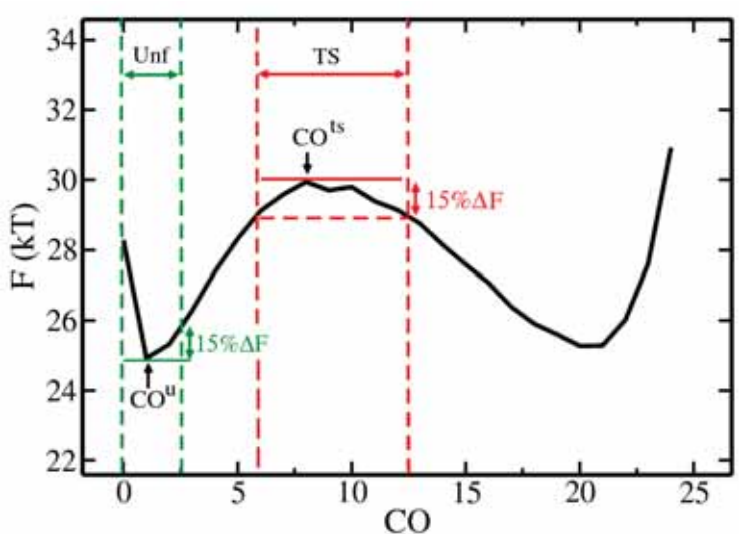

Figure 1

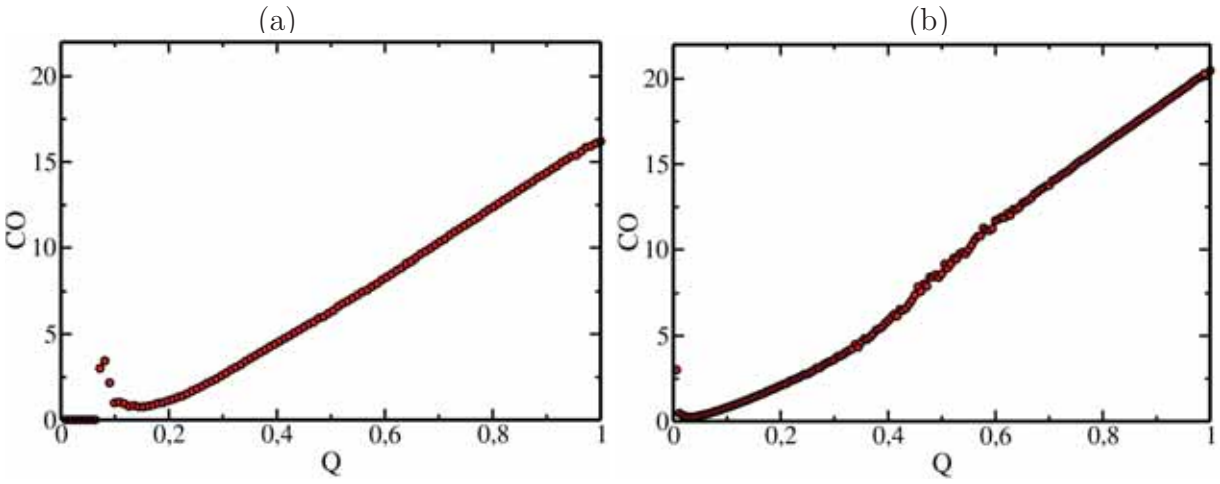

Figure 2

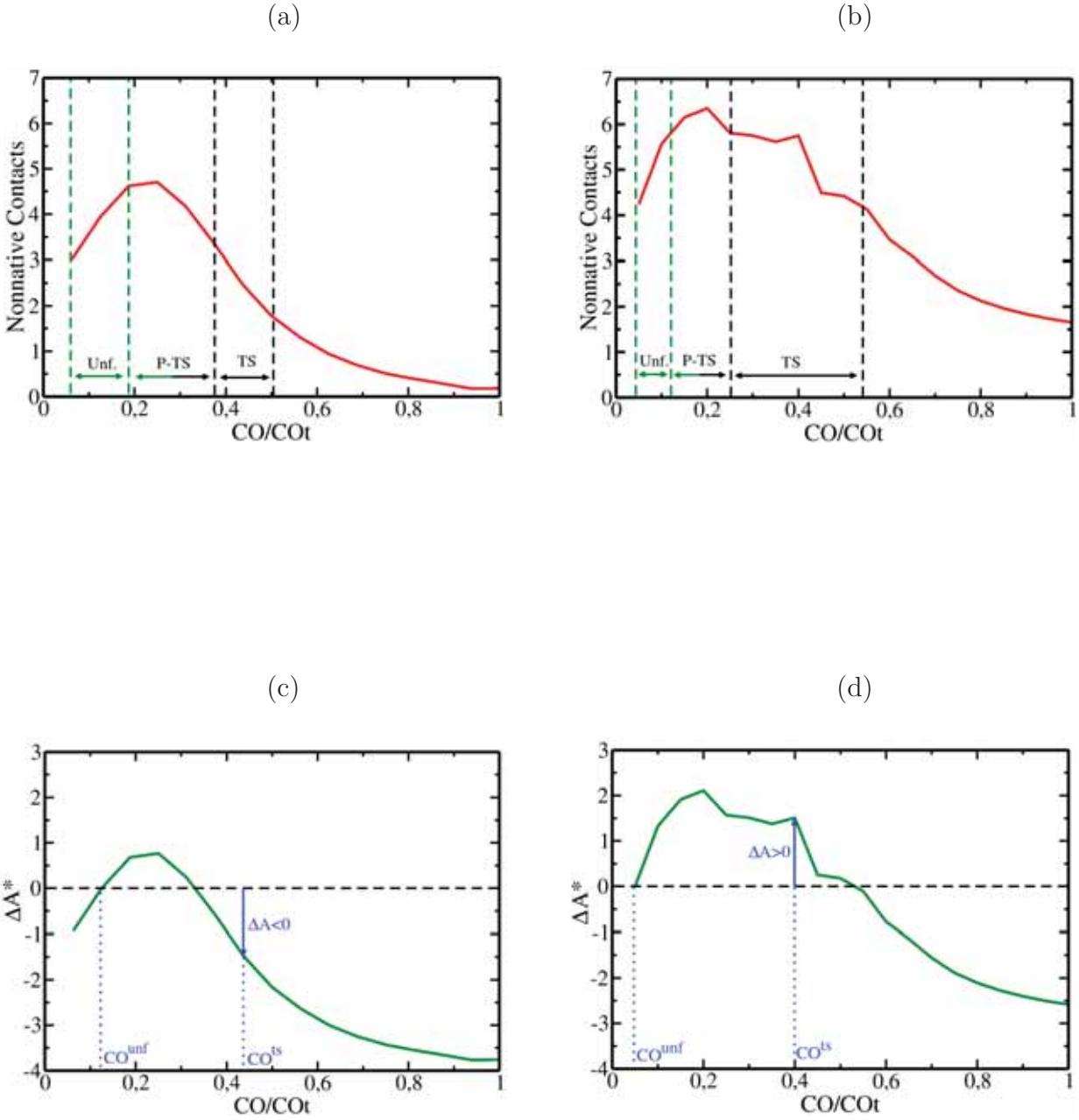
(a)
(b)
Nonnative Contacts
CO/COt
Unf.
P-TS
TS
(c)
(d)
ΔA*
ΔA<0
ΔA>0
CO^unf
CO^ts

Figure 3

(a) (b)

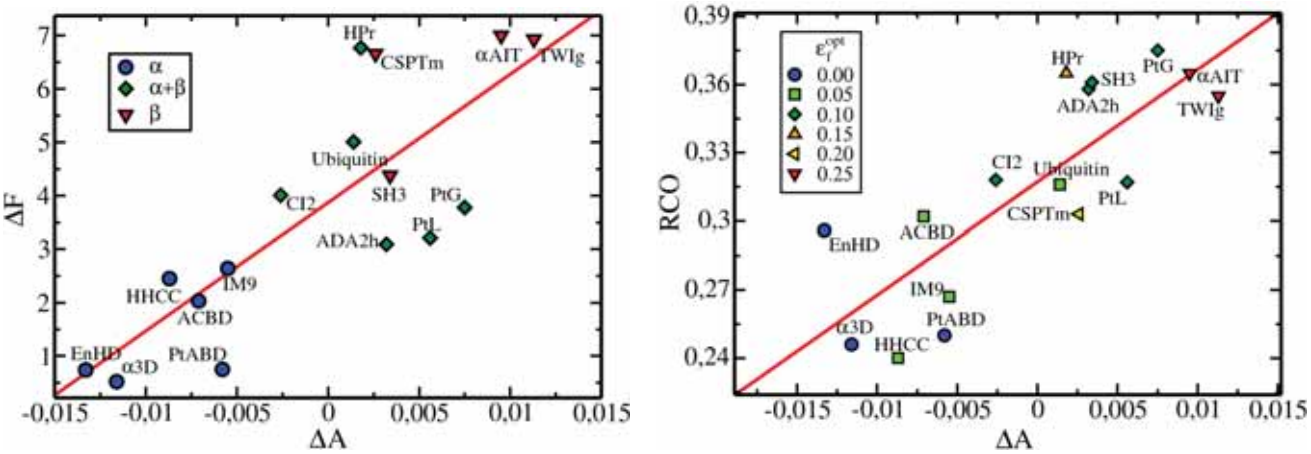

Figure 4

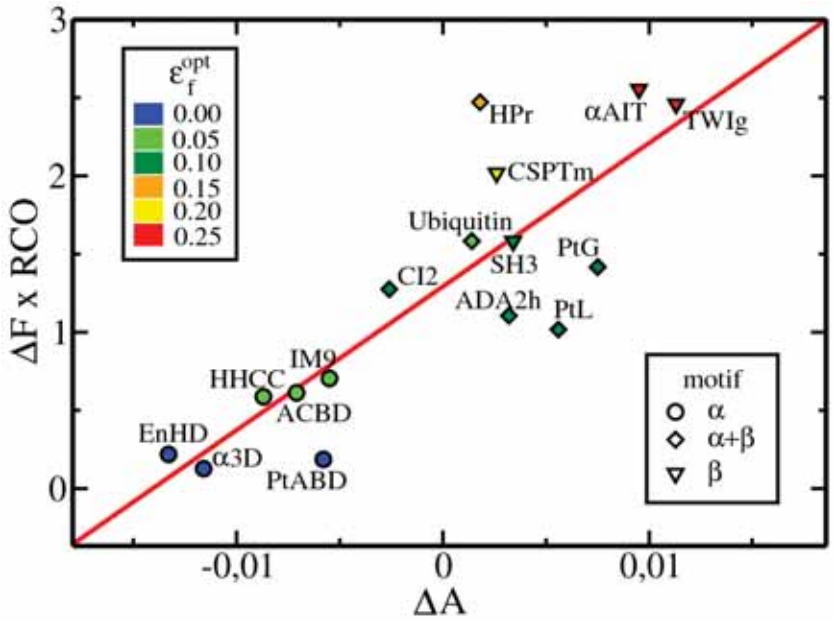

Figure 5

Support Information

Quadro 1 Proteínas Estudadas

NOME	PDB	# a.a.	M	Cartoon	NOME	PDB	# a.a.	M	Cartoon
ACR	1ARR	53	58		PtG	2K0P	56	139	
HP36	1VII	36	55		ADA2h	1PBA	81	175	
PSBD	2PDD	43	64		CI2	1CIS	66	152	
α3D	2A3D	73	136		SH3	1FMK	61	152	
PtABD	1BDC	60	102		Ubiquitin	1UBQ	76	188	
EnHD	1ENH	104	111		CSPTm	1G6P	66	180	
IM9	1IMP	86	178		αAIT	2AIT	74	196	
ACBD	2ABD	86	182		HPr	1HDN	85	222	
HHCC	1HRC	104	246		TWIg	1WIU	93	253	
PtL	2PTL	60	136						

#a.a. = Número de aminoácidos na cadeia. M = Contatos Nativos

Structure-based C_α model

O modelo Cα é dito um modelo baseado em estrutura visto que utiliza parâmetros obtidos da resolução estrutural de proteínas depositadas em bancos de dados, como o PDB [1], a

fim de construir o potencial energético que define as conformações de uma dada proteína. Durante a dinâmica no qual o modelo Cα é submetido, o potencial que define a energia das conformações é do tipo G$\bar{o}$ [2] no qual a ideia principal é dar importância para as interações entre aminoácidos que residem nos contatos nativos, e em seguida, escolher a energia dos contatos que minimizam a energia total do estado nativo. A simplicidade do modelo Cα está baseada na referência à cadeia de aminoácidos como uma cadeia de esferas simples centralizadas nas posições dos carbonos alfas. Tais esferas são mantidas juntas por meio de ligações e ângulos de interação, e a geometria nativa fica contida no potencial diedro e no termo de interação não-local.

A expressão que define a energia de uma configuração Γ baseada na conformação nativa Γ^0 para o modelo Cα é dada por:

$$\begin{aligned} V(\Gamma,\Gamma_o) &= \sum_{bonds} \epsilon_r (r-r_o)^2 \\ &+ \sum_{angles} \epsilon_\theta (\theta-\theta_o)^2 \\ &+ \sum_{dihedrals} \epsilon_\phi \left\{ [1-cos(\phi-\phi_o)] + \frac{1}{2}[1-cos(3(\phi-\phi_o))] \right\} \\ &+ \sum_{contacts} \epsilon_C \left[5\left(\frac{d_{ij}}{r_{ij}}\right)^{12} - 6\left(\frac{d_{ij}}{r_{ij}}\right)^{10} \right] \\ &+ \sum_{non-contacts} \epsilon_{NC} \left(\frac{\sigma_{NC}}{r_{ij}}\right)^{12} \end{aligned} \quad (1)$$

De acordo com a Equação 1, o potencial que rege este modelo baseado em estrutura é calculado via a contabilização de 5 termos referentes a algumas das interações que atuam na estrutura nativa. O primeiro termo remete a um potencial harmônico representando a ligação entre dois carbonos α adjacentes, no qual r_0 é a distância entre os dois carbonos α ligados entre si na estrutura nativa. A segunda somatória também forma um potencial

harmônico, mas desta vez, um potencial harmônico angular formado por três carbonos α em sequência na cadeia polipeptídica, no qual θ_0 é o ângulo formado pelos três resíduos na conformação nativa. Já o terceiro termo da expressão 5 leva em consideração a torção realizada pela cadeia, no qual, ϕ_0 é o angulo diédro formado por quatro carbonos α em sequência. O quarto termo contabiliza a interação entre o carbono α i e j que formam um contato na estrutura nativa, para isso é utilizado um potencial 10-12, no qual d_{ij} é o valor da distância entre estes carbonos que realizam um contato nativo. Por fim, o último termo remete a todos os carbonos α que não realizam um contato nativo. Este termo é utilizado para manter a distância máxima de aproximação entre os carbonos α, no qual σ_{NC} possui valor de $4\mathring{A}$ e está correlacionado com o volume ocupado por um carbono no modelo. As constantes: ϵ_r, ϵ_θ, ϵ_ϕ e ϵ_{NC} são todas dadas em unidades de ϵ_c e valem, respectivamente, 100, 20, 1 e 1. [2, 3]

Simulation details

Para testar a correlação entre a diferença na quantidade de contatos não-nativos e a presença de uma frustração ótima, utilizamos como objeto de estudo um grupo de 19 proteínas [3].

A partir dos dados estruturais depositados no PDB [1], extraímos o primeiro mapa de contato nativo utilizando o CSU [4].A dinâmica molecular utilizando o arquivo de topologia gerado pelo Structure-based Models in Gromacs (SMOG) disponível online [5] foi elaborada com os recursos computacionais do Shiva Cluster do departamento de Física do Instituto de Biociências, Letras e Ciências exatas da UNESP e realizada com o pacote de dinâmica molecular *GROMACS versão 4.5-5* [6]. As proteínas foram simuladas sobre 10^9 passos tendo as informações guardadas a cada 5000 passos, obtendo no total 200000 frames. A coordenada de reação utilizada para acompanhar o enovelamento foi a fração

de contatos nativos Q, sendo que de acordo com a dinâmica, um contato nativo era aceito em uma conformação Γ se a distância entre os aminoácidos nesta conformação era menor do que $1.2d_{ij}$, no qual d_{ij} é a distância entre os resíduos na estrutura nativa.

Os perfis de grandezas termodinâmicas, tais como, Energia térmica, Energia Livre, Entropia e calor específico foram calculados por meio do método dos múltiplos histogramas (WHAM - *Weighted Histograms Analysis Method [7, 8])sendo que a temperatura de enovelamento* T_F^0 *foi definida como a temperatura do maior valor referente ao calor específico.*

References

1. *Berman Helen M., Westbrook John, Feng Zukang, Gilliland Gary, Bhat T. N., Weissig Helge, Shindyalov Ilya N., Bourne Philip E.. The protein data bank. Nucleic Acids Res 2000;28:235–242.*

2. *Clementi C., Nymeyer H., Onuchic J. N.. 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 2000;298:937–953.*

3. *Contessoto V. G., Lima D. T., Oliveira R. J., Bruni A. T., Chahine J., Leite V. B. P.. Analyzing the effect of homogeneous frustration in protein folding. Proteins: Struct Funct Bioinf 2013;81:1727–1737.*

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

5. *Noel J. K., Whitford P. C., Sanbonmatsu K. Y., Onuchic J. N.. SMOG@ctbp: simplified deployment of structure-based models in GROMACS. Nucleic Acids Research 2010;.*

6. *Van Der Spoel D., Lindahl E., Hess B., Groenhof G., Mark A. E., Berendsen H. J. C.. GROMACS: fast, flexible, and free. J Comp Chem 2005;26:1701–1718.*

7. *Ferrenberg A. M., Swendsen R. H.. New monte carlo technique for studying phase transitions. Phys Rev Lett 1988;61:2635–2638.*

8. *Ferrenberg A. M, Swendsen R. H.. Optimized monte-carlo data analysis. Phys Rev Lett 1989;63:1195–1198.*

Autorizo a reprodução xerográfica para fins de pesquisa.

São José do Rio Preto, 24 / 04 / 2014

Paulo Ricardo Mauro

Assinatura