



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

---

Antonio Bento de Oliveira Junior

---

Métodos para visualização de superfície  
de energia do enovelamento de  
proteínas

---

São José do Rio Preto  
2017

**Antonio Bento de Oliveira Junior**

**Métodos para visualização de superfície de energia  
do enovelamento de proteínas**

Tese apresentada como parte dos requisitos para obtenção do título de Doutor 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.

Financiadora: CAPES DS

Orientador: Prof. Dr. Vitor B. Pereira Leite

**São José do Rio Preto  
2017**

Oliveira Junior, Antonio Bento de.

Métodos para visualização de superfície de energia do enovelamento de proteínas / Antonio Bento de Oliveira Junior. -- São José do Rio Preto, 2017

59 f. : il., tabs.

Orientador: Vitor B. Pereira Leite

Tese (doutorado) – Universidade Estadual Paulista “Júlio de Mesquita Filho”, Instituto de Biociências, Letras e Ciências Exatas

1. Biologia molecular. 2. Biofísica. 3. Enovelamento de proteínas.  
I. Universidade Estadual Paulista "Júlio de Mesquita Filho". Instituto de Biociências, Letras e Ciências Exatas. II. Título.

CDU – 577.3

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

**Antonio Bento de Oliveira Junior**

**Métodos para visualização de superfície de energia  
do enovelamento de proteínas**

Tese apresentada como parte dos requisitos para obtenção do título de Doutor 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.

Financiadora: CAPES DS

**Comissão Examinadora**

Prof. Dr. Vitor B. Pereira Leite

UNESP - São José do Rio Preto - SP

Orientador

Prof. Dr. Jorge Chahine

UNESP - São José do Rio Preto - SP

Prof. Dr. Pedro Geraldo Pascutti

UFRJ - Rio de Janeiro - RJ

Prof. Dr. Laurent Emmanuel Dardenne

LNCC - Petrópolis - RJ

Prof. Dr. Antonio Francisco P. de Araújo

UNB - Brasília - DF

São José do Rio Preto, 05 de Dezembro de 2017

# Agradecimentos

*Gostaria de começar agradecendo uma pessoa em especial que, provavelmente, só conseguir chegar até aqui graças a ela, a minha mãe Roseli, sempre me apoiando, e me dando a força necessária para tomar decisões importantes na minha vida, muito obrigado!*

*À galera da pós-graduação em Biofísica Molecular, em especial, o pessoal do grupo do professor Vitor, pelas opiniões e sugestões que sempre contribuíram muito pra realização desse trabalho. E que me ajudaram a perceber que a formação acadêmica vai muito além da formação técnica.*

*Ao meu Orientador, Vitor Leite, que sempre me apoiou e acreditou em mim. Pela oportunidade oferecida e sua destreza na orientação, desenvolvendo várias discussões e sugestões e, principalmente, me ensinando a ter um pensamento científico.*

*Gostaria de agradecer a todos os meus amigos, que entre cervejas, rock'n roll e muitas conversas boas, me aturaram e me ajudaram a contornar todas as situações difíceis que passei durante todo o meu período de formação, obrigado a todos aqueles momentos inesquecíveis.*

*Aos meus professores da graduação e pós-graduação, pelo seus ensinamentos, tanto dentro como fora da sala de aula, e pela boa convivência durante meus anos no departamento. Agradeço aos professores Alexandre Suman de Araújo e Sidney Jurado de Carvalho por participarem da minha banca de Qualificação, me aconselhando para o término deste trabalho. Agradeço também a todos os funcionários do departamento de Física, por sempre estarem dispostos a ajudar.*

*Agradeço as agências financiadoras Capes e Fapesp, pelo suporte financeiro. Também agradeço o Gridunesp pelo tempo computacional que foi tão importante para a realização desse trabalho.*

*À todos vocês, meu muitíssimo obrigado!*

“There are two possible outcomes: if the result confirms the hypothesis, then you’ve made a measurement. If the result is contrary to the hypothesis, then you’ve made a discovery”

*Enrico Fermi*

# Resumo

O enovelamento de proteínas acontece em um espaço de fase multidimensional, onde o número conformações possíveis é exponencialmente alta. Uma forma comum de representar essas conformações é utilizar uma coordenada de reação efetiva (por exemplo, fração de contatos nativos). Porém, como a informação de cada conformação não é representada neste tipo de aproximação estatística, alguns mecanismos do enovelamento de proteínas não são possíveis de ser descritos ou analisados. Neste trabalho, usou-se uma métrica para descrever a distancia entre quaisquer duas conformações, essa métrica é calculada levando em conta as distâncias internas dos aminoácidos presentes em cada estrutura. Utilizando-se um método de projeção efetiva é possível ir além da representação em uma dimensão e visualizar a superfície de enovelamento da proteína em duas ou três dimensões. Para aplicar essa metodologia realizou-se simulações computacionais do enovelamento de proteínas utilizando o modelo baseado em estrutura, com aproximação para  $C\alpha$ . Três proteínas foram analisadas: CI-2, o Domínio SH3 e a Proteína A. Dos resultados, foi possível observar que para cada tipo de "motifs" estrutural (folha- $\beta$  e/ou  $\alpha$ -hélice) projetou funis de enovelamento distintos. A partir da visualização foi possível analisar o processo de enovelamento em detalhes, sendo possível identificar a conectividade entre as conformações assim como, possíveis rotas de enovelamento (*foldings pathways*). Analisou-se também as diferenças estruturais da rota dominante no domínio SH3 e a competitividade entre a estrutura do estado nativo e do estado espelhado que acontece em proteínas que possuem somente  $\alpha$ -hélice, como é o caso da proteína A.

**Palavras-chave:** Enovelamento de Proteína, Superfícies de energia, Redução Multidimensional.

# Abstract

Protein folding occurs in a very high dimensional phase space, in which an exponentially large number of states is represented in terms of one effective reaction coordinate. Since the role of each local minimum is not considered in this statistical approach, the folding mechanism is unveiled by describing the local minima in an effective one-dimensional representation. In this work, we used a metric to describe the distance between any two conformations, which is based on internal distances between amino acids in each conformation. A effective projection method allows to go beyond the one-dimensional representation and visualizing a 2D folding funnel representation. Computer simulations of protein folding were performed using  $C\alpha$  structure-based model. Three proteins have been studied: CI2, SH3 Domain and Protein-A. Distinct funnels have been generated according to the major motifs in each proteins, ( $\beta$ -sheet or/and  $\alpha$ -helix). The visualization allows assessing the folding process in detail, e.g. by identifying the connectivity between conformations and establishing the paths that lead to the native state and we analyzed structural differences in the dominant route of SH3 and the competitiveness between the native and mirror structures in protein A.

**Keywords:** Protein Folding, Energy Landscapes, Multidimensional Reduction





# Sumário

<b>Lista de Figuras</b>	<b>xi</b>
<b>Lista de Tabelas</b>	<b>xvii</b>
<b>1 Introdução</b>	<b>1</b>
1.1 Aspectos gerais de proteínas . . . . .	1
1.2 O enovelamento de proteínas . . . . .	3
1.3 A teoria de superfície de energia . . . . .	5
1.4 Visualização do funil de enovelamento de proteínas . . . . .	7
<b>2 Metodologia</b>	<b>11</b>
2.1 Modelo baseado em estrutura . . . . .	11
2.2 Matriz de dissimilaridades . . . . .	13
2.3 Processamento dos dados . . . . .	14
2.3.1 Clusterização multidimensional . . . . .	15
2.4 Projeção multidimensional . . . . .	17
2.5 Proteínas estudadas . . . . .	19
<b>3 Resultados e Discussões</b>	<b>21</b>
3.1 Descrição da superfície em 2D . . . . .	21
3.2 Visualização da proteína CI2, Domínio SH3 e Proteína A . . . . .	25

## Sumário

---

3.3	Energia livre em função de uma coordenada de reação . . . . .	26
3.4	Rotas dominantes de enovelamento no Domínio SH3 . . . . .	27
3.5	Análise das imagens espelhadas na proteína A . . . . .	31
<b>4</b>	<b>Conclusões</b>	<b>35</b>
	<b>Referências Bibliográficas</b>	<b>39</b>
<b>A</b>	<b>Artigos</b>	<b>47</b>
<b>B</b>	<b>Método do histograma</b>	<b>57</b>

# Lista de Figuras

- 1.1 (a) Representação estrutural do estado nativo da proteína Ubiquitina. A informação sobre a estrutura é obtida pelo PDB - *Protein Data Bank* ([www.rcsb.org](http://www.rcsb.org)) com o código de identificação (PDB ID) 1UBQ. As Letras N e C indicam as terminações da proteína com início no N-terminal e final no C-terminal. As cores indicam os diferentes tipos de estruturas secundárias. Em laranja é representada a estrutura de hélice  $\alpha$ , em verde a estrutura de folhas  $\beta$ , em azul as voltas ou *loops* e em cinza claro as regiões sem estruturas definidas, ditas como aleatórias ou *random coil*.  
(b) Representação da estrutura primária da proteína Ubiquitina. As cores e as figuras geométricas indicam a representação unidimensional das estruturas secundárias. As diferentes letras indicam os diferentes tipos de resíduos de aminoácidos que compõem essa proteína. . . . . 2
- 1.2 (a) Ilustração do funil de estruturas do processo de enovelamento de uma proteína ao longo da coordenada de reação  $Q$  (Figura adaptada de [1, 2]). (b) Representação do perfil de energia livre  $F(Q)$  em função da coordenada de reação  $Q$ . É destacado a existência do estado desenovelado  $D$ , do estado nativo  $N$  e da barreira de energia  $\Delta F$  existente entre esses dois estado. Esta curva foi obtida por meio da simulação computacional do modelo  $C_\alpha$  [3] (descrito com mais detalhes na seção de modelos baseados em estrutura) para a proteína *CI2 - Chymotrypsin Inhibitor 2* PDB ID: 1YPA [4]. . . . . 7

## Lista de Figuras

---

- 2.1 Exemplo de duas estruturas ( $k$  e  $l$ ) onde são indicadas as distâncias internas ( $r_{i,j}^{k(l)}$ ) entre os aminoácidos (1, 7, 13, 33 e 60). As diferenças estruturais estão relacionadas com as variações destas distâncias entre todos os aminoácidos, de acordo com a Equação 2.2. . . . . . 15
- 2.2 Exemplo de uma trajetória da simulação da proteína SH3 em  $T_f$ , é notável que existe dois estados frequentemente visitados (com a energia  $E \approx 50$  e  $E \approx 250$ ), representando os estados enovelado e desenovelado, respectivamente. A linha vermelha marca o estado de transição, calculado posteriormente pelo método do WHAM. As linhas em azul representam os intervalos  $\Delta$ , onde é escolhido a estrutura de menor energia. . . . . . 16
- 2.3 Inicialização do sistema para a projeção em duas dimensões. Neste caso, a disposições das conformações foram baseada em energias. . . . . . 18
- 2.4 Estrutura cristalográfica das proteína de interesse: (a) *Chymotrypsin Inhibitor 2* Proteína CI2 (PDB: 1YPA); (b), Domínio SH3 (PDB: 1FMK) e; (c) Proteína A (PDB: 1BDD). . . . . . 19
- 3.1 Exemplo de visualização em 2D: No lado esquerdo da figura, tem-se a visualização da superfície de enovelamento da proteína *ww domain*, usada nesse trabalho somente para explicar a visualização em 2D. Cada ponto neste gráfico refere-se a uma conformação (ou cluster de conformações) gerada durante o processamento de dados (Seção 2.3). No lado direito, estão representadas quatro conformações (em vermelho), e suas respectivas distâncias multidimensionais ( $\delta^{k,l}$ , dado pela equação 2.3). Uma estrutura exemplo de cada uma delas está representada sobreposta com a estrutura nativa da proteína (em verde). . . . . . 22
- 3.2 Visualização em 2D da proteína *ww domain*: Ao colorir a projeção com o fração de contatos nativos, é possível definir regiões na projeção. O estado nativo (em vermelho) forma o principal cluster, e em geral, localiza-se no centro da projeção. Os estados desenovelados (em azul) se apresentam ao redor desse cluster principal. . . . . . 23

3.3	Nesta figura são apresentadas três projeções em 2D da proteína <i>ww domain</i> . Para cada projeção foi utilizado um conjunto de conformações distintas. Em (a) foram utilizadas $2,0 \times 10^4$ estruturas, em (b) foram utilizados $1,0 \times 10^4$ estruturas e em (c) são representadas $1,0 \times 10^3$ . A escala de cor representa o número de contatos nativos feitos em cada conformação. . . . .	24
3.4	Nesta figura é apresentada as projeções em 2D para seguintes proteínas: (a) CI2; (b) SH3; e (c) Proteína A. Cada gráfico é colorido pela coordenada $Q$ (número de contatos nativos). Em detalhe, é exibido um exemplo de uma estrutura nativa e uma estrutura desenovelada para cada proteína. . . . .	25
3.5	Energia Livre vs Coordenada de Reação. Cada gráfico, (a), (b) e (c) representa a energia livre do enovelamento para as proteínas CI2, SH3 e Proteína A, respectivamente. A linha tracejada em vermelho é a energia livre calculada pelo método WHAM [5]. A linha em preto é a energia livre calculada usando a função de distribuição radial (FDR). A FDR é calculada usando a estrutura nativa como referência. . . . .	27
3.6	(a) Projeção em 2D da proteína SH3 colorida pela coordenada de reação $Q_{path}$ . Esta coordenada representa os contatos necessários para a proteína se enovelar através da rota dominante. Seus valores variam desde valores negativos (enovelamento reverso) até valores positivos (enovelamento nativo) [6]. (b) Gráfico de curva de nível da superfície de energia livre, utilizando a projeção em 2D como referência. Neste caso, a cor representa os valores calculados pela equação 3.1. Os três pontos em destaque (A, B e C) representam as regiões onde possui rotas preferenciais para o enovelamento. . . . .	28

## Lista de Figuras

---

- 3.7 Análise do mapa de contato das rotas do enovelamento da proteína SH3. (a) Representa o mapa de contato da proteína SH3 colorido pela fração de contatos formados no estado de transição. Dois grupos são realçados: *Early group* (em verde) representando os contatos mais prováveis (entre as folhas  $\beta_b$ ,  $\beta_c$  e  $\beta_d$ ) e o *late group* (em vermelho), apresentando os contatos raramente formado no estado de transição (entre as folhas  $\beta_c$  e  $\beta_d$ ). Os gráficos (b), (c) e (d) apresentam conformações representativas de cada rota marcadas na figura 3.6 (b). Note que os contatos feitos em (b) (entre as folhas  $\beta_c$  e  $\beta_d$ ) diferem dos contatos feitos em (c) e em (d) (muito similar aos contatos feitos no *early group*. . . . . 30
- 3.8 Comparação entre os potenciais de Lennard-Jones e Gaussiano: As curvas em preto representam os potencial de Lennard-Jones, com mínimos em 6 e 10 Å. As curvas em vermelho representam os potenciais gaussianos, enquanto que a linha tracejada representa o modelo *Multi-basin*. . . . . 32
- 3.9 Análise das imagens espelhadas da proteína A. (a) Visualização em 2D da Proteína A proveniente de simulações usando o modelo *SBM multi-basin* com potencial gaussiano. Além disso, utilizou-se as mesmas configurações encontrada no artigo de Noel et. al. [7]. A cor representa a coordenada de reação  $Q_N - Q_M$ , onde os pontos em vermelho referem-se a formação de contatos nativos e os pontos em azul representam a formação dos contatos da estrutura espelhadas (*mirror state*). (b) A curva de nível que representa a densidade de estados calculada pelo histograma em 2D. Três regiões são realçadas: (A) Estruturas nativas, (B) Estrutura quase-nativas (*native-like*) e, (C) estruturas espelhadas. As linhas tracejadas representam as probabilidades da conversão entre esse estados e o estado desenovelado. . . . . 32

4.1 Visualização do enovelamento de proteínas em 3D. Nesta figura é apresentado uma visão geral dos resultados apresentados neste trabalho, a projeção em 2D é utilizado como plano XY, o eixo Z representa a energia e a cor é definida pela coordenada de reação apropriada para cada proteína (Essas coordenadas são discutidas em suas respectivas seções). (a) Os resultados obtidos para o domínio SH3. (b) os resultados obtidos para Proteína A. . . . . 36





# Lista de Tabelas

3.1	Configuração da simulação para a proteína WW Domain . . . . .	21
3.2	Configuração da simulação para as proteínas CI2, domínio SH3 e Proteína A . . . . .	25



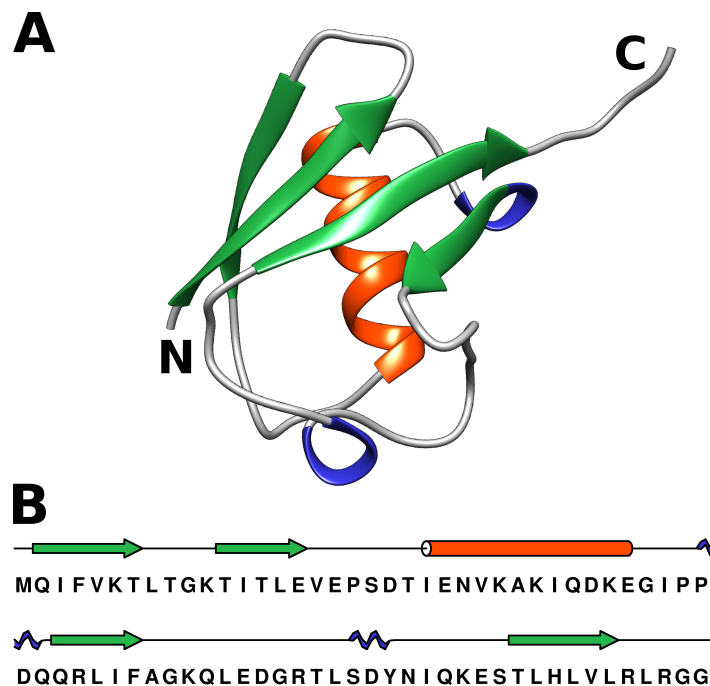
# Capítulo 1

## Introdução

### 1.1 Aspectos gerais de proteínas

A compreensão dos mecanismos biomoleculares tem proporcionado um grande desafio para os cientistas de diversas áreas do conhecimento. Uma das classes de macromoléculas biológicas de grande importância para os seres vivos são as proteínas. Elas constituem a maior parte da massa celular seca e são responsáveis por inúmeras funções em meio fisiológico. A construção celular, a regulação de entrada e de saída de íons, o transporte e a recepção de sinais, são exemplos de algumas funções realizadas pelas proteínas [8]. Algumas doenças estão diretamente relacionadas com o mal enovelamento das proteínas e com a falha dos mecanismos de controle de qualidade da célula [9], como o mal de Alzheimer, a doença de Huntington, a encefalopatia espongiforme (doença da vaca louca), a diabetes tipo II, entre outras. A grande importância das proteínas para a maquinaria biológica e para a fisiologia dos organismos vivos torna esta classe de macromoléculas um alvo interessante de estudo.

De uma maneira simplificada, a proteína pode ser representada por um heteropolímero linear em que cada monômero de sua cadeia é constituído por um resíduo de aminoácido. Em geral, existem 20 diferentes tipos de aminoácidos que podem compor a cadeia proteica. Os aminoácidos possuem um carbono na cadeia principal denominado carbono alfa ( $C_\alpha$ ), que realiza quatro ligações covalentes: *i*) com um átomo de hidrogênio; *ii*) com um grupo amina; *iii*) com um grupo carboxila e; *iv*) com um grupo radical também conhecido como a cadeia lateral do aminoácido. O grupo



**Figura 1.1** (a) Representação estrutural do estado nativo da proteína Ubiquitina. A informação sobre a estrutura é obtida pelo PDB - *Protein Data Bank* ([www.rcsb.org](http://www.rcsb.org)) com o código de identificação (PDB ID) 1UBQ. As Letras N e C indicam as terminações da proteína com início no N-terminal e final no C-terminal. As cores indicam os diferentes tipos de estruturas secundárias. Em laranja é representada a estrutura de hélice  $\alpha$ , em verde a estrutura de folhas  $\beta$ , em azul as voltas ou *loops* e em cinza claro as regiões sem estruturas definidas, ditas como aleatórias ou *random coil*. (b) Representação da estrutura primária da proteína Ubiquitina. As cores e as figuras geométricas indicam a representação unidimensional das estruturas secundárias. As diferentes letras indicam os diferentes tipos de resíduos de aminoácidos que compõem essa proteína.

radical diferencia todos os aminoácidos encontrados na natureza quanto à estrutura, tamanho, carga elétrica e hidrofobicidade. Cada aminoácido é ligado ao seu vizinho por uma ligação covalente e uma sequência com combinações distintas dos diferentes aminoácidos formam cada proteína [10, 11].

Existem quatro níveis de organização que descrevem estruturalmente as proteínas. A estrutura primária consiste na representação linear da sequência de resíduos de aminoácidos representada pelas diferentes letras na Figura 1.1 (b). A estrutura secundária se refere às estruturações locais da proteína, como por exemplo, hélice  $\alpha$  e folhas  $\beta$  (em laranja e em verde, respectivamente, na Figura 1.1 (a)). A estrutura terciária representa o arranjo tridimensional de todos os átomos em suas estruturas

secundárias, como representado na Figura 1.1 (a). A estrutura quaternária é o arranjo espacial de subunidades de uma proteína, formando um complexo proteico [8, 11].

Cada proteína possui uma sequência unidimensional particular de resíduos de aminoácidos. O processo em que esta cadeia linear encontra seu estado nativo é chamado de enovelamento ou dobramento de proteínas. Os estudos sobre o problema do enovelamento procuram entender como a informação presente na sequência unidimensional de resíduos de aminoácidos faz com que a proteína encontre seu estado nativo tridimensional e funcional. A compreensão das interações envolvidas e dos princípios físico-químicos presentes no processo de enovelamento é um grande desafio da ciência atual e envolve pesquisadores de diversas áreas do conhecimento.

## 1.2 O enovelamento de proteínas

Os estudos de Anfinsen [12] na década de 60 tiveram um papel fundamental para o desenvolvimento do estudo do enovelamento proteico e consistiam em medir, por meio de reações físico-químicas, a atividade de uma proteína, no caso, a ribonuclease. Anfinsen mostrou que uma proteína depois de desnaturada (desenovelada), poderia espontaneamente se enovelar, ou seja, restabelecer sua forma nativa em condições fisiológicas adequadas, reativando sua função biológica. Anfinsen então concluiu que a sequência de aminoácidos possuía toda a informação necessária para definir sua estrutura tridimensional e, portanto, sua função biológica. Também concluiu, com esses experimentos, que o processo de desnaturação e renaturação são processos reversíveis. Com seus estudos sobre a desnaturação da ribonuclease, Anfinsen estabeleceu o que foi chamado de hipótese termodinâmica. Essa hipótese diz que a estrutura de uma proteína, no seu estado nativo e em condições fisiológicas normais, é o estado energético mais baixo de todo o sistema.

Os resultados de Anfinsen foram o ponto de partida para um novo problema, no qual ainda hoje não há uma resposta completa: como determinar a estrutura tridimensional de uma proteína partindo apenas do conhecimento da sequência de aminoácidos que constitui sua estrutura primária? É claro que este problema pode ser tomado como um corolário da questão fundamental, que consiste em compreender

## Introdução

---

os mecanismos envolvidos neste processo biológico e desvendar quais as forças que governam o enovelamento.

Em 1968, Cyrus Levinthal [13] argumentou sobre qual seria a forma (cinética) em que a proteína encontra o seu estado nativo. Por meio de um exemplo, Levinthal mostrou que esta procura não poderia ser de forma aleatória. O argumento dele foi o seguinte: considere uma pequena proteína com 100 aminoácidos e suponha que cada aminoácido pode acessar apenas dois estados (por exemplo, só pode tomar duas orientações diferentes). Nestas condições, a proteína teria acesso a um total de  $2^{100} \approx 10^{30}$  conformações, dentre as quais, obviamente, a estrutura nativa incluída.

Como a proteína não pode passar de uma conformação para outra em menos de um picossegundo ( $ps$ ), que é o tempo de uma vibração térmica, seriam necessários  $2^{100} ps$ , ou seja,  $3,9 \times 10^{10}$  anos, no mínimo, para explorar exaustivamente todo o espaço conformacional e encontrar a conformação correspondente ao estado nativo. Como se pode constatar, esta escala de tempo é da ordem de grandeza da idade do universo, estimada em  $1,4 \times 10^{10}$  anos. Depara-se, dessa forma, com um problema, uma vez que o enovelamento de proteínas deste tamanho acontece em, no máximo, alguns segundos e tipicamente ocorre na escala temporal de microssegundos. A conclusão que se chega é que a hipótese termodinâmica não consegue explicar a escala de tempo característica do processo de enovelamento de proteínas. Por razões óbvias, este problema ficou conhecido como paradoxo de Levinthal e foi o próprio Levinthal o primeiro a sugerir uma solução para esse problema.

Levinthal teorizou a existência de um caminho específico, composto por estados intermediários – de forma semelhante ao do que ocorre numa reação química comum – no fim do qual se encontra o estado nativo. Como consequência da proposta de Levinthal, até ao início da década de 90, a investigação experimental sobre o enovelamento de proteínas foi, em grande parte, dominada pela procura de estados intermediários suficientemente estáveis para que pudessem ser isolados e devidamente caracterizados. No entanto, rapidamente foi mostrado que a existência de estados intermediários estáveis não é um requisito essencial para a rapidez do processo de enovelamento. A perspectiva clássica do enovelamento de proteínas se baseia na dicotomia termodinâmica *versus* cinética, e na abordagem tradicional da bioquímica, que considera cada molécula um sistema único, sendo, por isso, necessária uma descrição detalhada, em escala atômica, do seu caminho de enovelamento.

Um ponto de vista alternativo foi introduzido no final da década de 80 [14, 15]. Nesta abordagem é sugerido que o principal fator para se entender o processo de enovelamento deve ser uma visão global sobre a superfície de energia da proteína. O enovelamento da proteína seria organizado por um conjunto de estruturas similares, ao invés da existência de um rota preferencial passando por estados intermediários. Desta forma, uma abordagem estatística do processo de enovelamento pode ser empregada para descrever a superfície de energia da proteína.

### 1.3 A teoria de superfície de energia

A teoria sobre a superfície de energia do enovelamento de proteína, introduzida por Onuchic, Wolynes e colaboradores [16, 2, 17, 18], procura, de maneira simplificada, descrever o processo de enovelamento e os princípios gerais que governam este mecanismo [19–23]. Ela se baseia na natureza estatística do processo de enovelamento [14, 15, 24], na qual se observa a superfície de energia global da proteína.

A superfície de energia tem caráter multidimensional e uma maneira simplificada de analisa-la é utilizando uma coordenada de reação. A coordenada de reação é necessária para acompanhar o processo de enovelamento ao longo da superfície de energia [25] e a sua escolha é fundamental para uma boa abordagem ao problema estudado [26]. Uma coordenada de reação amplamente utilizada é a fração de contatos nativos  $Q$ . Esta coordenada é a razão entre os contatos nativos formados em um conformação desenovelada  $D$  e os contatos nativos presentes na estrutura enovelada  $N$ . Desta forma, esta coordenada mede, ao longo do processo de enovelamento, o grau de similaridade entre uma estrutura desenovelada e o estado nativo. Os valores de  $Q$  variam entre 0 e 1, onde 0 significa que a proteína está totalmente desenovelada e 1 significa que a proteína está no seu estado nativo. A coordenada de reação  $Q$  é baseada na topologia do estado nativo da proteína e descreve de maneira satisfatória o processo de enovelamento [25].

Nesta perspectiva, a superfície de energia possui uma topografia afunilada com um gradiente de energia direcionado para a região do estado nativo da proteína [27–29] - Figura 1.2 (a).

Nesta abordagem, o mecanismo de enovelamento pode ser entendido como um processo de difusão ao longo da superfície de energia quando se acompanha a coordenada



## Introdução

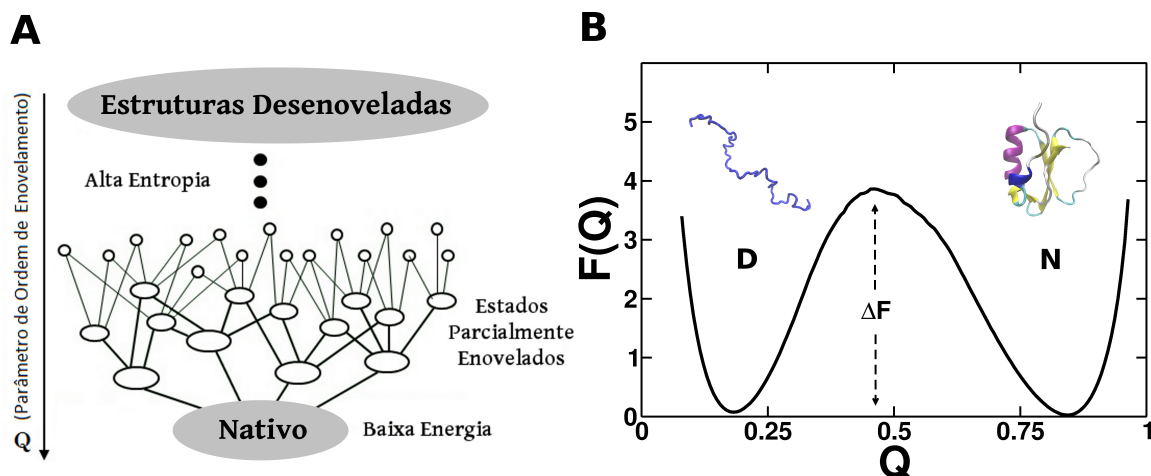
---

de reação [30–32]. A Figura 1.2 (a) apresenta a coordenada de reação  $Q$  como o parâmetro de ordem para se acompanhar o enovelamento. No topo estão descritas as estruturas desenoveladas e a alta entropia referente essa quantidade de estados. O conjunto de estruturas similares são agrupadas ao longo do processo de enovelamento, onde a proteína passa por estados parcialmente enovelados e quantidade de estruturas possíveis diminui [1, 25, 18, 33]. Este processo ocorre até que a proteína atinja seu mínimo energia na parte de baixo do funil de estruturas, atingindo o seu estado nativo. Pela descrição do funil de estruturas, não há um caminho preferencial no processo de enovelamento, mas uma multiplicidade de possíveis rotas [2].

Desta perspectiva, durante o processo de enovelamento, existe uma competição entre a contribuição energética  $E(Q)$  e a contribuição entrópica  $S(Q)$ , que são dependentes da coordenada de reação  $Q$ . O processo pode ser mapeado por um potencial efetivo que leve em consideração a competição entre essas duas grandezas. O potencial termodinâmico dado pela energia livre  $F(Q)$  é utilizado para descrever esta competição:

$$F(Q) = E(Q) - TS(Q). \quad (1.1)$$

O perfil de energia livre  $F(Q)$  é construído em função da coordenada de reação  $Q$  (Figura 1.2 (b)). A curva de energia livre destaca três regiões principais: o estado desenovelado  $D$ , o estado nativo  $N$  e a barreira de energia  $\Delta F$  que existe entre os dois estados. A temperatura  $T$  utilizada na construção da curva é a temperatura em que ocorre enovelamento. Nesta temperatura, conhecida como temperatura de enovelamento  $T_f$ , a probabilidade da proteína se encontrar no estado desenovelado é a mesma de encontra-la no estado nativo. Experimentalmente, a temperatura de enovelamento  $T_f$  é conhecida como a temperatura de *melting*  $T_m$ . Esta temperatura pode ser obtida por experimentos de calorimetria na construção de uma curva de desnaturação térmica ou de calor específico [34–36].



**Figura 1.2** (a) Ilustração do funil de estruturas do processo de enovelamento de uma proteína ao longo da coordenada de reação  $Q$  (Figura adaptada de [1, 2]). (b) Representação do perfil de energia livre  $F(Q)$  em função da coordenada de reação  $Q$ . É destacado a existência do estado desenovelado  $D$ , do estado nativo  $N$  e da barreira de energia  $\Delta F$  existente entre esses dois estado. Esta curva foi obtida por meio da simulação computacional do modelo  $C_\alpha$  [3] (descrito com mais detalhes na seção de modelos baseados em estrutura) para a proteína *CI2* - *Chymotrypsin Inhibitor 2* PDB ID: 1YPA [4].

## 1.4 Visualização do funil de enovelamento de proteínas

Os funis de enovelamento teorizados por Onichic, Wolynes e colaboradores apresentam características importantes do processo de enovelamento, sendo de grande interesse a sua visualização detalhada. No entanto, uma das grandes dificuldades encontrada para realizar a visualização desses funis foi a de que eles ocorrem em uma superfície multidimensional. A multidimensionalidade dos funis das proteínas, mesmo restringindo a representação e a visualização nas proximidades do seu estado nativo, impõe certas restrições, o que exige a busca por métodos que permitam a análise qualitativa dessas estrutura, permitindo, por meio dessa análise, estudar as características do enovelamento de proteínas. Muitos trabalhos vêm sendo desenvolvidos afim de caracterizar a superfície de energia. Técnicas como *PCA* (*Principal Component Analysis*) [37], *hierarchical model of networks* [38], *Markov State Models* [39–41] e *disconnectivity graphs* [42, 43] buscam caracterizar a superfície de energia e obter informações sobre o processo de enovelamento.

## Introdução

---

Em abordagens que vão além da representação em uma coordenada de reação, os mínimos locais são endereçados individualmente, possibilitando o uso de abordagens como, por exemplo, a Análise de Componentes Principais (*PCA*) [37, 44]. Nestes trabalhos, Becker e Karplus desenvolveram uma técnica chamada *disconnectivity graphs*, que tinha como finalidade visualizar a organização geral da superfície de energia. Esta visualização é representada em termos de mínimos locais de energia e seus respectivos estados de transição [23]. Esta técnica mostrou-se eficaz na visualização de superfícies de energia, com aplicação em vários estudos, como por exemplo, sistemas que apresentam características de “multi funil” [45–47], análise dos efeitos de *gatekeepers* [48] e a eficiência do enovelamento de proteínas com e sem frustração [49].

Estes trabalhos utilizaram, como base de suas análises, as informações cinética (mínimos locais e estados de transição) para a construção das superfícies de energia. Porém, para esses tipos de análises, não foi utilizado nenhum tipo de informação estrutural do sistema, o que constitui uma deficiência da técnica, uma vez que negligencia parte da informação da superfície de energia [50]. Smeeton et. al. apresentaram métodos para incluírem a informação estrutural juntamente com a informação cinética produzida por *disconnectivity graphs* [50]. Paralelamente a esse trabalho, foi desenvolvida uma metodologia, durante o mestrado do autor, que focou na organização estrutural das conformações, calculando as diferenças conformacionais entre elas. Neste trabalho, foi proposta uma métrica que refletia o comportamento das estruturas, permitindo o seu mapeamento em duas e em três dimensões. Dos resultados obtidos, foi observado que proteínas distintas apresentaram diferentes padrões de superfície, possibilitando a identificação não só de rotas de enovelamento, mas também de efeitos de mutações pontuais. Esses dados foram obtidos por meio de simulações de enovelamento de proteínas em modelo de rede cúbica.

Durante o primeiro ano de doutorado do autor, realizou-se as últimas simulações em modelo de rede, alguns ajustes no manuscrito para a finalização do texto, que posteriormente foi publicado [51]. Todos detalhes deste trabalho, pode ser lido no Apêndice A. deste trabalho a finalização no trabalho que culminou no focou-se em [51].

Neste trabalho de doutorado, é apresentada a atualização e generalização da metodologia de visualização do enovelamento de proteínas, desenvolvido durante o mestrado do autor. Esta atualização permite estender a análise estrutural para modelos mais complexos como, por exemplo, o modelo baseado em estrutura e o

## 1.4 Visualização do funil de enovelamento de proteínas

---

modelo com todos os átomos. Aqui será discutido o resultado para três proteínas que são representantes de classes básicas de estruturas (*motifs*). A primeira delas é formada somente por  $\alpha$ -hélices (Proteína A); a segunda formada somente por folhas- $\beta$  (Domínio SH3) e a terceira possui um misto de  $\alpha$ -hélice e folhas- $\beta$  (Proteína CI2). Serão discutidos também quais informação que se pode obter a partir da visualização do enovelamento proteínas. Essas informações são usadas para acrescentar detalhes sobre o enovelamento, como por exemplo, a análise de rotas dominantes e a competição entre estados energeticamente estáveis.



# Capítulo 2

## Metodologia

### 2.1 Modelo baseado em estrutura

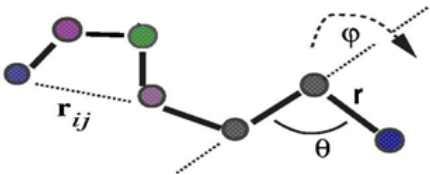
Simulações computacionais que utilizam modelos minimalistas permitem explorar uma grande faixa de parâmetros dos sistemas a um custo computacional relativamente baixo. Esse tipo de simulação tem se mostrado eficiente no estudo e no entendimento do enovelamento de proteínas [52–56]. A utilização de modelos mais realísticos envolve uma infinidade de parâmetros e de interações atômicas, fazendo com que o custo computacional necessário para extrair as informações importantes do sistema aumente consideravelmente.

Esse tipo de modelagem requer algumas simplificações no que se refere ao potencial de interação entre os componentes do sistema e aos seus próprios componentes. No que tange à simplificação dos componentes do sistema, pode-se utilizar, por exemplo, o modelo com todos os átomos ou o modelo somente com carbonos alfa ( $C_\alpha$ ). A simplificação utilizada no modelo  $C_\alpha$  consiste na substituição de todos os resíduos de aminoácidos da cadeia polipeptídica por esferas centradas na posição dos carbonos alfa ( $C_\alpha$ ) correspondentes, mantendo-se um volume de exclusão. Desse modo, apenas os carbonos alfa da cadeia principal são representados explicitamente. O modelo  $C_\alpha$  é amplamente utilizado e testado pela comunidade científica, permitindo uma diminuição no tempo computacional de até uma ordem de grandeza quando se comparado com o modelo com todos os átomos [57, 58].

## Metodologia

A simplificação no potencial de interação do modelo baseado em estrutura, o qual foi utilizado neste trabalho, é construído a partir da estrutura nativa da proteína, que é obtido por meio de banco de dados de proteínas (PDB), que por sua vez, foram obtidas por técnicas experimentais como cristalografia, difração de raio-x, ressonância magnética, entre outras.

Os modelos cujos potenciais são construídos a partir da estrutura nativa são conhecidos como modelos baseado em estruturas. Em especial, quando qualquer contato favorece a formação do estado nativo, esse modelo pode ser chamado de  $G\bar{o}$  [59]. Desse modo, o potencial de uma dada configuração  $\Gamma$  da proteína, sendo  $\Gamma_0$  sua configuração no estado nativo, será dado pela expressão:

$$\begin{aligned}
 V(\Gamma, \Gamma_0) = & \sum_{bonds} \epsilon_r (r - r_o)^2 \\
 & + \sum_{angles} \epsilon_\theta (\theta - \theta_o)^2 \\
 & + \sum_{dihedrals} \epsilon_\phi \left\{ [1 - \cos(\phi - \phi_o)] + \frac{1}{2} [1 - \cos(3(\phi - \phi_o))] \right\} \\
 & + \sum_{contacts} \epsilon_C \left[ 5 \left( \frac{d_{ij}}{r_{ij}} \right)^{12} - 6 \left( \frac{d_{ij}}{r_{ij}} \right)^{10} \right] + \sum_{non-contacts} \epsilon_{NC} \left( \frac{\sigma_{NC}}{r_{ij}} \right)^{12} \quad (2.1)
 \end{aligned}$$


onde  $\epsilon_r = 100$ ,  $\epsilon_\theta = 20$ ,  $\epsilon_\phi = \epsilon_0$ ,  $\epsilon_C = \epsilon_0$ ,  $\epsilon_{NC} = \epsilon_0$  e  $\epsilon_0$  é a energia de interação entre os contatos, com valor igual a 1 em unidades reduzidas. Na Equação 2.1 o parâmetro  $r_o$  representa a distância nativa entre dois átomos diretamente ligados;  $\theta_o$  é o ângulo entre três átomos consecutivos da estrutura nativa e;  $\phi_o$  o ângulo diedral entre quatro átomos consecutivos. O termo  $d_{ij}$  de Van der Waals é obtido por meio do mapa de contato utilizando o algoritmo *Contact of Structural Units* (CSU) [60]. Para impedir a sobreposição dos componentes da proteína, utiliza-se um termo repulsivo (*non-contacts*), no qual  $\sigma_{NC} = 4\text{\AA}$  (distância de máxima aproximação). Com esses parâmetros, que são obtidos a partir da estrutura nativa depositada no PDB, o mínimo global de energia da macromolécula ocorre quando ela atinge o seu estado nativo. Portanto, independente da estrutura da proteína no início da simulação, tanto aberta quanto fechada, se obtém, ao final da simulação, a estrutura enovelada e em seu estado nativo.

Nas simulações realizadas para este trabalho, foi utilizado o software Gromacs [61], em sua versão 4.6.5. Os arquivos necessários para a simulação foram gerados por meio da ferramenta online *SMOG* [62]. Foram realizadas duas configurações de simulações, cada qual com o seu conjunto de configurações. A primeira delas é usada para encontrar a temperatura ótima de enovelamento ( $T_f$ ). Portanto, são realizadas várias simulações curtas ( $5,0 \times 10^8$  frames) em um intervalo de temperaturas. A  $T_f$  é definida como a temperatura em que a probabilidade de encontrar a proteína enovelada e desenovelada é a mesma ( $P_{fold} = P_{unfold} = 0,5$ ). Outra forma de encontrar  $T_f$  é calculando o calor específico em função da temperatura. Esse cálculo pode ser realizado utilizando a ferramenta *eSBMTools* [63]. Neste caso,  $T_f$  é definido com a temperatura onde tem-se o maior valor do calor específico.

A segunda configuração de simulação é realizada para se obter o maior número de conformações possíveis. Portanto, realiza-se uma simulação longa ( $5,0 \times 10^9$  frames) na temperatura de enovelamento ( $T_f$ ). Os resultados dessa simulação serão utilizados nas próximas etapas, que serão explicada nas próximas seções.

## 2.2 Matriz de dissimilaridades

Com os resultados das simulações do enovelamento de proteínas descritas na seção 2.1, são obtidos os seguintes arquivos: *trajetoria.xtc* e *energia.edr* (ambos formato do Gromacs). Através do arquivo de trajetória, é possível obter a coordenada de reação  $Q$  em função do tempo (um exemplo por de ser visto na Figura 2.2).

O primeiro passo para a construção da visualização é definir uma métrica que descreva de maneira robusta a sua informação. É necessário que a métrica leve em conta características importantes dos dados em questão. Além disso, devido às diferenças nessas características, faz-se necessário atribuir um valor que distinga tais informações.

Uma característica relevante que se tem no enovelamento proteico é a informação estrutural (posição de cada aminoácido da proteína em função do tempo). As posições dos aminoácidos são essenciais para a elaboração da métrica, pois diferentes combinações de posição resultam em diferentes estruturas e, conseqüentemente, em diferentes



## Metodologia

---

conformações. De maneira formal, a medida de comparação entre duas conformações  $(k, l)$  pode ser escrita como:

$$q^{k,l} = \frac{1}{N} \sum_N \exp \frac{-(r_{i,j}^k - r_{i,j}^l)^2}{\sigma_{i,j}^2}, \quad (2.2)$$

onde  $r_{i,j}^{k(l)}$  é a distância entre o aminoácido  $i$  e  $j$  da conformação  $k(l)$  e  $N$  é o fator de normalização para que o valor de  $q^{k,l}$  seja um número entre 0 e 1 e  $\sigma_{i,j} = \sigma_0|i-j|^\epsilon$ , sendo  $\sigma_0$  e  $\epsilon$  valores ajustáveis. Essa equação foi apresentada primeiramente por Wolynes e colaboradores, onde foi utilizada para a predição da estrutura de proteínas [64]. Na Figura 2.1 é apresentada um esquema de como as distâncias são calculada para duas conformações diferentes (no caso,  $k$  e  $l$ ). Para encontrar o valor de dissimilaridade, basta utilizar a seguinte expressão:

$$\delta^{k,l} = 1 - q^{k,l}, \quad (2.3)$$

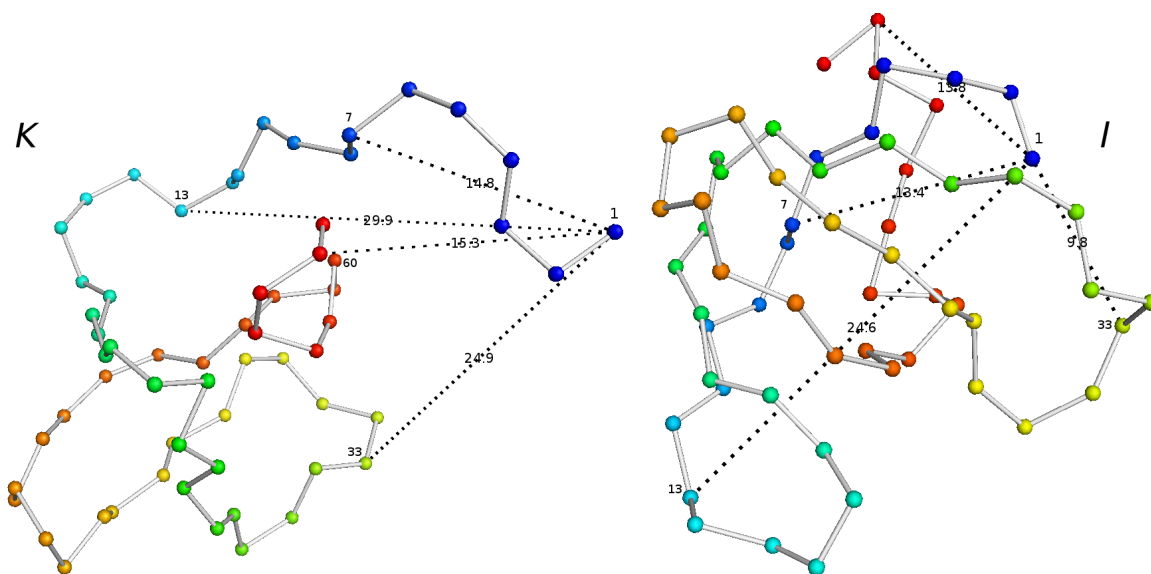
sendo  $\delta^{k,l}$  um valor compreendido entre 0 e 1, onde valores próximos de 0 significam que as duas conformações são muito similares. Em contrapartida, valores próximos a 1 significam que as estruturas são estruturalmente diferentes.

A medida de dissimilaridade é calculada para todos os pares de estruturas  $k$  e  $l$  de uma dada trajetória, resultando em uma matriz triangular  $n \times n$ , onde  $n$  é o número total de conformações analisadas.

## 2.3 Processamento dos dados

Após o cálculo da matriz de dissimilaridades, realiza-se o processamento dos dados. Devido ao tamanho da matriz (na ordem de  $5 \times 10^8$  estruturas), o cálculo da projeção multidimensional fica inviável, uma vez que esse tipo de tratamento tem um grande custo computacional por necessitar de uma grande quantidade de memória RAM. Para contornar esse problema, foram propostas duas soluções:

- i)* A primeira solução proposta foi criar um *ensemble* de estruturas, onde as conformações escolhidas para se fazer a projeção multidimensional são definidas como a estrutura que possui a menor energia (ou maior  $Q$ ) em um determinado intervalo



**Figura 2.1** Exemplo de duas estruturas ( $k$  e  $l$ ) onde são indicadas as distâncias internas ( $r_{i,j}^{k(l)}$ ) entre os aminoácidos (1, 7, 13, 33 e 60). As diferenças estruturais estão relacionadas com as variações destas distâncias entre todos os aminoácidos, de acordo com a Equação 2.2.

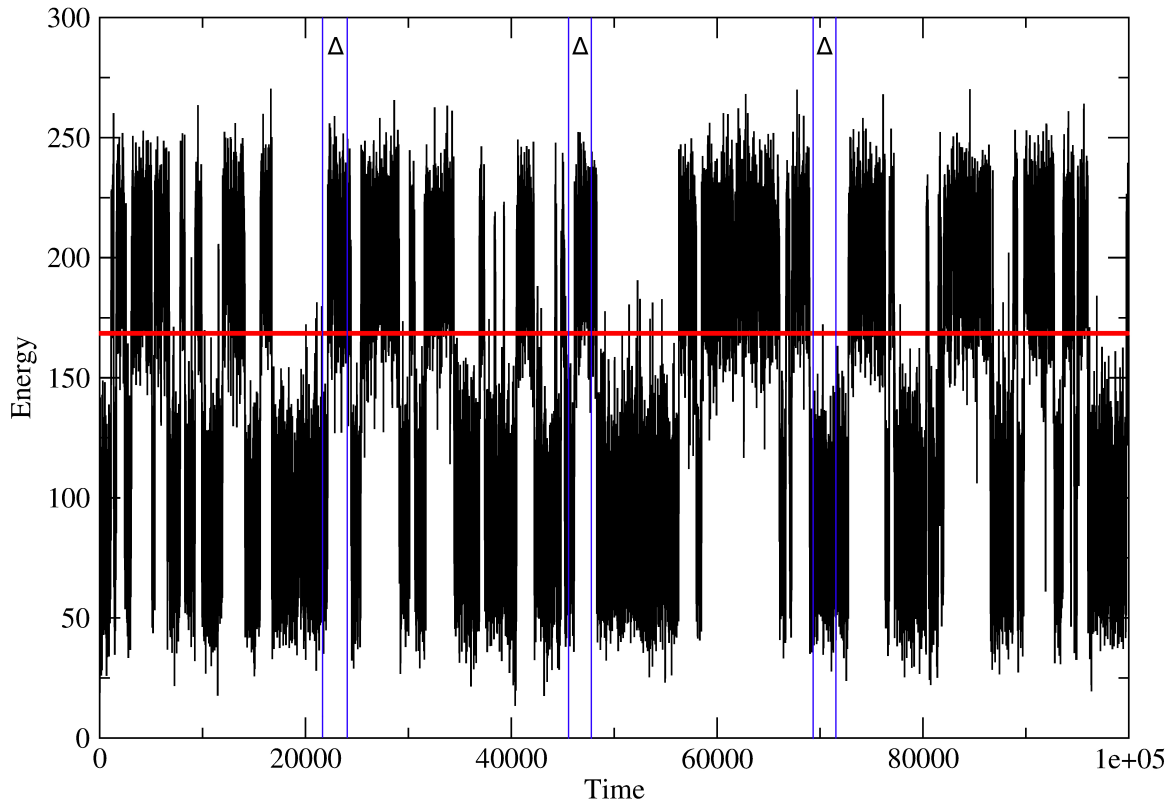
$\Delta$  (como pode ser visto na Figura 2.2). O valor de  $\Delta$  é ajustado de forma a se obter o número desejado de estruturas<sup>1</sup>. O ponto negativo desse método é que não há garantia que o *ensemble* de estruturas escolhidas represente todo o espaço conformacional visitado durante a simulação.

ii) Para contornar essa questão e maximizar a amostragem do espaço conformacional, foi desenvolvido um método para realizar uma clusterização de todas as estruturas obtidas na simulação. Dessa forma, garante-se que todo o espaço de fase visitado durante a simulação seja utilizado para gerar a visualização da superfície de energia. Essa clusterização é explicada na seção seguinte.

### 2.3.1 Clusterização multidimensional

Ao construir a matriz de dissimilaridade é possível observar uma grande quantidade de estruturas semelhantes ( $\delta^{k,l} < 0,2$ ). Com o intuito de diminuir o tamanho da matriz, é realizada uma clusterização, utilizando-se como parâmetro o valor de  $\delta^{k,l}$ . O algoritmo para esse cálculo é apresentado abaixo:

<sup>1</sup>No Capítulo de resultados são apresentado algumas características devido à escolha do valor de  $\Delta$ .



**Figura 2.2** Exemplo de uma trajetória da simulação da proteína SH3 em  $T_f$ , é notável que existe dois estados frequentemente visitados (com a energia  $E \approx 50$  e  $E \approx 250$ ), representando os estados enovelado e desenovelado, respectivamente. A linha vermelha marca o estado de transição, calculado posteriormente pelo método do WHAM. As linhas em azul representam os intervalos  $\Delta$ , onde é escolhido a estrutura de menor energia.

1. Ler a matriz de dissimilaridades;
2. Define-se o valor de *cutoff* e o número final de estruturas;
3. Para cada conformação  $k$  e  $l$ , verifica se  $\delta^{k,l} < \textit{cutoff}$ :
  - i*) Se sim, clusteriza o par  $(k, l)$ , transformando-o em uma única conformação (*nodo*);
  - ii*) Se  $\delta^{k,l}$  for maior que o *cutoff* para todas conformações  $k$ , cria-se um novo nodo com a conformação  $k$ ;
4. Verifica-se o número de nodos finais:
  - i*) Se o número de nodos for maior do que o número de estruturas desejada, aumenta-se o *cutoff* e repete-se o item 3;

- ii) Se o número de nodos for menor que o número de estruturas desejadas, é definido que houve convergência;
- 5. Escreve a matriz de dissimilaridades clusterizada e um arquivo contendo o número de estruturas em cada nodo e suas respectivas identificações (ID).

## 2.4 Projeção multidimensional

A etapa mais complexa para a criação deste modelo é a redução do espaço multidimensional original dos dados para um espaço dimensional compreensível para a percepção humana, isto é, um espaço visual de, no máximo, três dimensões. Idealmente, o modelo visual criado deve se assemelhar ao modelo visual proposto pela teoria do funil de enovelamento. Para isso, deve-se manter, ao máximo, a correspondência entre os dados antes e depois da redução dimensional. Técnicas de projeção multidimensional tratam essa questão, pois mapeiam os dados em espaços  $p$ -dimensionais, com  $p = \{1, 2, 3\}$ , preservando, ao máximo, a informação sobre as relações de distâncias ou dissimilaridade entre os dados [65].

Uma técnica de projeção multidimensional pode ser definida da seguinte forma: seja  $X$  um conjunto de objetos  $R^m$  com  $\delta : R^m \times R^m \rightarrow R$  um critério de proximidade entre dois objetos em  $R^m$ , e  $Y$  um conjunto de objetos em  $R^p$  para  $p = 1, 2, 3$  e  $d : R^p \times R^p \rightarrow R$  um critério de proximidade em  $R^p$ . Uma técnica de projeção multidimensional pode ser descrita como uma função  $f : X \rightarrow Y$  que visa tornar  $|\delta(x_i, x_j) - d(x_i, x_j)|$  o mais próximo possível de zero,  $\forall x_i, x_j \in X$ .

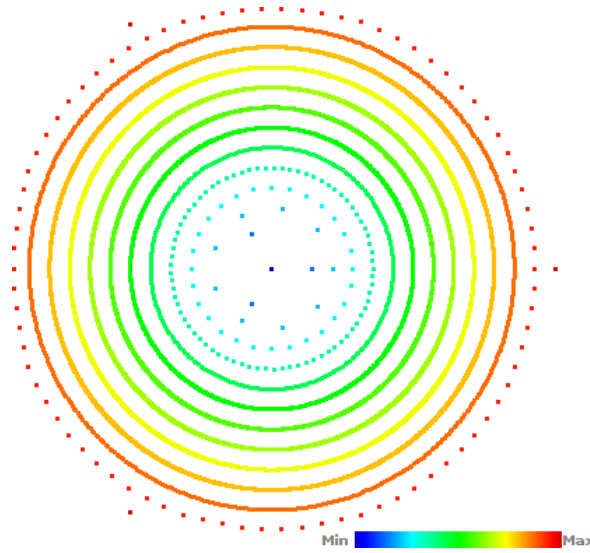
Em uma projeção bem construída, a proximidade dos pontos indica a semelhança entre os objetos que representam. Pontos próximos indicam instâncias semelhantes de acordo com a medida de distância  $\delta$ . Intuitivamente, pontos distantes representam objetos com pouca relação, também de acordo com  $\delta$ . Assim, a questão principal para a construção de uma boa projeção está diretamente relacionada com a forma com que as distâncias entre os objetos multidimensionais ( $\delta$ ) são calculadas. Dentre as várias técnicas de projeção que podem ser utilizadas, adotou-se, neste trabalho, a técnica *Force Scheme* [65], pois ela propõe um balanceamento entre precisão e desempenho computacional. Esta técnica estabelece um sistema de forças, onde, inicialmente, posicionam-se os objetos de forma aleatória ou por meio de alguma heurística e, em

## Metodologia

---

seguida, forças de atração e repulsão entre os objetos levam o sistema a um estado de equilíbrio.

Este trabalho estabelece uma inicialização do sistema baseada na energia das conformações. Este fato não interfere no resultado obtido, mas age no sentido de acelerar a convergência da técnica. O posicionamento inicial dos pontos pode ser visto na Figura 2.3, onde são definidos círculos de energia para cada valor de energia encontrado durante o processo de envelhecimento. As conformações são espalhadas uniformemente em seus respectivos círculos. Energias menores são representadas por círculos concêntricos dentro de círculos de maior energia.



**Figura 2.3** Inicialização do sistema para a projeção em duas dimensões. Neste caso, a disposição das conformações foram baseada em energias.

Após o posicionamento inicial dos pontos, a técnica *Force Scheme* realiza iterações para aproximar as distâncias entre os objetos projetados, a partir das distâncias  $\delta$  entre as conformações. Na primeira iteração, a técnica considera como conjunto de entrada,  $Y$ , a projeção definida na inicialização. Para cada ponto projetado  $y_i \in Y$ , calcula-se um vetor  $\vec{v}_{i,j} = (y_j - y_i)$ ,  $\forall y_j \neq y_i$ . Move-se, então,  $y_i$  na direção de  $\vec{v}$ , acrescentando-se a fração de  $\Delta$ , onde  $\Delta$  é dado por:

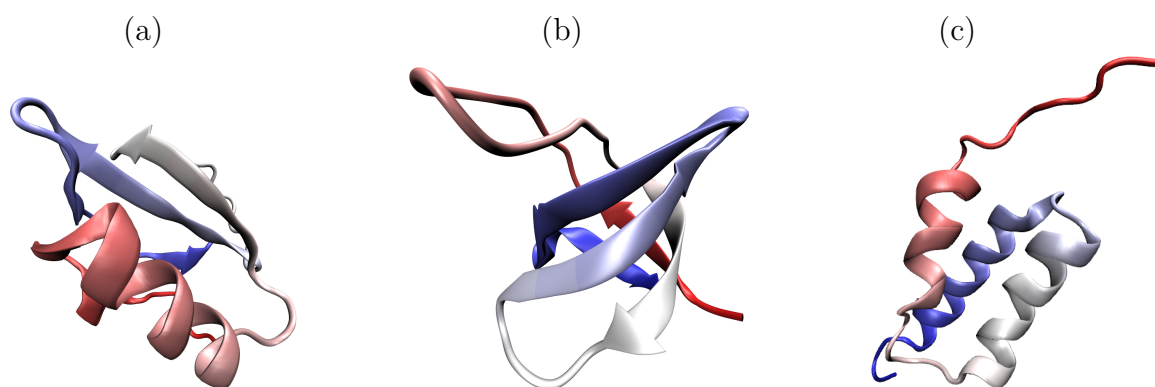
$$\Delta = \left\{ \left[ 1 + (2j + 1)^{\frac{1}{2j+1}} \right] \right\} - 1, \quad (2.4)$$

onde  $j$  é o número de iterações realizadas até um dado momento.

Ao término da iteração, cada objeto sofreu um deslocamento na direção de cada outro objeto, aproximando a distância entre eles com a distância calculada entre as conformações que eles representam. As iterações são repetidas sucessivamente até o sistema atingir o equilíbrio. O número de iterações pode ser definido arbitrariamente ou, então, pode-se parar o sistema através de algum critério específico. Aqui, estabeleceu-se como critério de parada a verificação de quanto o sistema está em um estado próximo ao equilíbrio, comparando iterações sucessivas, ou seja, é comparada a quantidade de deslocamentos dos objetos entre iterações subsequentes. Se esse valor for pequeno, abaixo de um limiar estabelecido (neste trabalho, adotou-se  $10^{-5}$  como limiar), o sistema é considerado em equilíbrio e a projeção final é obtida.

## 2.5 Proteínas estudadas

Como o objetivo principal deste trabalho é realizar uma análise conformacional da superfície de energia do enovelamento de proteínas, escolheu-se proteínas com *motifs* estruturais distintos entre-si. A primeira escolha foi a proteína CI2, formada por um misto de folhas- $\beta$  e  $\alpha$ -hélices; a segunda proteína analisada foi o domínio SH3, formada somente de folhas- $\beta$  e a terceira proteína escolhida é formada unicamente de  $\alpha$ -hélices, chamada de Proteína A. Abaixo será descrito cada uma delas com mais detalhes.



**Figura 2.4** Estrutura cristalográfica das proteína de interesse: (a) *Chymotrypsin Inhibitor 2* Proteína CI2 (PDB: 1YPA); (b), Domínio SH3 (PDB: 1FMK) e; (c) Proteína A (PDB: 1BDD).

A *Chymotrypsin Inhibitor 2* (CI2) é uma proteína de 64 resíduos, constituída por seis folhas- $\beta$  e por uma  $\alpha$ -hélice, formando um núcleo hidrofóbico (Figura 2.4 (a)).

## Metodologia

---

Estudos experimentais estabeleceram que o enovelamento da CI2 pode ser modelado pela cinética de dois estados [66, 67]. A estrutura do estado de transição para essa proteína tem sido extensivamente caracterizada, mostrando que cerca de metade das interações nativas são formadas nessa posição [68]. Outros estudos demonstram que, devido à distribuição dos valores- $\Phi$  serem uniformes, a maioria dos contatos hidrofóbicos estão formados no *ensemble* do estado de transição [69].

O domínio SH3 é um fragmento de 57 resíduos da proteína quinase que se estende a partir do aminoácido T84 até o S140. Esta proteína possui cinco folhas- $\beta$  em um arranjo antiparalelo, formando um Barril  $\beta$  parcial (Figura 2.4 (b)). Medidas experimentais têm mostrado que o domínio SH3 tem um enovelamento rápido e com mecanismos de dois estados [68]. A análise dos valores- $\Phi$  revelaram que duas regiões desta proteína são altamente estruturadas e estão acopladas no estado de transição, enquanto outras regiões apresentam uma fraca ordenação no *ensemble* do estado de transição [70]. Estudos por simulação mostraram que esta proteína possui uma rota específica de enovelamento, determinada principalmente por sua geometria [71, 72, 6]

A proteína A é uma componente da parede celular da *Staphylococcus aureus*, que é encontrada ligada especificadamente na porção Fc da imunoglobulina G (IgG) de várias espécies de mamíferos. Esta proteína possui 60 resíduos e seu estado nativo é formado por três  $\alpha$ -hélices, onde cada hélice está conectadas por voltas (Figura 2.4 (c)). O cálculo da energia livre por primeiros princípios mostrou que existem três “passos termodinâmicos” importantes no seu enovelamento. O primeiro deles é a interação da região amino-terminal com o motif hélice-volta-hélice, o segundo é a interação da hélice I com a hélice II e finalmente o encaixe da hélice III no subdomínio da Hélice I-II. Também foi encontrado um pequeno estado metaestável na região em que a proteína possuía um volume igual a 1,5 do volume do estado nativo. [73]

# Capítulo 3

## Resultados e Discussões

### 3.1 Descrição da superfície em 2D

Nesta seção será apresentada uma descrição de como interpretar os gráficos produzidos ao realizar a projeção multidimensional. Seguindo a metodologia descrita no capítulo anterior, cada resultado será apresentado em duas partes: (1) Uma tabela contendo todas as informações utilizadas na visualização (a proteína utilizada, o tempo de simulação, a temperatura em unidades reduzidas, o tipo de processamento de dados (Delta ou Cluster) de acordo com a seção 2.3 e o número de estruturas projetadas); (2) Uma figura onde é apresentado um gráfico do tipo *scatter map* representando a projeção em 2D.

Para realizar a demonstração da projeção em 2D, utilizou-se, como modelo, a proteína *WW-Domain* (PDB: 2N8S). Esta proteína possui 39 aminoácidos, e tem sua estrutura formada por três folhas- $\beta$ , ocupando cerca de 30% de sua estrutura. As informações da visualização desta proteína estão na Tabela 3.1.

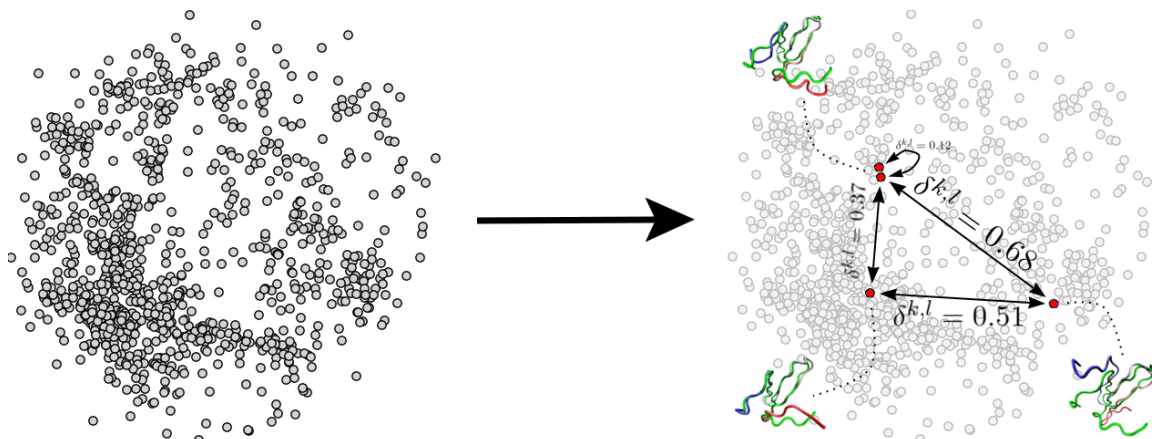
**Tabela 3.1** Configuração da simulação para a proteína WW Domain

Proteína	PDB	Tempo (ns)	$T_f$ <sup>1</sup>	PD <sup>2</sup>	Frames	Projeção
WW Domain	2N8S	250	112	Delta	$1 \times 10^5$	$1 \times 10^3$

<sup>1</sup> Temperatura de *folding* em unidades reduzidas;

<sup>2</sup> Tipo de processamento dos dados (Seção 2.3).

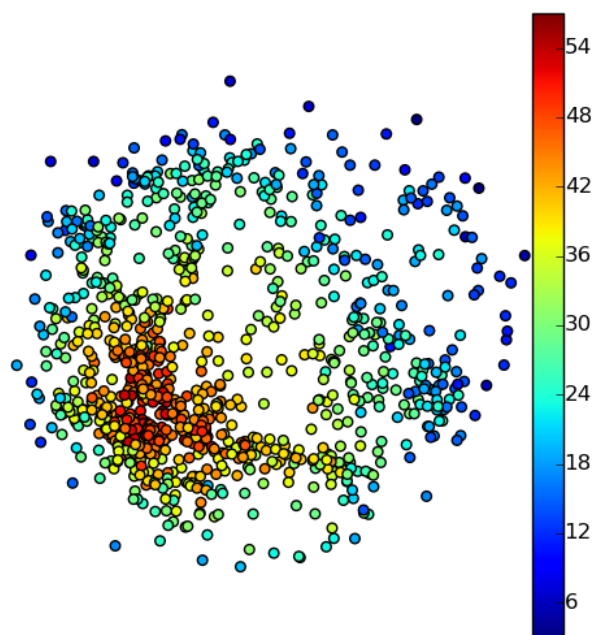




**Figura 3.1** Exemplo de visualização em 2D: No lado esquerdo da figura, tem-se a visualização da superfície de enovelamento da proteína *ww domain*, usada nesse trabalho somente para explicar a visualização em 2D. Cada ponto neste gráfico refere-se a uma conformação (ou cluster de conformações) gerada durante o processamento de dados (Seção 2.3). No lado direito, estão representadas quatro conformações (em vermelho), e suas respectivas distâncias multidimensionais ( $\delta^{k,l}$ , dado pela equação 2.3). Uma estrutura exemplo de cada uma delas está representada sobreposta com a estrutura nativa da proteína (em verde).

Na Figura 3.1 está representada a projeção em 2D da proteína *ww domain*. Nos gráficos de projeção que serão apresentados nesse trabalho, os eixos x e y serão removidos, uma vez que não apresentam nenhum significado físico. Neste tipo de projeção, cada ponto no gráfico representa uma conformação que a proteína visitou durante a simulação (ou cluster de conformações, dependendo do tipo de processamento de dados discutido na seção 2.3). A principal informação que pode-se obter desse tipo de gráfico é a distância euclidiana entre duas conformações quaisquer ( $k$  e  $l$ ). Esta distância ( $d^{k,l}$ ) é a projeção em 2D da distância multidimensional ( $\delta^{k,l}$ ) calculado pela equação 2.3, que reflete a dissimilaridade estrutural entre essas estruturas. Em outras palavras, pode se dizer que pontos próximos significam estruturas muito parecidas, ao passo que, pontos distantes representam estruturas completamente diferentes.

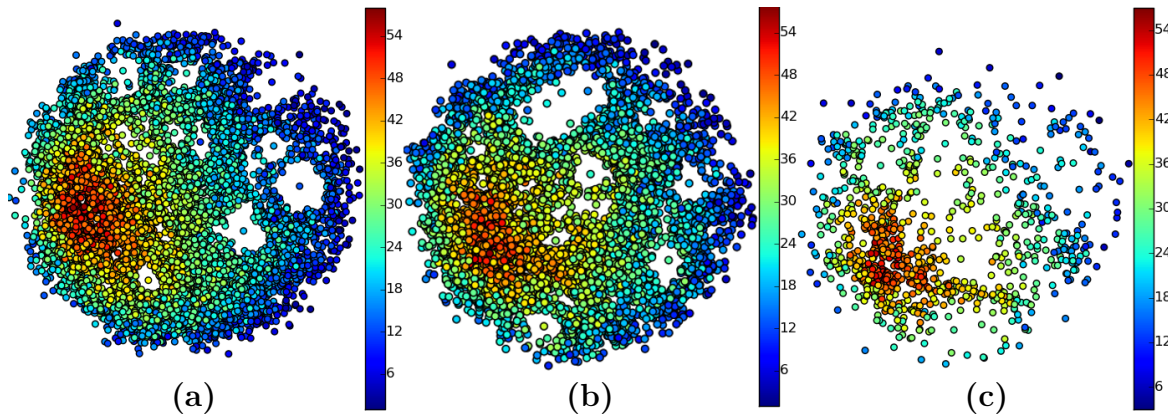
No lado direito da Figura 3.1, quatro estruturas foram realçadas em vermelho e as distâncias multidimensionais ( $\delta^{k,l}$ ) entre elas estão apresentadas sobre suas respectivas setas. Nota-se que, para o caso das duas estruturas muito próximas, o valor de  $\delta^{k,l}$  é bem pequeno (0,12), o que reflete em conformações estruturalmente muito parecidas. Nesta figura, também é apresentada a estrutura de cada conformação em comparação com a estrutura nativa da proteína (em verde). É possível observar clusters de conformações



**Figura 3.2** Visualização em 2D da proteína *ww domain*: Ao colorir a projeção com o fração de contatos nativos, é possível definir regiões na projeção. O estado nativo (em vermelho) forma o principal cluster, e em geral, localiza-se no centro da projeção. Os estados desenovelados (em azul) se apresentam ao redor desse cluster principal.

em várias regiões da projeção. Esses clusters significam que, durante a simulação, a proteína gastou um determinado tempo naquele estado (com estruturas similares). No exemplo citado, a proteína foi simulada na temperatura de folding, onde ela tem a probabilidade de 50% de se encontrar no estado nativo e 50% de encontrar-se no estado desenovelado, o que permite prever que o maior cluster provavelmente represente o estado nativo. Para verificar essa característica, calcula-se o valor de  $Q$  (fração de contatos nativos) para cada conformação, e assim, pode-se colorir a projeção de acordo com o grau de enovelamento. Esse resultado é apresentado na Figura 3.2.

Quando é atribuído o valor de  $Q$  para cada conformação, fica evidente o cluster do estado nativo (em vermelho), uma vez que as estruturas enoveladas possuem um alto grau de similaridade. Já no caso das estruturas parcialmente enoveladas, existe uma gama de possibilidades, que aumenta gradualmente conforme a proteína desenovela. Essas conformações se distribuem de modo que a ficarem ao redor do cluster nativo. Esse tipo de distribuição é consistente com as ideias associadas com a teoria de superfície de energia e do funil de enovelamento de proteínas.



**Figura 3.3** Nesta figura são apresentadas três projeções em 2D da proteína *ww domain*. Para cada projeção foi utilizado um conjunto de conformações distintas. Em (a) foram utilizadas  $2,0 \times 10^4$  estruturas, em (b) foram utilizados  $1,0 \times 10^4$  estruturas e em (c) são representadas  $1,0 \times 10^3$ . A escala de cor representa o número de contatos nativos feitos em cada conformação.

Outra característica que pode ser discutida nessa seção é a robustez da visualização em função do número de estruturas projetadas. No resultado apresentado anteriormente, utilizou-se  $1 \times 10^3$  conformações para realizar a projeção, o que equivale a 1% do total de estruturas visitadas durante a simulação. A pergunta levantada aqui é a seguinte: 1% do *ensemble* de conformações representa, de forma significativa, a superfície de enovelamento da proteína? Na Figura 3.3 são apresentados projeções para números diferentes de estruturas.

De acordo com a Figura 3.3, é possível perceber que a variação no número de estruturas não interfere no padrão final da visualização. A variância observada aparece na forma de um ajuste fino, ou seja, quanto maior o número de estruturas, mais detalhes podem ser observados. Porém, fatores importantes como rotas de enovelamento, estados metaestáveis, entre outros fatores, sempre são observados, mesmo para configurações com um número reduzido de estruturas.

No resultados a seguir, utilizou-se como modelo padrão para visualização, a configuração correspondente a 10% do total de estruturas simuladas. Tal escolha foi definida devido à dificuldade no processo de projeção multidimensional, sendo este o gargalo computacional<sup>1</sup>, mesmo utilizando uma técnica que permita trabalhar com um número alto de informações, que é o caso da técnica *Force Scheme* [65].

<sup>1</sup>Ver seção 2.3 e 2.4

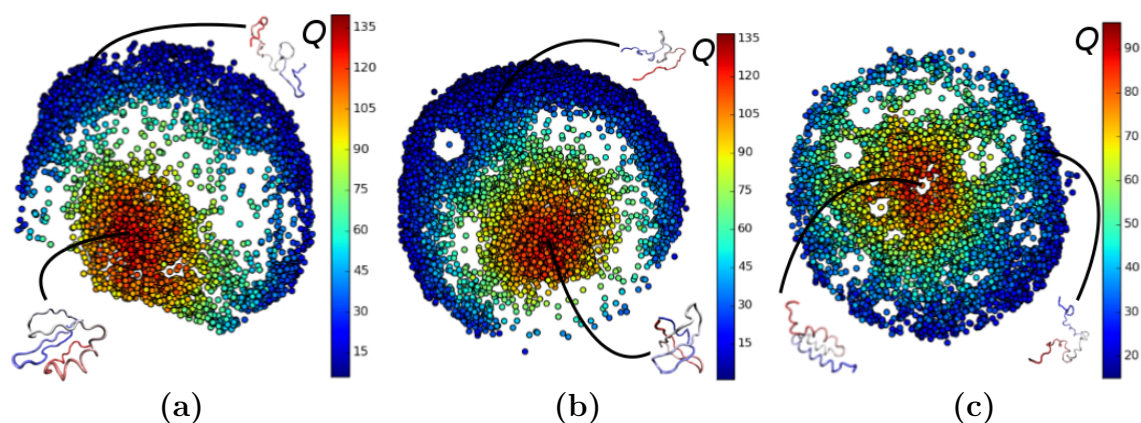
## 3.2 Visualização da proteína CI2, Domínio SH3 e Proteína A

**Tabela 3.2** Configuração da simulação para as proteínas CI2, domínio SH3 e Proteína A

Proteína	PDB	Tempo (ns)	$T_f$ <sup>1</sup>	PD <sup>2</sup>	Frames	Projeção
CI2	1YPA	500	125	Delta	$2 \times 10^5$	$2 \times 10^4$
SH3	1FMK	500	134	Delta	$2 \times 10^5$	$2 \times 10^4$
Proteína A	1BDD	500	115	Delta	$2 \times 10^5$	$2 \times 10^4$

<sup>1</sup> Temperatura de *folding* em unidades reduzidas;

<sup>2</sup> Tipo de processamento dos dados (Seção 2.3).



**Figura 3.4** Nesta figura é apresentada as projeções em 2D para seguintes proteínas: (a) CI2; (b) SH3; e (c) Proteína A. Cada gráfico é colorido pela coordenada  $Q$  (número de contatos nativos). Em detalhe, é exibido um exemplo de uma estrutura nativa e uma estrutura desenovelada para cada proteína.

## 3.2 Visualização da proteína CI2, Domínio SH3 e Proteína A

Nesta seção são apresentadas as visualizações para as proteínas de interesse deste trabalho. De acordo com a seção 2.5, essas proteínas possuem diferentes *motifs* estruturais, portanto, a primeira análise que pretende-se investigar é se existe variação no padrão de visualização para diferentes *motifs*. As visualizações são apresentadas na Figura 3.4.

A partir das visualizações, é possível extrair algumas informações estruturais do enovelamento dessas proteínas. Por exemplo, na Figura 3.4 (a), uma vez que a proteína está desenovelada (pontos azuis), existem duas rotas principais (pontos em verde, que representam os estados de transição) que levam ao estado nativo. Estas rotas

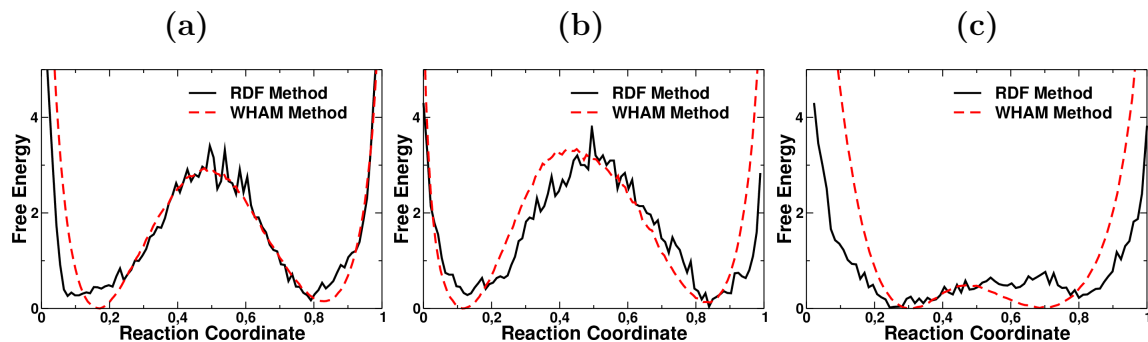
de enovelamentos estão relativamente distantes, significando que são estruturalmente diferentes entre si. Tal fato demonstra que a proteína CI2 possui múltiplos caminhos para atingir o estado nativo, o que está de acordo com trabalhos anteriores [58].

Já no caso da proteína A (Figura 3.4 (c)), é possível perceber que a distribuição da sua superfície de enovelamento é bem homogênea, quando comparada com as outras duas proteínas estudadas, ou seja, no caso da proteína A, as possibilidades (ou caminhos) de enovelamentos são diversos, sendo possível começar o seu enovelamento por qualquer uma das suas três  $\alpha$ -hélices. Logo, pode-se inferir que esta proteína possui um rápido enovelamento, uma vez que ela possui um alto número de conformações em seu estado de transição.

### 3.3 Energia livre em função de uma coordenada de reação

Uma maneira de testar a metodologia desenvolvida nesse trabalho é realizar o cálculo da energia livre em função de uma coordenada de reação. Para comparar os resultados, realizou-se o cálculo da energia livre por dois métodos distintos. O primeiro deles foi por meio da maneira mais divulgada na literatura para o cálculo de energia livre, o método dos múltiplos histogramas (WHAM) [5]. O segundo método utilizou-se como base a projeção em 2D. Partindo de uma estrutura nativa (qualquer conformação com o  $Q$  próximo ao  $Q$  nativo), calcula-se a função de distribuição radial (FDR). Essa função descreve como a densidade de conformações varia em função da distância de um ponto definido inicialmente (neste caso, o estado nativo). Assim, calcula-se o número de estruturas contidas na área do anel à uma distância  $r + dr$  do ponto de referência. O incrementado na distância ( $r + dr$ ) é dado por:  $0, 1\rho$ , onde  $\rho$  é definido como a distância da conformação mais distante do estado de referência. A correspondência entre esses dois métodos pode ser visto na Figura 3.5

Para as três proteínas, os resultados foram altamente correlacionado, sendo possível, de maneira qualitativa, recuperar as barreiras de transição. Esse tipo de análise se torna interessante, uma vez que não é necessário definir uma coordenada de reação (como por exemplo,  $Q$ , RMSD, distância entre aminoácidos, etc). A informação estrutural, que normalmente é usada nessas coordenadas, está implícita na projeção em 2D. Outro



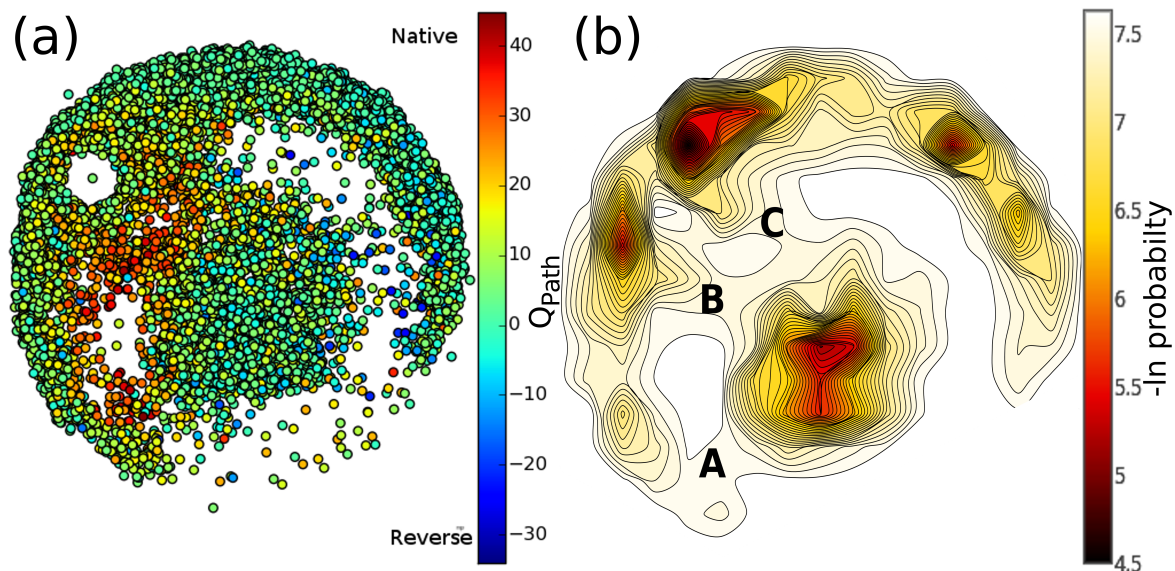
**Figura 3.5** Energia Livre vs Coordenada de Reação. Cada gráfico, (a), (b) e (c) representa a energia livre do enovelamento para as proteínas CI2, SH3 e Proteína A, respectivamente. A linha tracejada em vermelho é a energia livre calculada pelo método WHAM [5]. A linha em preto é a energia livre calculada usando a função de distribuição radial (FDR). A FDR é calculada usando a estrutura nativa como referência.

fator importante é que o cálculo de energia livre por meio da FDR, utilizou as projeções em 2D apresentadas na Figura 3.4, que são precedentes de uma simulação longa na temperatura de folding ( $T_f$ ).

### 3.4 Rotas dominantes de enovelamento no Domínio SH3

Trabalhos anteriores apresentaram resultados que mostram que o enovelamento do domínio SH3 é predominantemente guiado por fatores geométricos [71, 72]. No trabalho de Lammert, H. et. al. [6], foi mostrado, por meio de simulações, que essa proteína possui rotas dominantes que minimizam o tempo que a proteína leva para enovelar-se. Esta rota dominante é definida por alguns contatos que possuem uma alta frequência de formação no estado de transição. No trabalho citado, é apresentada uma coordenada de reação que mede se os contatos que a proteína realiza fazem parte ou não desta rota dominante. Essa coordenada será chamada de  $Q_{path}$ .

Para descrever detalhes específicos sobre essa rota, investigou-se como ela está distribuída na projeção em 2D. Para comparar os resultados, realizou-se as simulações de enovelamento com os mesmos parâmetros usados no trabalho de Lammert. Inicialmente a projeção em 2D foi colorido pela coordenada de reação  $Q$  (Figura 3.4 (b)), nesta



**Figura 3.6** (a) Projeção em 2D da proteína SH3 colorida pela coordenada de reação  $Q_{path}$ . Esta coordenada representa os contatos necessários para a proteína se enovelar através da rota dominante. Seus valores variam desde valores negativos (enovelamento reverso) até valores positivos (enovelamento nativo) [6]. (b) Gráfico de curva de nível da superfície de energia livre, utilizando a projeção em 2D como referência. Neste caso, a cor representa os valores calculados pela equação 3.1. Os três pontos em destaque (A, B e C) representam as regiões onde possui rotas preferenciais para o enovelamento.

figura, é possível observar três rotas bem definidas no estado de transição (pontos verdes). O próximo passo foi colorir a projeção em 2D usando a coordenada de reação  $Q_{path}$ . De acordo com a sua definição, essa coordenada varia desde valores negativos, os quais correspondem a enovelamentos feitos por rotas diferentes da rota dominantes (*Reverse*), até valores positivos, que correspondem aos enovelamentos que concordam com a rota dominante (*Native*).

Na Figura 3.6(a) é apresentado a projeção em 2D colorida pela coordenada de reação  $Q_{path}$ . As conformações em vermelho representam os estados de transição com valores positivos de  $Q_{path}$ , no lado direito do gráfico é possível observar as conformações em azul, estas conformações representam os valores negativos de  $Q_{path}$ . Note que as três rotas definidas anteriormente são, predominantemente, conformações em vermelho. De acordo com a projeção, os caminhos possíveis para o enovelamento do domínio SH3 podem ser separado em dois grupos: O enovelamento nativo (conformações a esquerda do estado nativo) e o enovelamento reverso (conformações a direita do estado nativo), Como essas duas regiões estão diametralmente opostas na projeção, pode-se afirmarm

### 3.4 Rotas dominantes de enovelamento no Domínio SH3

---

que a suas estruturas são bastante distintas, o que corrobora com os resultados de Lammert.

Para entender as possíveis diferenças e aprofundar o entendimento dessas três rotas, calculou-se a superfície de energia livre. Usando como base a projeção em 2D, calculou-se o histograma em 2D, ou seja, a densidade de estados por unidade de grid (o tamanho do grid é definido arbitrariamente, de acordo com a resolução requerida). A energia livre por unidade de grid é dado por:

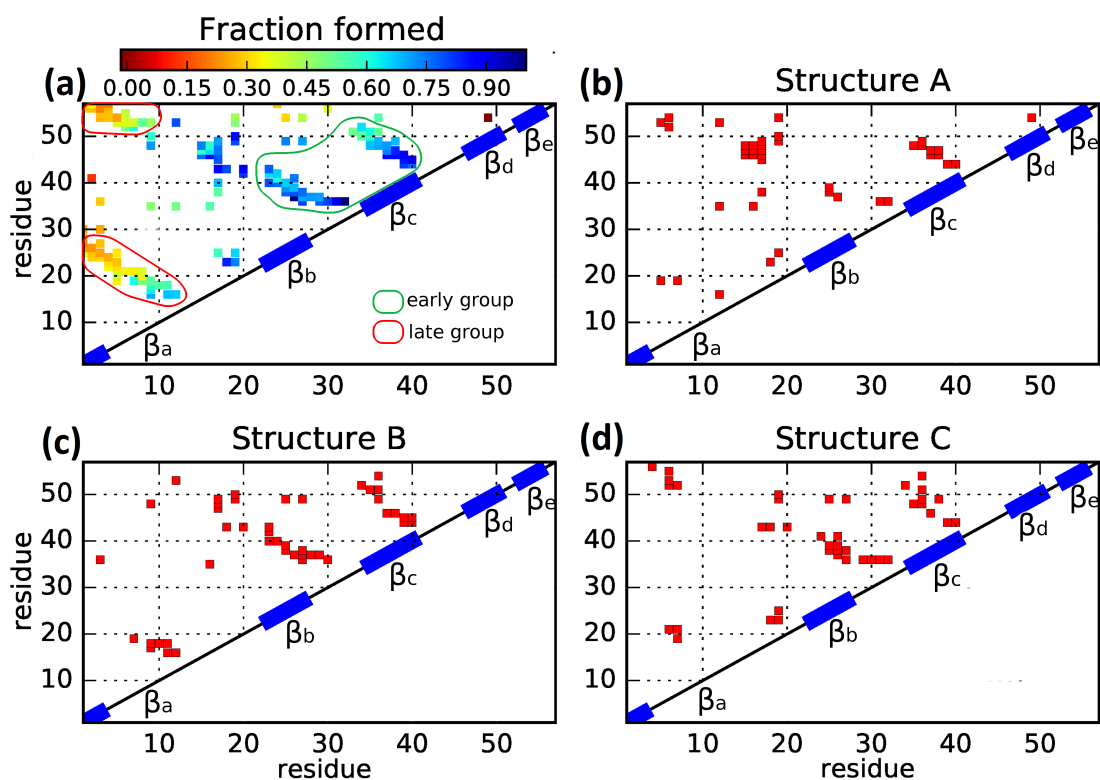
$$\Delta\mathcal{F} = -\ln\left(\frac{n_i}{N}\right), \quad (3.1)$$

onde  $n_i$  é a densidade de estados calculada pelo histograma em 2D e  $N$  é o número total de conformações projetadas. Na Figura 3.6(b) é apresentado um gráfico de curva de nível, onde a cor representa a energia livre calculada pela equação 3.1

Na Figura 3.6 (b) é possível identificar exatamente três rotas de enovelamento (A, B e C) que possuem baixas barreiras de transição dos estados desenovelado para o estado nativo. Note que as rotas B e C possuem barreiras menores (próximo de 7) quando comparadas com a rota A (próximo de 7,5). Já no caso do enovelamento reverso (lado direito da figura) é possível notar que possui uma barreira muito maior (com valores maiores que 7,5), quando comparado com o enovelamento nativo. Esses resultados estão de acordo com o que foi proposto por Lammert et. al.

Adicionalmente, analisaram-se as diferenças entre os mapas de contato de conformações representativas em cada uma das regiões realçadas na figura 3.6 (b). Este resultado pode ser observado na Figura 3.7. Note que as três rotas possuem os contatos nomeados por Lammert como *early group*, no entanto, as rotas B e C (estruturas B and C) possuem uma formação predominante de contatos entre as folhas  $\beta_b$  e  $\beta_c$ , enquanto a rota A possui preferencialmente a formação de contatos entre as folhas  $\beta_c$  e  $\beta_d$ . Essa diferença provavelmente está associada a um custo energético maior, como apresentado no gráfico de energia livre.





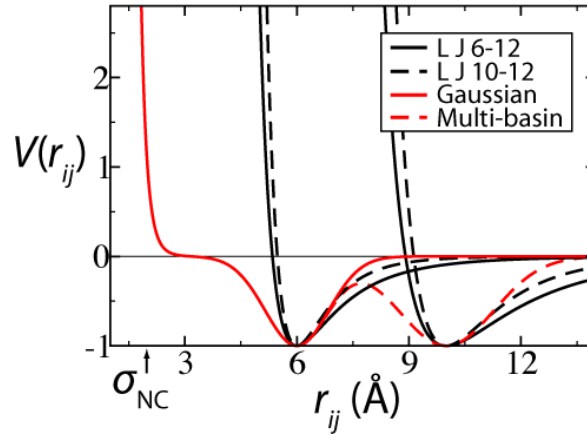
**Figura 3.7** Análise do mapa de contato das rotas do enovelamento da proteína SH3. (a) Representa a o mapa de contato da proteína SH3 colorido pela fração de contatos formados no estado de transição. Dois grupos são realçados: *Early group* (em verde) representando os contatos mais prováveis (entre as folhas  $\beta_b$ ,  $\beta_c$  e  $\beta_d$ ) e o *late group* (em vermelho), apresentando os contatos raramente formado no estado de transição (entre as folhas  $\beta_c$  e  $\beta_d$ ). Os gráficos (b), (c) e (d) apresentam conformações representativas de cada rota marcadas na figura 3.6 (b). Note que os contatos feitos em (b) (entre as folhas  $\beta_c$  e  $\beta_d$ ) diferem dos contatos feitos em (c) e em (d) (muito similar aos contatos feitos no *early group*).

## 3.5 Análise das imagens espelhadas na proteína A

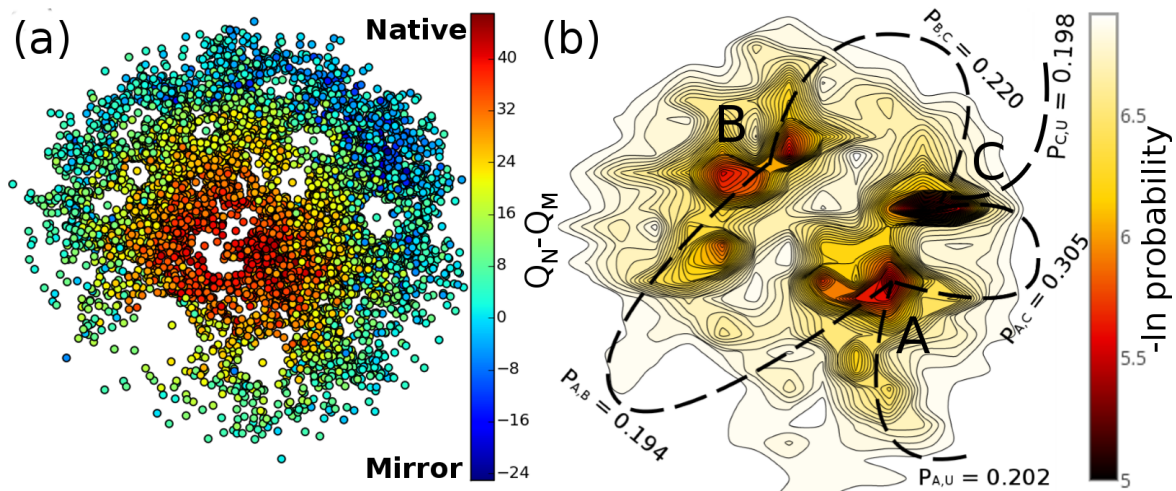
Imagens espelhadas (*mirror image*) são formadas quando a estrutura da proteína, além de múltiplas rotas de enovelamento, possui mais de uma estrutura compatível com os contatos nativos, o que permite a essas estruturas serem energeticamente competitivas com o estado nativo [74, 75]. Para entender a competitividade entre os estados nativo/espelhado da proteína A, Noel et. al. realizou simulações e mostrou que o estado nativo e seu estado espelhado têm uma estabilidade entálpica similar e que são termodinamicamente competitivas [7].

Como essas diferenças são intrinsecamente estruturais, buscou-se entender como a imagem espelhada espalha-se em uma projeção em 2D. Para investigar esse processo, realizaram-se simulações utilizando o mesmo modelo utilizado por Noel. Trata-se de um modelo baseado em estrutura, com a diferença de que, ao invés de utilizar um potencial que leva diretamente ao estado nativo (modelo  $G\bar{o}$ ), agora o potencial tem dois mínimos igualmente estáveis (esse modelo é conhecido como *Multi-basin*). Neste caso, um mínimo corresponde ao estado nativo e o outro mínimo de energia é o estado espelhado. De maneira mais formal, o termo de Lennard-Jones da equação 2.1 é substituído por um potencial gaussiano com dois mínimos igualmente prováveis. A Figura 3.8 ilustra essa diferença, detalhes sobre simulações com esse modelo podem ser encontradas na referência [62].

Na Figura 3.9(a) é apresentada a visualização em 2D da Proteína A, no qual foi utilizada a mesma metodologia desenvolvida neste trabalho, exceto pela diferença no potencial, considerado conforme explicado no parágrafo anterior. Para entender como as conformações nativa/espelhada estão espalhadas na projeção, coloriu-se o gráfico com a coordenada de reação  $Q_N - Q_M$ , que representa a soma de todos os contatos terciários realizados no estado nativo  $Q_N$  e no estado espelhado  $Q_M$  (detalhes sobre a escolha dessa coordenada podem ser encontrado na referência [7]). As conformações em azul representam os estados similares ao estado espelhado, ao passo que as conformações em vermelho representam os estados similares ao estado nativo. Nesta figura é possível ver um aumento no cluster do estado nativo, quando comparado com a projeção apresentada na Figura 3.4(c). Este “alargamento” é devido a competitividade entre o estado nativo e o estado espelhado.



**Figura 3.8** Comparação entre os potenciais de Lennard-Jones e Gaussiano: As curvas em preto representam os potenciais de Lennard-Jones, com mínimos em 6 e 10 Å. As curvas em vermelho representam os potenciais gaussianos, enquanto que a linha tracejada representa o modelo *Multi-basin*.



**Figura 3.9** Análise das imagens espelhadas da proteína A. (a) Visualização em 2D da Proteína A proveniente de simulações usando o modelo *SBM multi-basin* com potencial gaussiano. Além disso, utilizou-se as mesmas configurações encontrada no artigo de Noel et. al. [7]. A cor representa a coordenada de reação  $Q_N - Q_M$ , onde os pontos em vermelho referem-se a formação de contatos nativos e os pontos em azul representam a formação dos contatos da estrutura espelhadas (*mirror state*). (b) A curva de nível que representa a densidade de estados calculada pelo histograma em 2D. Três regiões são realçadas: (A) Estruturas nativas, (B) Estrutura quase-nativas (*native-like*) e, (C) estruturas espelhadas. As linhas tracejadas representam as probabilidades da conversão entre esse estados e o estado desenovelado.

### 3.5 Análise das imagens espelhadas na proteína A

---

Para entender o significado dessas diferenças, calculou-se o gráfico de curva de nível seguindo o mesmo método descrito na seção anterior. Na figura 3.9(b) são observados três mínimos bem distintos: *i)* O cluster nativo (Região A); *ii)* os estados próximo ao estado nativo, *native-like* (região B) e; *iii)* o cluster de estados espelhado (região C).

Para entender a transição entre esses estados e comparar as barreiras entre eles, calcularam-se as probabilidades de transição entre essas conformações e também a probabilidade de, a partir do estado nativo ou do estado espelhado, a proteína desenovelar. Tais probabilidades estão de acordo com as barreiras encontradas, ou seja, quanto maior a barreira entre conformações, menor a probabilidade de transição.



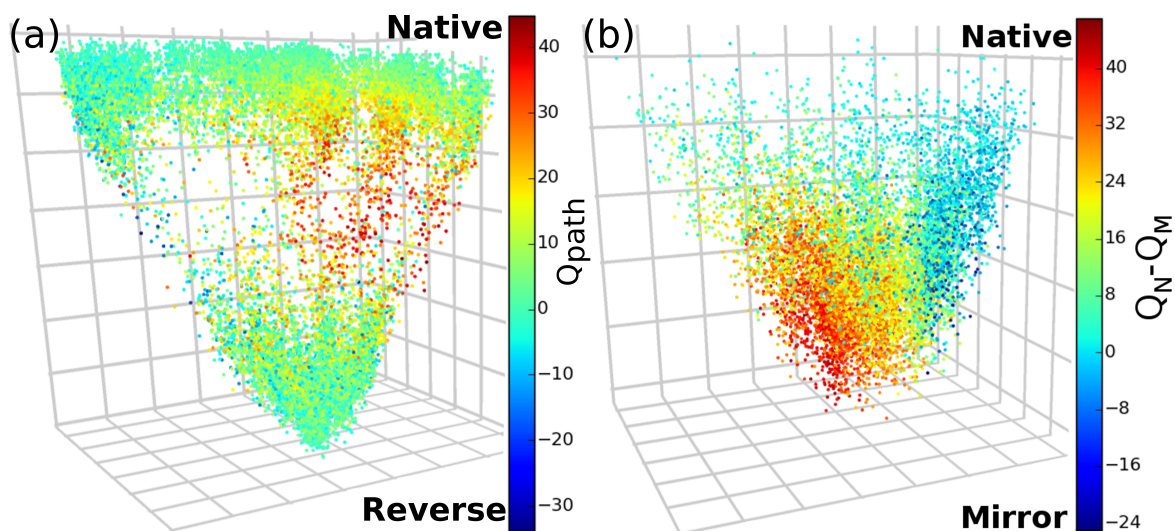
# Capítulo 4

## Conclusões

Neste trabalho foi apresentado uma nova metodologia para visualização de superfícies de energia. O método proposto baseia-se em aplicar uma métrica que leva em conta a informação estrutural de proteínas para diferenciá-las e representá-las em um espaço dimensional reduzido. Os resultados apresentados mostraram que cada proteína apresenta um padrão de visualização, sendo possível observar rotas de enovelamento, assim como suas particularidades. Tais particularidades são negligenciadas quando, por exemplo, o enovelamento é representado utilizando uma coordenada de reação (como a fração de contatos nativos  $Q$ , RMSD, etc).

Uma vez que a projeção em 2D é baseada na dissimilaridade conformacional, o primeiro resultado apresentado mostrou que é possível recuperar a energia livre, calculando a função de distribuição radial, e utilizando como referencia o estado nativo (Figura 3.5). Esse resultado mostra que a projeção em 2D é robusta, uma vez que as conformações utilizadas nesse cálculo são provenientes de uma simulação realizada na temperatura de folding ( $T_f$ ). Portanto, é esperado que o perfil de energia livre calculado pelo método tradicional WHAM esteja em acordo, pelo menos qualitativamente, com a função de distribuição radial.

Por levar em conta as características estruturais de cada conformação para a descrição da superfície de enovelamento, foi possível descrever os estados de transição com mais detalhes. Por exemplo, resultados anteriores mostraram que o domínio SH3 possui uma rota dominante para enovelar-se [6]. Quando analisou-se a projeção em



**Figura 4.1** Visualização do enovelamento de proteínas em 3D. Nesta figura é apresentado uma visão geral dos resultados apresentados neste trabalho, a projeção em 2D é utilizado como plano XY, o eixo Z representa a energia e a cor é definida pela coordenada de reação apropriada para cada proteína (Essas coordenadas são discutidas em suas respectivas seções). (a) Os resultados obtidos para o domínio SH3. (b) os resultados obtidos para Proteína A.

2D da SH3, percebeu-se que esta rota dominante pode ser dividida em dois subgrupos nesta rota dominante (o caminho A e o caminho B/C na Figura 3.6).

No caso da Proteína A, foi analisada a competição entre o estado nativo e o estado espelhado, conforme apresentado por Noel et. al. [57]. O resultado foi reproduzido neste trabalho e com base na a projeção em 2D (Figura 3.9), foi possível observar a formação do cluster do estado espelhado (pontos em azul). A partir desse cluster, foi possível calcular as barreiras de transição (e as probabilidades) entre esses estados (nativo e espelhado). Esse tipo de análise é interessante, pois não é necessário definir uma coordenada de reação, uma vez que a informação está implícita na projeção.

Para representar os resultados de maneira geral, construiu-se um gráfico onde é apresentada toda informação obtida pela metodologia proposta neste trabalho. A Figura 4.1 utiliza a projeção em 2D como plano XY, o eixo Z representa a energia e a cor representa a coordenada específica para cada proteína ( $Q_{path}$  para o domínio SH3 e  $Q_M - Q_N$  para a Proteína A).

A visualização em 2D do enovelamento de proteínas baseada na métrica de similaridades pode apresentar grande ganho no estudo de proteínas que possuem importantes detalhes estruturais, como é o caso de proteínas intrinsecamente desordenadas (IDPs),

---

nos quais o estado nativo não é muito bem definido. Acredita-se que, ao realizar um escaneamento no mapa conformacional dessas proteínas, será possível encontrar clusters que definam as regiões onde pode acontecer atividades nessa proteínas e, portanto, entender como são formados tais regiões. Outra caso bastante interessante para ser alvo dessa metodologia é o Prion. Prion possui um enovelamento múltiplo, onde, ao menos, uma dessas possibilidades de enovelamento possui um caráter infeccioso parecido com a infecção viral. Baseado nos estudos apresentados aqui, nota-se que existe uma gama de possibilidade para aplicação do método desenvolvido, principalmente por não utilizar uma coordenada de reação específica e, possuindo como única distinção, a similaridade das conformações visitadas durante a simulação.





# Referências Bibliográficas

- [1] Joseph D Bryngelson, José Nelson Onuchic, Nicholas D Socci, and Peter G Wolynes. Funnels, pathways, and the energy landscape of protein folding: A synthesis. *Proteins: Structure, Function, and Bioinformatics*, 21(3):167–195, March 1995.
- [2] J N Onuchic, Z Luthey-Schulten, and P G Wolynes. Theory of protein folding: the energy landscape perspective. *Annual Review of Physical Chemistry*, 48:545–600, 1997. PMID: 9348663.
- [3] Vinícius de Godoi Contessoto. *Estudo do efeito da adição de frustração no enovelamento de proteínas utilizando modelos baseados em estruturas*. Dissertação de Mestrado, Universidade Estadual Paulista (UNESP), 2012.
- [4] Y Harpaz, N Elmasry, A R Fersht, and K Henrick. Direct observation of better hydration at the n terminus of an alpha-helix with glycine rather than alanine as the n-cap residue. *Proceedings of the National Academy of Sciences*, 91(1):311–315, January 1994.
- [5] Alan M. Ferrenberg and Robert H. Swendsen. New monte carlo technique for studying phase transitions. *Physical Review Letters*, 61(23):2635–2638, December 1988.
- [6] Heiko Lammert, Jeffrey K. Noel, and JosÃ© N. Onuchic. The Dominant Folding Route Minimizes Backbone Distortion in SH3. *PLOS Comput Biol*, 8(11):e1002776, November 2012.
- [7] Jeffrey K. Noel, Alexander Schug, Abhinav Verma, Wolfgang Wenzel, Angel E. Garcia, and JosÃ© N. Onuchic. Mirror Images as Naturally Competing Conformations in Protein Folding. *The Journal of Physical Chemistry B*, 116(23):6880–6888, June 2012.
- [8] Donald Voet, Judith G. Voet, and Charlotte W. Pratt. *Principles of biochemistry*. J. Wiley & sons, April 2008.
- [9] C.M. Dobson. Protein folding and misfolding. *Nature*, 426(6968):884–890, 2003.
- [10] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. *Molecular Biology of the Cell*. Artmed, 4 edition, 2004.

## Referências Bibliográficas

---

- [11] Albert L. Lehninger, David Lee Nelson, and Michael M. Cox. *Lehninger principles of biochemistry*. W.H. Freeman, 2005.
- [12] C B Anfinsen. Principles that govern the folding of protein chains. *Science (New York, N.Y.)*, 181(4096):223–230, July 1973. PMID: 4124164.
- [13] Cyrus Levinthal. Are there pathways for protein folding? *Extrait du Journal de Chimie Physique*, 65(1), 1968.
- [14] J D Bryngelson and P G Wolynes. Spin glasses and the statistical mechanics of protein folding. *Proceedings of the National Academy of Sciences*, 84(21):7524–7528, November 1987.
- [15] Joseph D. Bryngelson and Peter G. Wolynes. Intermediates and barrier crossing in a random energy model (with applications to protein folding). *The Journal of Physical Chemistry*, 93(19):6902–6915, 1989.
- [16] P E Leopold, M Montal, and J N Onuchic. Protein folding funnels: a kinetic approach to the sequence-structure relationship. *Proceedings of the National Academy of Sciences of the United States of America*, 89(18):8721–8725, September 1992. PMID: 1528885 PMCID: 49992.
- [17] Ken A. Dill, S. Banu Ozkan, M. Scott Shell, and Thomas R. Weikl. The protein folding problem. *Annual Review of Biophysics*, 37(1):289–316, 2008. PMID: 18573083.
- [18] D. Thirumalai, Edward P. O’Brien, Greg Morrison, and Changbong Hyeon. Theoretical perspectives on protein folding. *Annual Review of Biophysics*, 39(1):159–183, 2010. PMID: 20192765.
- [19] H Frauenfelder, SG Sligar, and PG Wolynes. The energy landscapes and motions of proteins. *Science*, 254(5038):1598–1603, December 1991.
- [20] José Nelson Onuchic, Hugh Nymeyer, Angel E. García, Jorge Chahine, and Nicholas D. Socci. The energy landscape theory of protein folding: Insights into folding mechanisms and scenarios. In *Advances in Protein Chemistry*, volume 53, pages 87–152. Elsevier, 2000.
- [21] Steven S. Plotkin and José N. Onuchic. Understanding protein folding with energy landscape theory part i: Basic concepts. *Quarterly Reviews of Biophysics*, 35(02), August 2002.
- [22] Steven S. Plotkin and José N. Onuchic. Understanding protein folding with energy landscape theory part II: quantitative aspects. *Quarterly Reviews of Biophysics*, 35(03), January 2003.
- [23] David Wales. *Energy Landscapes: Applications to Clusters, Biomolecules and Glasses*. Cambridge University Press, 2003.
- [24] E I Shakhnovich and A M Gutin. Formation of unique structure in polypeptide chains. theoretical investigation with the aid of a replica approach. *Biophysical Chemistry*, 34(3):187–199, November 1989.

- [25] Samuel S. Cho, Yaakov Levy, and Peter G. Wolynes. P versus q: Structural reaction coordinates capture protein folding on smooth landscapes. *Proceedings of the National Academy of Sciences of the United States of America*, 103(3):586–591, January 2006.
- [26] Harold A. Scheraga, Mey Khalili, and Adam Liwo. Protein-folding dynamics: Overview of molecular simulation techniques. *Annual Review of Physical Chemistry*, 58(1):57–83, 2007.
- [27] P. Wolynes, J. Onuchic, and D. Thirumalai. Navigating the folding routes. *Science*, 267(5204):1619–1620, 1995.
- [28] Jorge Chahine, Hugh Nymeyer, Vitor B. P. Leite, Nicholas D. Socci, and José Nelson Onuchic. Specific and nonspecific collapse in protein folding funnels. *Physical Review Letters*, 88(16):168101, April 2002.
- [29] J. Wang, R. J. Oliveira, X. Chu, P. C. Whitford, J. Chahine, W. Han, E. Wang, J. N. Onuchic, and V. B. P. Leite. Topography of funneled landscapes determines the thermodynamics and kinetics of protein folding. *Proceedings of the National Academy of Sciences*, 109(39):15763–15768, September 2012.
- [30] N. D Socci, J. N Onuchic, and P. G Wolynes. Diffusive dynamics of the reaction coordinate for protein folding funnels. *arXiv:cond-mat/9601091*, January 1996.
- [31] Jorge Chahine, Ronaldo J Oliveira, Vitor B. P Leite, and Jin Wang. Configuration-dependent diffusion can shift the kinetic transition state and barrier height of protein folding. *Proceedings of the National Academy of Sciences*, 104(37):14646–14651, September 2007.
- [32] R.J. Oliveira, P.C. Whitford, J. Chahine, V.B.P. Leite, and J. Wang. Coordinate and time-dependent diffusion dynamics in protein folding. *Methods*, 52(1):91–98, 2010.
- [33] David J Wales. Energy landscapes: some new horizons. *Current Opinion in Structural Biology*, 20(1):3–10, February 2010.
- [34] Jose M. Sanchez-Ruiz. Probing Free-Energy Surfaces with Differential Scanning Calorimetry. *Annual Review of Physical Chemistry*, 62(1):231–255, 2011.
- [35] Jie Wen, Kelly Arthur, Letha Chemmalil, Salman Muzammil, John Gabrielson, and Yijia Jiang. Applications of differential scanning calorimetry for thermal stability analysis of proteins: Qualification of DSC. *Journal of Pharmaceutical Sciences*, 101(3):955–964, March 2012.
- [36] Christopher M. Johnson. Differential scanning calorimetry as a tool for protein folding and stability. *Archives of Biochemistry and Biophysics*, 531(1–2):100–109, March 2013.
- [37] Oren M Becker and Martin Karplus. The topology of multidimensional potential energy surfaces: Theory and application to peptide structure and kinetics. *The Journal of Chemical Physics*, 106(4):1495–1517, January 1997.

## Referências Bibliográficas

---

- [38] Frank Noe, Illia Horenko, Christof Schutte, and Jeremy C. Smith. Hierarchical analysis of conformational dynamics in biomolecules: Transition networks of metastable states. *126(15):155102*, 2007.
- [39] Diego Prada-Gracia, Jesús Gómez-Gardeñes, Pablo Echenique, and Fernando Falo. Exploring the free energy landscape: From dynamics to networks and back. *PLoS Comput Biol*, 5(6):e1000415, June 2009.
- [40] Alex Dickson and Charles L. Brooks. Native states of fast-folding proteins are kinetic traps. *Journal of the American Chemical Society*, 135(12):4729–4734, March 2013.
- [41] Frank Noé and Stefan Fischer. Transition networks for modeling the kinetics of conformational change in macromolecules. *Current Opinion in Structural Biology*, 18(2):154–162, April 2008.
- [42] Michael C. Prentiss, David J. Wales, and Peter G. Wolynes. The energy landscape, folding pathways and the kinetics of a knotted protein. *PLoS Comput Biol*, 6(7):e1000835, July 2010.
- [43] Francesco Rao and Amedeo Caflisch. The protein folding network. *Journal of molecular biology*, 342(1):299–306, September 2004. PMID: 15313625.
- [44] Oren M Becker. Quantitative visualization of a macromolecular potential energy “funnel”. *Journal of Molecular Structure: THEOCHEM*, 398–399:507–516, June 1997.
- [45] Mark A. Miller, Jonathan P. K. Doye, and David J. Wales. Structural relaxation in atomic clusters: Master equation dynamics. *Physical Review E*, 60(4):3701–3718, October 1999.
- [46] Jonathan Doye, Mark Miller, and David Wales. The double-funnel energy landscape of the 38-atom lennard-jones cluster. *The Journal of Chemical Physics*, 110(14):6896, 1999. arXiv:cond-mat/9808265.
- [47] Jonathan Doye, Mark Miller, and David Wales. Evolution of the potential energy surface with size for lennard-jones clusters. *The Journal of Chemical Physics*, 111(18):8417, 1999. arXiv:cond-mat/9903305.
- [48] David J. Wales and Peter E. J. Dewsbury. Effect of salt bridges on the energy landscape of a model protein. *The Journal of Chemical Physics*, 121(20):10284–10290, November 2004.
- [49] Mark A. Miller and David J. Wales. Energy landscape of a model protein. *The Journal of Chemical Physics*, 111(14):6610–6616, October 1999.
- [50] Lewis C. Smeeton, Mark T. Oakley, and Roy L. Johnston. Visualizing energy landscapes with metric disconnectivity graphs. *Journal of Computational Chemistry*, 35(20):1481–1490, July 2014.

- [51] Antonio B. Oliveira, Jr., Francisco M. Fatore, Fernando V. Paulovich, Osvaldo N. Oliveira, Jr., and Vitor B. P. Leite. Visualization of protein folding funnels in lattice models. *PLoS ONE*, 9(7):e100861, July 2014.
- [52] P.C. Whitford, J.K. Noel, S. Gosavi, A. Schug, K.Y. Sanbonmatsu, and J.N. Onuchic. An all-atom structure-based potential for proteins: Bridging minimal models with all-atom empirical forcefields. *Proteins: Struct. Func. Bioinf.*, 75(2):430–441, 2009.
- [53] P.C. Whitford, O. Miyashita, Y. Levy, and J.N. Onuchic. Conformational transitions of adenylate kinase: Switching by cracking. *J. Mol. Biol.*, 366:1661–1671, 2007.
- [54] R.B. Best, Y. Chen, and G. Hummer. Slow protein conformational dynamics from multiple experimental structures: The helix/sheet transition of arc repressor. *Structure*, 13:1755–1763, 2005.
- [55] D.M. Zuckerman. Simulation of an ensemble of conformational transitions in a united-residue model of calmodulin. *J. Phys. Chem. B*, 108:5127–5137, 2004.
- [56] Q. Lu and J. Wang. Single molecule conformational dynamics of adenylate kinase: Energy landscape, structural correlations, and transition state ensembles. *J. Am. Chem. Soc.*, 130(14):4772–4783, 2008.
- [57] Jeffrey K. Noel and JosÃ© N. Onuchic. The many faces of structure-based potentials: From protein folding landscapes to structural characterization of complex biomolecules. In Nikolay V. Dokholyan, editor, *Computational Modeling of Biological Systems*, Biological and Medical Physics, Biomedical Engineering, pages 31–54. Springer US. DOI: 10.1007/978-1-4614-2146-7\_2.
- [58] C. Clementi, H. Nymeyer, and J.N. Onuchic. Topological and energetic factors: What determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? An investigation for small globular proteins. *J. Mol. Biol.*, 298(5):937–953, 2000.
- [59] Y. Ueda, H. Taketomi, and N. Gō. Studies on protein folding, unfolding and fluctuations by computer simulation. I. The effects of specific amino acid sequence represented by specific inter-unit interactions. *Int. J. Peptide Res.*, 7:445–459, 1975.
- [60] V. Sobolev, R. Wade, G. Vried, and M. Edelman. Molecular docking using surface complementarity. *Proteins: Struct. Func. Genet.*, 25:120–129, 1996.
- [61] Sander Pronk, Szilárd Páll, Roland Schulz, Per Larsson, Pr Bjelkmar, Rossen Apostolov, Michael R. Shirts, Jeremy C. Smith, Peter M. Kasson, David van der Spoel, Berk Hess, and Erik Lindahl. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*, 29(7):845–854, April 2013.
- [62] <http://smog-server.org/>.

## Referências Bibliográficas

---

- [63] Benjamin Lutz, Claude Sinner, Geertje Heuermann, Abhinav Verma, and Alexander Schug. eSBMTools 1.0: enhanced native structure-based modeling tools. *Bioinformatics (Oxford, England)*, 29(21):2795–2796, November 2013.
- [64] Corey Hardin, Michael P. Eastwood, Michael C. Prentiss, Zadia Luthey-Schulten, and Peter G. Wolynes. Associative memory hamiltonians for structure prediction without homology:  $\beta/\beta$  proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 100(4):1679–1684, February 2003.
- [65] Eduardo Tejada, Rosane Minghim, and Luis Gustavo Nonato. On improved projection techniques to support visual exploration of multi-dimensional data sets. *Information Visualization*, 2(4):218–231, December 2003.
- [66] Sophie E. Jackson and Alan R. Fersht. Folding of chymotrypsin inhibitor 2. 2. influence of proline isomerization on the folding kinetics and thermodynamic characterization of the transition state of folding. *Biochemistry*, 30(43):10436–10443, October 1991.
- [67] S E Jackson, N elMasry, and A R Fersht. Structure of the hydrophobic core in the transition state for folding of chymotrypsin inhibitor 2: a critical test of the protein engineering method of analysis. *Biochemistry*, 32(42):11270–11278, October 1993. PMID: 8218192.
- [68] C Clementi, H Nymeyer, and J N Onuchic. Topological and energetic factors: what determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? an investigation for small globular proteins. *Journal of molecular biology*, 298(5):937–953, May 2000. PMID: 10801360.
- [69] A. Li and V. Daggett. Identification and characterization of the unfolding transition state of chymotrypsin inhibitor 2 by molecular dynamics simulations. *Journal of Molecular Biology*, 257(2):412–429, March 1996.
- [70] V. P. Grantcharova, D. S. Riddle, J. V. Santiago, and D. Baker. Important role of hydrogen bonds in the structurally polarized transition state for folding of the src SH3 domain. *Nature Structural Biology*, 5(8):714–720, August 1998.
- [71] J. C. MartÁnez and L. Serrano. The folding transition state between SH3 domains is conformationally restricted and evolutionarily conserved. *Nature Structural Biology*, 6(11):1010–1016, November 1999.
- [72] Jose M. Borreguero, Feng Ding, Sergey V. Buldyrev, H. Eugene Stanley, and Nikolay V. Dokholyan. Multiple Folding Pathways of the SH3 Domain. *Biophysical Journal*, 87(1):521–533, July 2004.
- [73] E. M. Boczko and C. L. Brooks. First-principles calculation of the folding free energy of a three-helix bundle protein. *Science (New York, N.Y.)*, 269(5222):393–396, July 1995.
- [74] K. A. Olszewski, A. Kolinski, and J. Skolnick. Folding simulations and computer redesign of protein A three-helix bundle motifs. *Proteins*, 25(3):286–299, July 1996.

- [75] Khatuna Kachlishvili, Gia G. Maisuradze, Osvaldo A. Martin, Adam Liwo, Jorge A. Vila, and Harold A. Scheraga. Accounting for a mirror-image conformation as a subtle effect in protein folding. *Proceedings of the National Academy of Sciences*, 111(23):8458–8463, June 2014.
- [76] Alan M. Ferrenberg and Robert H. Swendsen. Optimized monte carlo data analysis. *Physical Review Letters*, 63(12):1195–1198, September 1989.





# Apêndice A

## Artigos



# Visualization of Protein Folding Funnels in Lattice Models

Antonio B. Oliveira Jr.<sup>1</sup>, Francisco M. Fatore<sup>2</sup>, Fernando V. Paulovich<sup>2</sup>, Osvaldo N. Oliveira Jr.<sup>3</sup>, Vitor B. P. Leite<sup>1\*</sup>

**1** Departamento de Física, Instituto de Biociências, Letras e Ciências Exatas, Universidade Estadual Paulista, São José do Rio Preto, São Paulo, Brazil, **2** Instituto de Ciências Matemáticas e Computação, Universidade de São Paulo, São Carlos, São Paulo, Brazil, **3** Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, São Paulo, Brazil

## Abstract

Protein folding occurs in a very high dimensional phase space with an exponentially large number of states, and according to the energy landscape theory it exhibits a topology resembling a funnel. In this statistical approach, the folding mechanism is unveiled by describing the local minima in an effective one-dimensional representation. Other approaches based on potential energy landscapes address the hierarchical structure of local energy minima through disconnectivity graphs. In this paper, we introduce a metric to describe the distance between any two conformations, which also allows us to go beyond the one-dimensional representation and visualize the folding funnel in 2D and 3D. In this way it is possible to assess the folding process in detail, e.g., by identifying the connectivity between conformations and establishing the paths to reach the native state, in addition to regions where trapping may occur. Unlike the disconnectivity maps method, which is based on the kinetic connections between states, our methodology is based on structural similarities inferred from the new metric. The method was developed in a 27-mer protein lattice model, folded into a  $3 \times 3 \times 3$  cube. Five sequences were studied and distinct funnels were generated in an analysis restricted to conformations from the transition-state to the native configuration. Consistent with the expected results from the energy landscape theory, folding routes can be visualized to probe different regions of the phase space, as well as determine the difficulty in folding of the distinct sequences. Changes in the landscape due to mutations were visualized, with the comparison between wild and mutated local minima in a single map, which serves to identify different trapping regions. The extension of this approach to more realistic models and its use in combination with other approaches are discussed.

**Citation:** Oliveira AB Jr., Fatore FM, Paulovich FV, Oliveira ON Jr., Leite VBP (2014) Visualization of Protein Folding Funnels in Lattice Models. PLoS ONE 9(7): e100861. doi:10.1371/journal.pone.0100861

**Editor:** Yaakov Koby Levy, Weizmann Institute of Science, Israel

**Received:** February 21, 2014; **Accepted:** May 31, 2014; **Published:** July 10, 2014

**Copyright:** © 2014 Oliveira et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Funding:** This work was supported by FAPESP, CNPq, CAPES and nBioNet network (Brazil). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

\* Email: vleite@sjrp.unesp.br

## Introduction

Understanding the processes leading to a protein folding into its native (functional) state is one of the important problems in molecular biophysics. In the 1960s, Anfinsen hypothesized that a protein in its native state and under physiological conditions would adopt such a structure with the lowest possible energy [1]. Though this hypothesis turned out to be correct, no explanation was offered to explain the large range of characteristic folding times, which may vary from milliseconds to seconds. In what became known as the Levinthal Paradox, in 1969 Levinthal argued that, due to an exponentially large number of states, a random search for the native structure would take cosmological times [2]. The solution to this paradox came from the energy landscape theory [3–7], which embeds the statistical nature of the folding process. The folding happens in a very high dimensional space, but in one of the possible descriptions, the complex landscape theory is projected along the reaction folding coordinate. The effective folding landscape topology is like a funnel, which has an energy gradient toward the native state region. This theory explained quantitatively the data for the folding of several proteins [8–14], and the funnel topology is correlated with the thermodynamics

and kinetics of folding [15]. Many aspects of the folding funnel can be inferred from this approach, such as analysis of conformational maps [16,17], folding mechanisms involving mutants [18], and topological features in the transition state [19].

In other approaches, local minima are individually addressed and go beyond one-dimensional representation [20,21]. Visualization of distances between local minima is a very appealing way of showing the underlying structure of the funnel. However, visualizing the local minima poses a significant challenge owing to the multidimensional nature of the system. Among the motivations to investigate the funnel details and its visualization is the potential help in understanding the role of metastable states, kinetic routes and conformational changes associated with protein function [22–24]. The visualization of potential and free energy surfaces is not essential for calculating any dynamic or thermodynamic properties, but it can certainly help in providing insights as to what those properties might be [20,25,26]. Methods such as Principal Component Analysis (PCA) have been used in funnel visualization for isobutyryl-(ala)<sub>3</sub>-NH-methyl (IAN) [27], where disconnectivity graphs were used to visualize the overall organization of the landscape [28]. The potential energy surface is represented in

terms of local minima and the transition states that connect them, providing a convenient coarse-grained representation of the corresponding landscape [29]. This method has been applied to a wide number of systems. For example, Lennard-Jones clusters present multi-funnel characteristics [30–32]. Disconnectivity graphs are able to reveal the effects of gatekeepers in the potential energy surface by raising the energies of low-lying minima relative to the global minimum [33]. The differences in folding efficiencies can also be inferred in proteins with and without frustration for structure based models [34]. Disconnectivity graphs can also be extended for the visualization of free energy landscape, maintaining the description of barriers faithfully [26,35,36]. When rate constants are associated with the rearrangements mediated by each transition state, a kinetic transition network can be defined [37,38]. So the kinetics and thermodynamics of complex transitions can be modeled in terms of transitions between the relevant conformational substates [39–41], in which kinetic transition networks are constructed from geometry optimization and molecular dynamics simulations. These examples show that this method overcomes the fundamental limitations of reaction-coordinate-based methods. Most of these approaches emphasize the kinetic path between probed states, and are able to indicate, for example, the funnel aspect of the landscape against a hub-like hypothesis [41].

In this paper we focus on the structural organization of conformations, looking at the difference of contacts in each conformation. We propose a suitable conformation metric that reflects the underlying landscape in which the kinetics takes place. The method is tested in a 27-mer protein lattice model, folded into a  $3 \times 3 \times 3$  cube, which has been extensively used in protein folding studies [3,42,43], and in particular for visualization methods [44]. We restricted the visualization to local minima of regions from around the transition-state to the native state. These partially folded states are the relevant ones in the study of metastable states and function-related conformation changes. The data obtained from computational simulations in a lattice model were projected on a 2D or 3D plot with the Force-Scheme method [45], which allowed us to map the connectivity of conformations (local minima). The choice of a metric is essential in order to reach a sensible connection between the original data and the projection, and it must efficiently distinguish between pairs of conformations. From the analyses, we noted that distinct sequences lead to different patterns, from which folding routes could be established and the effects from mutations could be probed.

## Results and Discussion

The simulation of the folding dynamics probes the conformations associated with local minima within given time intervals. We are interested in mapping the partially folded states, associated with conformations from the transition-state to the native configuration. The transition state was inferred from the free energy as a function of degree of nativeness (see Supporting Information) for the protein-like sequences A, Af, B, C and D. Conformational states are characterized by the energy and non-bonding contact points for each monomer of the sequence. The dataset thus generated is multidimensional, and its visualization requires dimension reduction projection methods. A crucial point for the projection is to establish a metric for the distance between two conformations. We tried several possibilities, including the Minkowski family of metrics [46], of which the Euclidean distance is one example. These did not lead to physically plausible results since the computation of such metrics considers that lack-of-contact comparisons define similar elements. In the lattice case,

the absence of contact (“0” comparisons) occurs when two conformations do not present contacts. In this scenario a binary distance is a better choice, *i.e.*, only contacts (“1” comparisons) are relevant.

The measure between two conformations  $i$  and  $j$  has to satisfy commutativity and null distance to itself, *i.e.*,

$$M(i,j) = M(j,i) \quad \text{and} \quad M(i,i) = 0. \quad (1)$$

The structural measure or distance shown to be most effective was the ratio between the dissimilarity ( $D_{i,j}$ ) and similarity ( $C_{i,j}$ ) between  $i$  and  $j$ , which is equivalent to the ratio between the Jaccard index and the Jaccard distance [47], defined as

$$M_s(i,j) = \frac{D_{i,j}}{C_{i,j}}, \quad (2)$$

$$C(i,j) = \frac{|\{i\} \cap \{j\}|}{|\{i\} \cup \{j\}|},$$

$$D(i,j) = \frac{|\{i\} \cup \{j\}| - |\{i\} \cap \{j\}|}{|\{i\} \cup \{j\}|}.$$

$D_{i,j}$  ( $C_{i,j}$ ) is given by the number of different (common) non-bonded contacts between conformations given by the set of contacts  $\{i\}$  and  $\{j\}$ .  $M_s$  takes into account all the contacts whether they are native or not. Comparing  $M_s$  with other variables often used, the usual reaction coordinate  $Q(A)$  (given by the fraction of native contacts formed in conformation  $A$ ) cannot satisfy Eq.(1), since  $Q(A)$ , given a native reference  $N$ , is different from  $Q(N)$ , given a reference conformation  $A$ . Root Mean Square deviation (RMSD) satisfies the Eq.(1) conditions, but compares the overall conformation, which may not properly account for local details.

One could argue that this topological distance, which could capture static features of the conformation space, may not cope with details of folding. Folding process is an intrinsically dynamic process, which is also the basis of the the discontinuity graphs discussed in the Introduction. Moreover, two structurally similar conformations could differ in terms of the dynamics for folding. We therefore incorporated in the simulations a dynamic measurement defined by

$$M_d(i,j) = \min_{\{paths\}} n(i,j), \quad (3)$$

where  $n(i,j)$  is the number of local minimum intermediates required to go from  $i$  to  $j$  conformations.  $M_d(i,j)$  corresponds to the minimum calculated over all the paths going from  $i$  to  $j$  (or vice-versa). The measurement is normalized upon dividing by the largest distance encountered. This approach resembles the method using to determine kinetic transition networks [48–50]. In subsidiary simulations we noted that using an effective distance  $M_{ef}$  (in Eq.(4)), which takes into account the dynamic measurement, yields essentially the same results as with our initial measurement defined in Eq.2. Therefore the use of the latter appears to embed the underlying landscape of the system.

### Visualizing the folding funnel

The protein funnel was obtained by projecting the multidimensional local minima, distributed according to the effective metric distance, onto a 2D surface. The 5 sequences investigated, viz. A, Af, B, C and D, are described in detail in the Methods. Figure 1 shows the funnel representation of sequence A, in which the minima are colored according to conformation energy in Figure 1a, or according to the reaction coordinate  $Q$  in Figure 1b. The steep convergence to the native state either in energy or  $Q$  representation is an indicative of the principle of minimal frustration associated with this sequence. The important information is the relative distance between two given points, and the axes were removed because the directions do not have any special meaning. Different regions in the 2D representation can be associated with different partially folded motifs, as shown in Figure 1a. As expected, different time intervals sample different minima, thus yielding varying local minima resolution, but the overall funnel pattern was maintained (see Figures S3, S4, S5 and S6 in the Supporting Information). The pattern preservation for distinct time intervals (in MCs) ensures that the sequence possesses a unique “signature”, with clusters of conformations becoming denser as the number of time intervals decreases (probing more fluctuations). For a 30 MCs interval, in particular, a more refined energy distribution can be visualized with the identification of higher energy conformations when compared with local minima in simulations with larger time intervals.

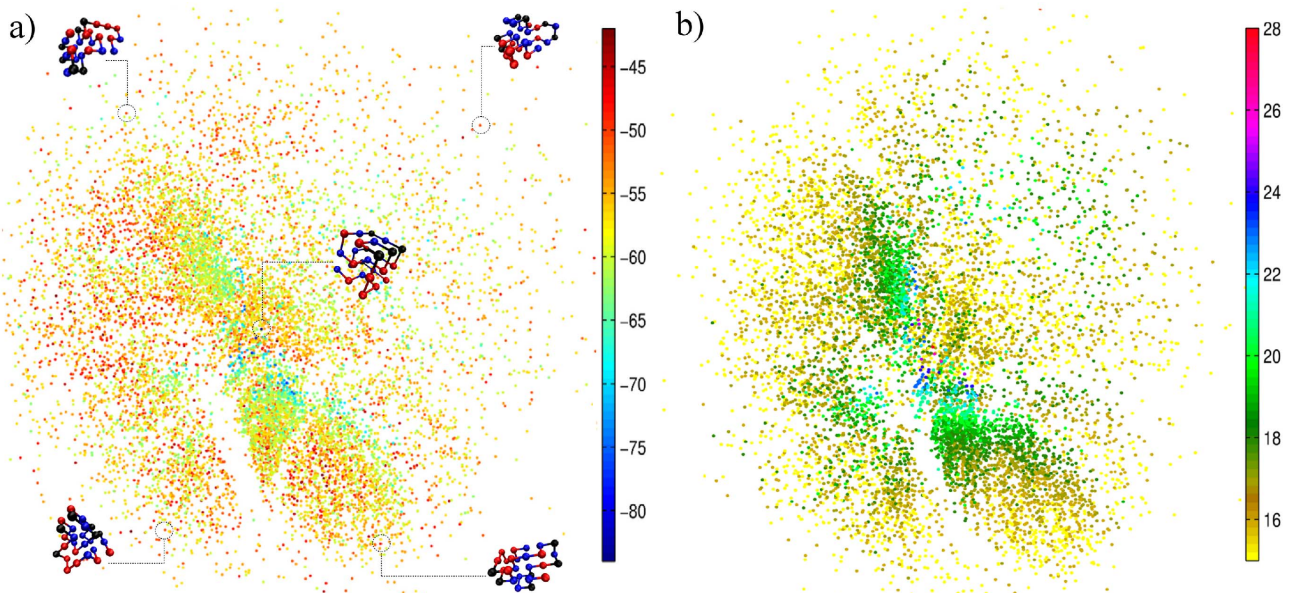
Figure 2 shows that the funnel landscape obviously depends on the protein sequence, with a unique native structure being represented by a unique funnel landscape. The sequence D, in particular, has a doubly degenerate native state, where the two lowest-energy conformations differ from each other by 5 native contacts. The existence of these two native states is reflected in two clusters of points in Figure 2d. For this sequence, a change from one region (native state) to the other native state requires unfolding (i.e. the need to move towards the periphery in the projection).

Note that, for sequences that are difficult to fold (Figure 2a and 2c), the number of conformations with intermediate energy (in the green light blue region) increases considerably, in comparison with the easily-foldable sequences (A and B) (Figure 1 and 2b). By the same token, the sequences with non-efficient folding funnels take a much longer average time to fold, as shown in Figure S2 in the Supporting Information.

In order to generate a 3D visualization for the funnel, the 2D representation was taken for the  $x$  and  $y$  axes, while the energy was taken as the  $z$  axis, with the lowest energy value corresponding to the native state. Color encodes the reaction coordinate  $Q$ , which is the degree of nativeness. Figure 3 shows the 3D picture of the funnel for the sequence A, while the figures for the other sequences are given in Figures S7 and S8 in the Supporting Information. It must be stressed that the result of the projection method is independent of the initial condition of the states in the 2D representation. The native conformation converges to the center of the funnel without any constraint or external force. The global minimum of the system, or native configuration, in the center of the 2D representation reinforces the funnel-like structure of the landscape.

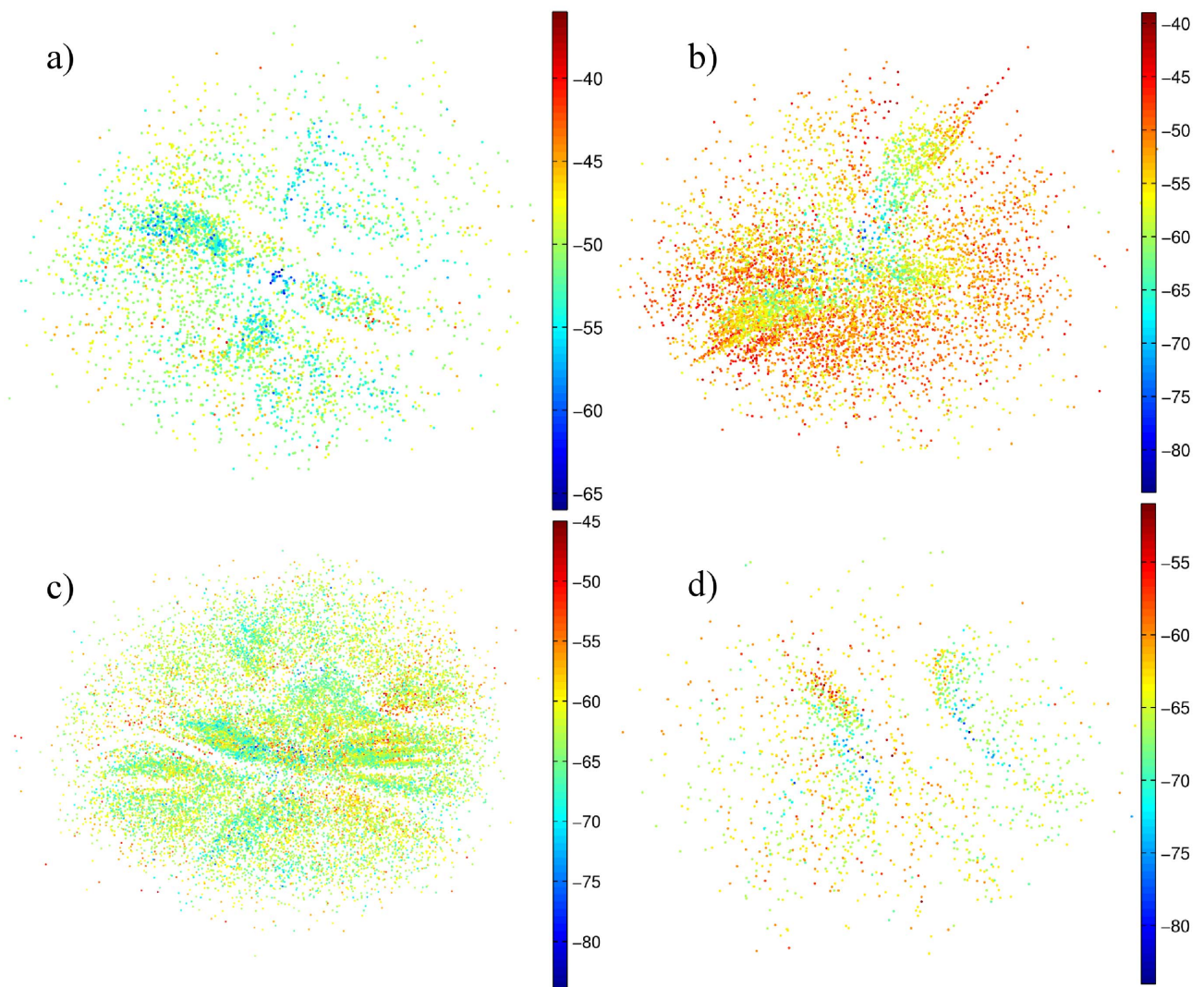
### Folding routes

The 2D and 3D visualizations of the folding funnels appear to confirm that the strategy proposed here is suitable for describing the folding process, but they do not suffice to ensure that the choice of the distance metrics is robust. The latter can be probed by analyzing the folding routes, for in a good funnel representation the folding route has to be represented by a sequence of small steps in the effective funnel representation. Figure 4a shows two routes generated from first passage time simulations, which show mostly small steps between successive minima. The details of this representation can be seen in different folding routes, which probe very distinct regions of the phase space (associated with different partially folded motifs). Also worth mentioning is that the



**Figure 1. Visualization in 2D of the conformation space for sequence A.** Each point represents one conformation (local minimum) and the distance between points refers to the projection of their effective distance. The axis directions do not have any special meaning and have been removed. In (a) the color is associated with the conformation energy. In (b) the color is associated with the reaction coordinate  $Q$ , where  $Q=28$  represents the native state.

doi:10.1371/journal.pone.0100861.g001



**Figure 2. 2D visualization for the sequences (a) Af; (b) B; (c) C and (d) D, obtained with a time interval of 1000 MCs.**  
doi:10.1371/journal.pone.0100861.g002

routes do not directly cross the empty regions, but go around them through neighboring connected states. Figure 4b shows that, for sequence A, the distances between two subsequent local minima in the 2D representation are almost always very small, which means that no drastic changes occur in conformation from one minimum to the next. This confirms the robustness of the approach presented here.

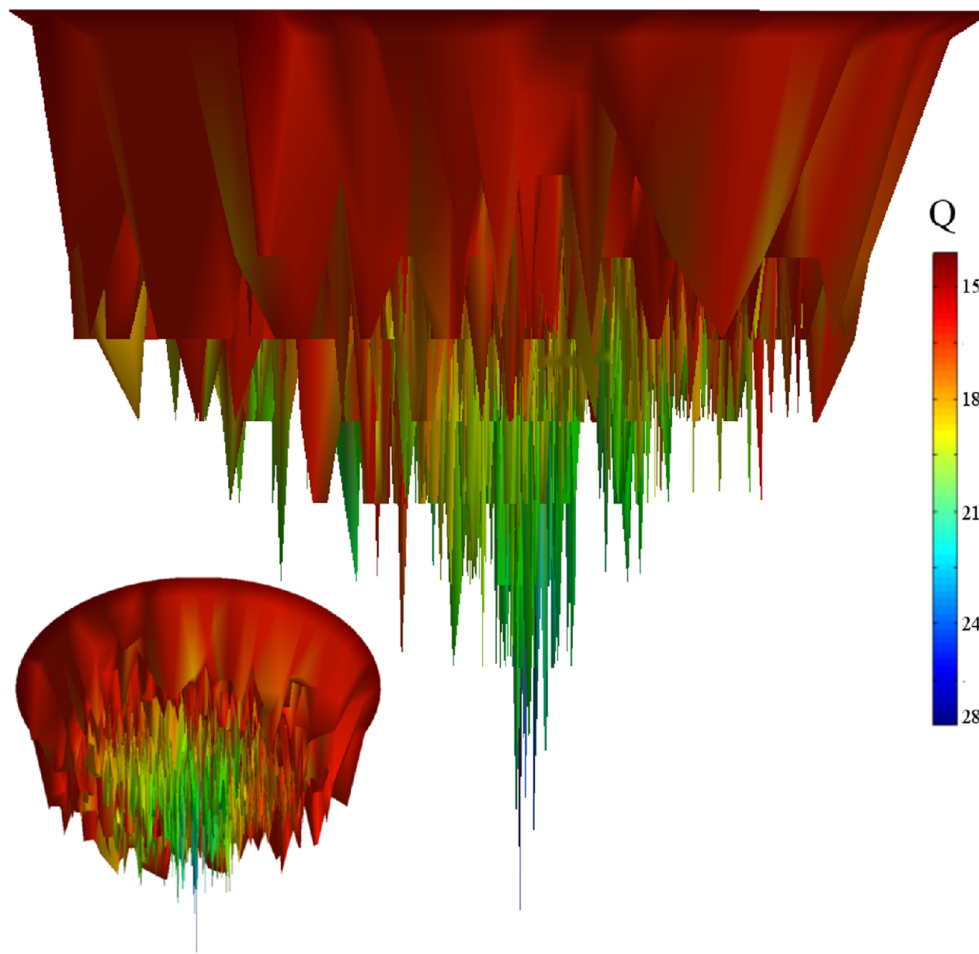
#### Analysis of a mutation

The 2D projection was also used to explore a mutation in sequence A, where two monomers were exchanged to yield a less stable sequence (see Table 1 in the Methods). The effects from the mutation can be evaluated by mapping the data of the two sequences in the same projection. Due to mutation a set of conformations is no longer energetically favorable for the folding. This can be seen in Figure 5a where the whole region on the left is missing for the mutated sequence (green points). One thousand (1000) folding routes were calculated for each sequence, with examples shown in Figures 5b and 5c. In contrast to the wild sequence (A), for the mutated sequence (Af) the routes normally

probe a significant part of conformational space before reaching the native state, with 95% of the pathways occurring on the right-hand part of the projection. The mutation stabilizes a different set of local minima, which hinders the folding process and causes a considerable increase in the average folding time (as seen in Figure S2). Note that most of the minima in the mutated sequence do not coincide with those of the wild sequence, thus indicating that they are structurally different, even though they have the same native state.

#### Conclusions

Visualization was based on the assumption that the distance between two conformations was the ratio between the Jaccard index and the Jaccard distance taking into account all non-bonded contact points. The suitability of the approach could be confirmed by comparing the funnels and folding routes for 5 sequences, where much larger folding times were estimated for sequences known to be difficult to fold. Furthermore, a doubly degenerate sequence yielded a funnel with two native states, as expected.



**Figure 3. 3D visualization of the funnel for sequence A, with two different views.** The third axis (depth) of the funnel is associated with the energy of the local minima, and the color map is the reaction coordinate  $Q$ .  
doi:10.1371/journal.pone.0100861.g003

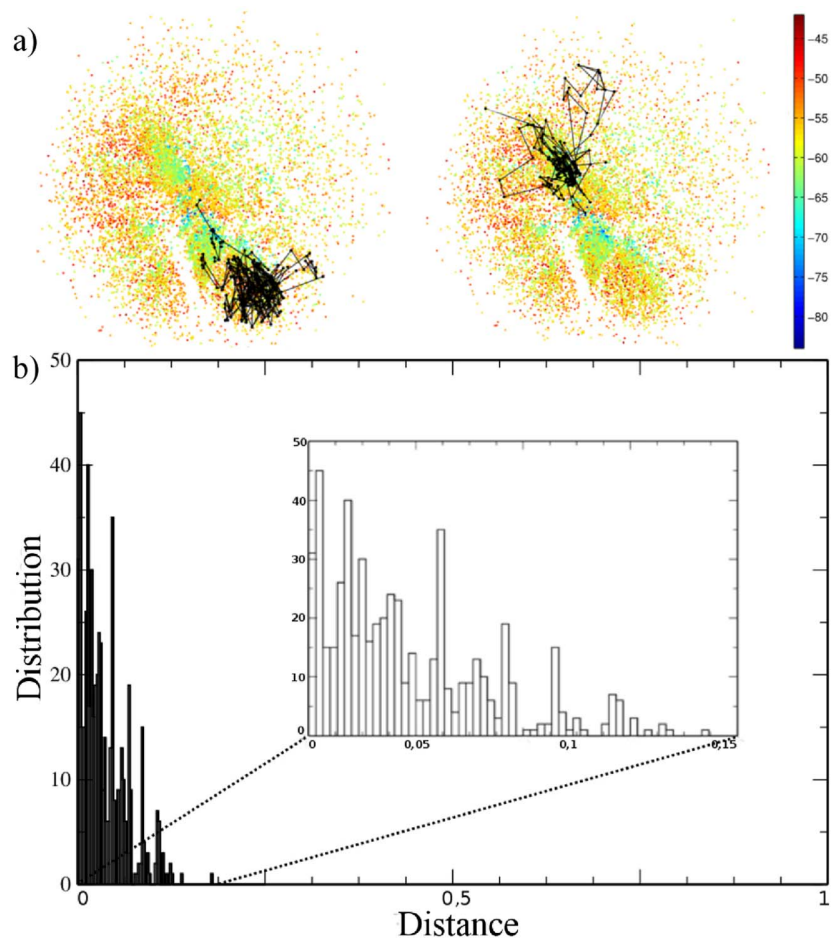
Since the methods employed are entirely generic, this approach is a potential tool to be used in association with other methods that efficiently probe the energy landscape, such as diffusion-map-directed MD (DM-d-MD) [51], disconnectivity graphs [20] and metadynamics [52]. The method was tested in a simple lattice model, in which the minima were sampled with variable time intervals. It will be straight forward to apply this methodology to realistic models and more meaningful sampling methods, such as those used by Wales [20,21,25]. In particular, our method may be helpful to probe details of folding trajectories and effects of mutation in the study of metastable states. As applications, previous work using disconnectivity graphs analyzed the potential energy landscapes of proteins involving gatekeeper residues [33,53,54]. By probing the gatekeeper residue contacts using our method we expect to be able to shed light into the nature of these peculiar conformational states.

## Methods

### Model

In this lattice model, a globular protein is modeled as a simplified heteropolymer made up of 27 monomers (or beads) covalently bonded. The monomers are placed on the vertices of a cubic lattice. These models are capable of accounting for several

features of protein folding [42], where the most compact (folded) structure is a  $3 \times 3 \times 3$  cube. One contact is defined for two monomers that are at nearest-neighbor distances but not connected covalently. In the lattice model the maximum number of contacts is 28. The energy of the system is given by  $E = n_l E_l + n_u E_u$ , where  $n_l$  is the number of (non-covalent) contacts of like monomers and  $n_u$  is the number of contacts between distinct monomers. The folding kinetics is performed with the Metropolis algorithm in a Monte Carlo simulation with typical motions in polymers [42]. Here we use a low hydrophobicity regime with  $E_l = -3$  and  $E_u = +3$  in arbitrary units. This regime was chosen to mimic the folding behavior where the sequence evolves toward its native state without going through a hydrophobic collapse [43,55]. Five sequences were chosen for the analysis, which exhibit very distinct features, as indicated in Table 1. For each conformation, the free energy was calculated as a function of the parameter  $Q$  (See Figure S1 in the Supporting Information). The data collected for the projection is restricted to conformations from around the transition state ( $Q_{TS} - 1$ ) to the native state ( $Q = 28$ ). The simulation temperature was set to  $1.1 T_f$ , in order for the conformational space to be visited as thoroughly as possible, thus avoiding the sequence having to spend long times in its native state. Local minima were obtained within time intervals segmented along the Monte Carlo trajectories. 4 time intervals



**Figure 4. Analysis of folding routes.** In (a) Folding routes for the sequence A, where the starting point was a random conformation and the final point corresponds to the native state. In (b) Histogram of the distribution of distances between two subsequent local minima in the 2D representation for very long trajectories.  
doi:10.1371/journal.pone.0100861.g004

were used: 30, 100, 300 and 1000 Monte Carlo steps (MCs). For each interval, the total time was set so that  $10^7$  minima were obtained. The conformation at each local minimum was stored in a  $27 \times 27$  binary matrix representing all the contacts. The conformational matrix is symmetrical and an element  $c_{ij}$  is 1 if there is a contact between monomers  $i$  and  $j$  and 0 otherwise.

### Metric

The projection of these multidimensional data was performed using a metric based on the conformational similarity (Jaccard index) and dissimilarity (Jaccard distance), referred to as the structural measurement:  $M_s$  (Eq. 2). We also tested a dynamic measurement in which the number of intermediate minima for going from one conformation to the other was taken into account. This latter metric was named dynamic measurement  $M_d$  (Eq. 3). Using these measurements one may calculate a normalized effective distance between any two conformations,

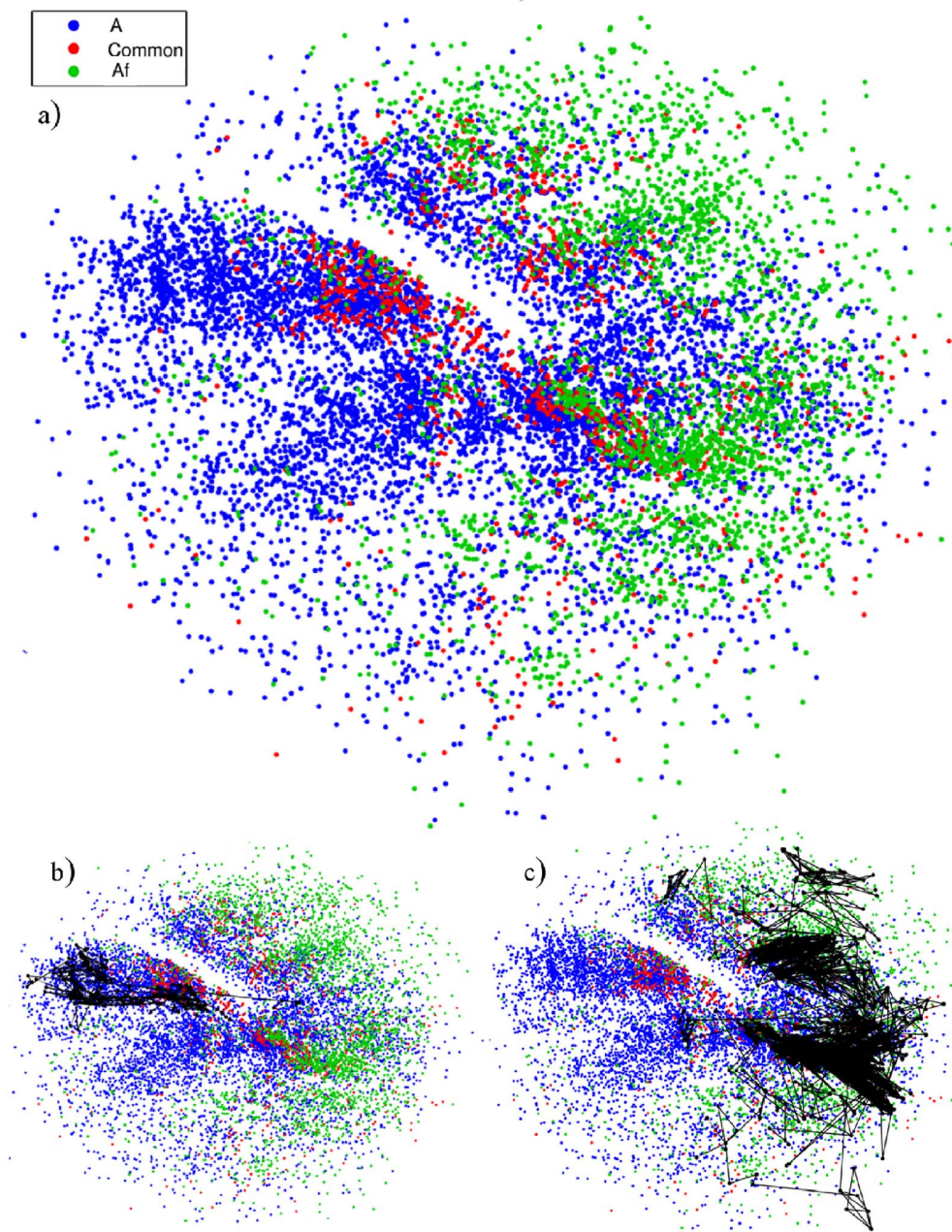
$$M_{ef}(i,j) = (1 + M_d(i,j)) M_s(i,j). \quad (4)$$

### Projection

Our goal is not to develop a technique for dimensionality reduction. We want to visualize the similarity between conformations according to our metric. Since the information of structures occurs in a multidimensional space, there is a need for projection into a lower dimension. As with any projection technique, we can create the projection in up to three dimensions [56]. The choice of two dimensions is simply for the ease of data interpretation. 3D projections are very difficult to interpret due to occlusions and overlaps which, in most cases, do not bring real gain compared to 2D [57].

The projection onto a 2D plot was made using the distance matrix with the Force-Scheme method [45], where the objects are initially placed in random positions, and then attraction and repulsion forces between the objects take the system to equilibrium according to a chosen heuristics. Here, the system was initialized with the conformation energies, which proved more efficient for convergence of the method. After the first placement of the objects, iterations within the Force-Scheme method are performed to preserve similarity in the original space into the projected space. In the first iteration, for each projected point  $y_i \in Y$ , (where  $Y$  is the input dataset) a vector is calculated  $\vec{v}_{i,j} = (y_j - y_i), \forall y_j \neq y_i$ . Then  $y_i$  is moved in the  $\vec{v}$  direction by a step  $\Delta$ , defined as:





**Figure 5. 2D Projection of sequence A (blue points) and its mutated form Af (green points), while the points in red are common to both A and Af sequences.** An example of a route for each of the sequences is presented: (b) sequence A and (c) sequence Af. doi:10.1371/journal.pone.0100861.g005

$$\Delta = \left[ (2k + 2)^{\frac{1}{2k+1}} - 1 \right], \quad (5)$$

where  $k$  is the number of previous iterations. After an iteration, each object should be moved closer to its similar ones until the system converges. The number of iterations may be defined arbitrarily or the scheme may be stopped when a threshold is reached. Here the process was stopped when the difference in distances for a given object between two consecutive iterations was below a threshold of  $10^{-4}$ . In order to build the 3D funnel, the

points in the 2D projection are shifted along a perpendicular axis according to their energies, thus generating a 3D structure where the lowest-energy states are placed on the bottom. We also performed tests with one of the most precise projection techniques in terms of distance preservations, referred to as Classical Multidimensional Scaling (MDS) [56]. The results were similar to those produced by the Force-Scheme in terms of distributing the points on the plane according to the similarity between conformations, with the final shape of the funnels also being very similar. The MDS technique, however, is much more costly in computational time, and in some cases ordinary microcomputers lack the power to obtain the funnels. Therefore, we opted for the

**Table 1.** Summary of sequences utilised.

Sequences	Zscore <sup>†</sup>	Representation	T <sub>f</sub>
A <sup>‡</sup>	6.75	ABABBBBCBACBABAABACACBACAACAB	1.89
Af <sup>§</sup>	5.91	ABABBBBCBACBABA <sup>^</sup> CACCBABAACAB	1.23
B	8.58	ABCDBEABBAEBDBCBAABCDBEAB	1.90
C	5.90	AAAAAABCAACBAABCAAACAAAAC	1.95
D	6.27	AAABAAAACABAAAABABACABAACABA	1.73

<sup>†</sup>Zscore is calculated according to methodology described by Dima et al. [58]. <sup>‡</sup>Sequence design by Shakhnovich et al. [59] which has been used in other studies [42,43]. <sup>§</sup>This sequence was obtained through a permutation of two monomers in A, which results in three frustrated contacts in the native structure.  
doi:10.1371/journal.pone.0100861.t001

Force-Scheme approach, which is much faster and allows one to process thousands of conformations in a few minutes with a simple PC.

## Supporting Information

**Figure S1 Free energy vs Native contacts (Q).** Free energy as a function of native contacts (Q) for four protein-like sequences A, Af, B and C. The simulation was performed at the folding transition temperature (T<sub>f</sub>).  
(TIF)

**Figure S2 Mean first-passage times.** Mean first-passage times as a function of the logarithm of the number of local minima needed to reach the native state. Note that the two proteins with high Zscore (A and B sequences), on average, fold more quickly. In contrast, in the sequences with a low Zscore (Af and C sequences), the number of conformations necessary to reach the native state is much greater.  
(TIF)

**Figure S3 Visualization in two dimensions for all time intervals for sequence A.** a) 30 MCs; b) MC 100; c) 300 MCs and d) 1000 MC.  
(TIF)

**Figure S4 Visualization in two dimensions for all time intervals of sequence Af.** a) 30 MCs; b) MC 100; c) 300 MCs and d) 1000 MC.  
(TIF)

**Figure S5 Visualization in two dimensions for all time intervals for sequence B.** a) 30 MCs; b) MC 100; c) 300 MCs and d) 1000 MC.  
(TIF)

**Figure S6 Visualization in two dimensions for all time intervals for sequence C.** a) 30 MCs; b) MC 100; c) 300 MCs and d) 1000 MC.  
(TIF)

**Figure S7 3D visualization of the funnel for sequence C.** A profile of the funnel is shown on the left, while details of the internal and external parts of the funnel are shown on the right.  
(TIF)

**Figure S8 3D visualization of the funnel for sequence D.** A profile of the funnel is shown on the left, while details of the internal and external parts of the funnel are shown on the right.  
(TIF)

## Acknowledgments

We thank Aline T. Bruni and Mariane L. Paiva for the initial study that motivated this work. We also thank Paul Whitford for helpful discussions and comments. This work was supported by FAPESP, CNPq, CAPES and nBioNet network (Brazil).

## Author Contributions

Conceived and designed the experiments: VBPL ABOJ. Performed the experiments: ABOJ FMF. Analyzed the data: ABOJ FMF FVP VBPL. Contributed reagents/materials/analysis tools: ABOJ FMF FVP. Wrote the paper: ABOJ VBPL ONOJ.

## References

- Anfinsen CB (1973) Principles that govern the folding of protein chains. *Science* (New York, NY) 181: 223–230.
- Levinthal C (1968) Are there pathways for protein folding? *Extrait du Journal de Chimie Physique* 65.
- Bryngelson JD, Onuchic JN, Socci ND, Wolynes PG (1995) Funnels, pathways, and the energy landscape of protein folding: A synthesis. *Proteins: Structure, Function, and Bioinformatics* 21: 167–195.
- Leopold PE, Montal M, Onuchic JN (1992) Protein folding funnels: a kinetic approach to the sequence-structure relationship. *Proceedings of the National Academy of Sciences of the United States of America* 89: 8721–8725.
- Thirumalai D, O'Brien EP, Morrison G, Hyeon C (2010) Theoretical perspectives on protein folding. *Annual Review of Biophysics* 39: 159–183.
- Dill KA, Ozkan SB, Shell MS, Weikl TR (2008) The protein folding problem. *Annual Review of Biophysics* 37: 289–316.
- Onuchic JN, Luthey-Schulten Z, Wolynes PG (1997) Theory of protein folding: the energy landscape perspective. *Annual review of physical chemistry* 48: 545–600.
- Klimov DK, Thirumalai D (1998) Linking rates of folding in lattice models of proteins with underlying thermodynamic characteristics. *The Journal of Chemical Physics* 109: 4119–4125.
- Sabelko J, Ervin J, Gruebele M (1999) Observation of strange kinetics in protein folding. *Proceedings of the National Academy of Sciences of the United States of America* 96: 6031–6036.
- Nymeyer H, Garcia AE, Onuchic JN (1998) Folding funnels and frustration in off-lattice minimalist protein landscapes. *Proceedings of the National Academy of Sciences* 95: 5921–5928.
- Onuchic JN, Nymeyer H, Garcia AE, Chahine J, Socci ND (2000) The energy landscape theory of protein folding: insights into folding mechanisms and scenarios. *Advances in protein chemistry* 53: 87–152.
- Schuler B, Lipman EA, Eaton WA (2002) Probing the free-energy surface for protein folding with single-molecule uorescence spectroscopy. *Nature* 419: 743–747.
- Lee CL, Stell G, Wang J (2003) First-passage time distribution and non-markovian diffusion dynamics of protein folding. *The Journal of Chemical Physics* 118: 959–968.
- Chavez LL, Onuchic JN, Clementi C (2004) Quantifying the roughness on the free energy landscape: entropic bottlenecks and protein folding rates. *Journal of the American Chemical Society* 126: 8426–8432.
- Wang J, Oliveira RJ, Chu X, Whitford PC, Chahine J, et al. (2012) Topography of funneled landscapes determines the thermodynamics and kinetics of protein folding. *Proceedings of the National Academy of Sciences* 109: 15763–15768.

16. Zhuravlev PI, Papoian GA (2010) Protein functional landscapes, dynamics, allostery: a tortuous path towards a universal theoretical framework. *Quarterly Reviews of Biophysics* 43: 295–332.
17. Potoyan DA, Papoian GA (2011) Energy landscape analyses of disordered histone tails reveal special organization of their conformational dynamics. *Journal of the American Chemical Society* 133: 7405–7415.
18. Izhaki LS, Otzen DE, Fersht AR (1995) The structure of the transition state for folding of chymotrypsin inhibitor 2 analysed by protein engineering methods: evidence for a nucleation-condensation mechanism for protein folding. *Journal of molecular biology* 254: 260–288.
19. Clementi C, Nymeyer H, Onuchic JN (2000) Topological and energetic factors: what determines the structural details of the transition state ensemble and “en-route” intermediates for protein folding? an investigation for small globular proteins. *Journal of molecular biology* 298: 937–953.
20. Wales DJ (2010) Energy landscapes: some new horizons. *Current Opinion in Structural Biology* 20: 3–10.
21. Wales D (2003) *Energy Landscapes: Applications to Clusters, Biomolecules and Glasses*. Cambridge University Press.
22. Shan Y, Arkhipov A, Kim ET, Pan AC, Shaw DE (2013) Transitions to catalytically inactive conformations in EGFR kinase. *Proceedings of the National Academy of Sciences of the United States of America* 110: 7270–7275.
23. Dobson CM (2003) Protein folding and misfolding. *Nature* 426: 884–890.
24. Reddy AS, Wang L, Singh S, Ling YL, Buchanan L, et al. (2010) Stable and metastable states of human amylin in solution. *Biophysical Journal* 99: 2208–2216.
25. Wales DJ (2012) Decoding the energy landscape: extracting structure, dynamics and thermodynamics. *Philosophical transactions Series A, Mathematical, physical, and engineering sciences* 370: 2877–2899.
26. Wales DJ, Bogdan TV (2006) Potential energy and free energy landscapes. *The Journal of Physical Chemistry B* 110: 20765–20776.
27. Becker OM, Karplus M (1997) The topology of multidimensional potential energy surfaces: Theory and application to peptide structure and kinetics. *The Journal of Chemical Physics* 106: 1495–1517.
28. Becker OM (1997) Quantitative visualization of a macromolecular potential energy “funnel”. *Journal of Molecular Structure: THEOCHEM* 398–399: 507–516.
29. Wales DJ, Miller MA, Walsh TR (1998) Archetypal energy landscapes. *Nature* 394: 758–760.
30. Miller MA, Doye JPK, Wales DJ (1999) Structural relaxation in atomic clusters: Master equation dynamics. *Physical Review E* 60: 3701–3718.
31. Doye J, Miller M, Wales D (1999) The double-funnel energy landscape of the 38-atom lennard-jones cluster. *The Journal of Chemical Physics* 110: 6896.
32. Doye J, Miller M, Wales D (1999) Evolution of the potential energy surface with size for lennardjones clusters. *The Journal of Chemical Physics* 111: 8417.
33. Wales DJ, Dewsbury PEJ (2004) Effect of salt bridges on the energy landscape of a model protein. *The Journal of Chemical Physics* 121: 10284–10290.
34. Miller MA, Wales DJ (1999) Energy landscape of a model protein. *The Journal of Chemical Physics* 111: 6610–6616.
35. Evans DA, Wales DJ (2003) The free energy landscape and dynamics of met-enkephalin. *The Journal of Chemical Physics* 119: 9947–9955.
36. Krivov SV, Karplus M (2002) Free energy disconnectivity graphs: Application to peptide models. *The Journal of Chemical Physics* 117: 10894–10903.
37. Noé F, Fischer S (2008) Transition networks for modeling the kinetics of conformational change in macromolecules. *Current Opinion in Structural Biology* 18: 154–162.
38. Prada-Gracia D, Gómez-Gardeñes J, Echenique P, Falo F (2009) Exploring the free energy landscape: From dynamics to networks and back. *PLoS Comput Biol* 5: e1000415.
39. Noé F, Horenko I, Schutte C, Smith J (2007) Hierarchical analysis of conformational dynamics in biomolecules: Transition networks of metastable states. *The Journal of Chemical Physics* 126.
40. Rao F, Caisch A (2004) The protein folding network. *Journal of molecular biology* 342: 299–306.
41. Dickson A, Brooks CL (2013) Native states of fast-folding proteins are kinetic traps. *Journal of the American Chemical Society* 135: 4729–4734.
42. Succi ND, Onuchic JN (1995) Kinetic and thermodynamic analysis of proteinlike heteropolymers: Monte carlo histogram technique. *The Journal of Chemical Physics* 103: 4732–4744.
43. Succi ND, Onuchic JN, Wolynes PG (1998) Protein folding mechanisms and the multidimensional folding funnel. *Proteins* 32: 136–158.
44. Garstecki P, Hoang TX, Cieplak M (1999) Energy landscapes, supergraphs, and folding funnel in spin systems. *Physical Review E* 60: 3219–3226.
45. Tejada E, Minghim R, Nonato LG (2003) On improved projection techniques to support visual exploration of multi-dimensional data sets. *Information Visualization* 2: 218–231.
46. Choi S, Cha S, Tappert C (2010) A survey of binary similarity and distance measures. *Journal on Systemics, Cybernetics and Informatics* 8: 43–48.
47. Tan PN, Steinbach M, Kumar V (2005) *Introduction to data mining*. Boston: Pearson Addison Wesley.
48. Wales DJ (2006) Energy landscapes: calculating pathways and rates. *International Reviews in Physical Chemistry* 25: 237–282.
49. Wales DJ (2002) Discrete path sampling. *Molecular Physics* 100: 3285–3305.
50. Wales DJ (2004) Some further applications of discrete path sampling to cluster isomerization. *Molecular Physics* 102: 891–908.
51. Zheng W, Rohrdanz MA, Clementi C (2013) Rapid exploration of configuration space with diffusion-map-directed molecular dynamics. *The journal of physical chemistry B*.
52. Laio A, Parrinello M (2002) Escaping free-energy minima. *Proceedings of the National Academy of Sciences* 99: 12562–12566.
53. Otzen DE, Oliveberg M (1999) Salt-induced detour through compact regions of the protein folding landscape. *Proceedings of the National Academy of Sciences* 96: 11746–11751.
54. Kurnik M, Hedberg L, Danielsson J, Oliveberg M (2012) Folding without charges. *Proceedings of the National Academy of Sciences* 109: 5705–5710.
55. Chahine J, Nymeyer H, Leite VBP, Succi ND, Onuchic JN (2002) Specific and nonspecific collapse in protein folding funnels. *Physical review letters* 88: 168101.
56. Cox TF, Cox MAA (2010) *Multidimensional Scaling, Second Edition*. CRC Press.
57. Ware C (2001) Designing with a 2 1/2d attitude. *Information Design Journal* 10: 2001.
58. Dima RI, Banavar JR, Cieplak M, Maritan A (1999) Statistical mechanics of protein-like heteropolymers. *Proceedings of the National Academy of Sciences* 96: 4904–4907.
59. Abkevich VI, Gutin AM, Shakhnovich EI (1994) Free energy landscape for protein folding kinetics: Intermediates, traps, and multiple pathways in theory and lattice model simulations. *The Journal of Chemical Physics* 101: 6052–6062.

# Apêndice B

## Método do histograma

Ao estudar um sistema deseja-se conhecer os valores das variáveis termodinâmicas não apenas em uma dada temperatura, mas como uma função da temperatura. O procedimento descrito pelo algoritmo de Metropolis se realiza com valor constante da temperatura. Por isso, sua aplicação direta no cálculo dos valores médio das variáveis exige tantas repetições do procedimento de simulação quantos forem os valores da temperatura para os quais deseja-se as referidas médias.

Um método baseado em uma ideia de Valleau e Card, mas desenvolvida principalmente por Ferrenberg e Swendsen [5, 76], se tornou conhecido como o "método do histograma". A ideia é obter as grandezas termodinâmicas de outras temperaturas a partir de uma única simulação em uma determinada temperatura.

A partir dessa técnica pode-se calcular uma densidade de estados aproximada do sistema com o qual é possível calcular qualquer grandeza termodinâmica para uma faixa de temperatura.

Para cadeias extremamente curtas em 2D é possível enumerar todas as conformações e obter a função de partição com a qual se calcula qualquer grandeza termodinâmica. Já para uma cadeia de 27 monômeros em uma rede cúbica é impossível enumerar todas as conformações, somente é possível enumerar todas as conformações maximamente compactadas. Então, o método do histograma é utilizado para calcular a densidade de estados. para isso, enumera-se quantas vezes ocorreu uma determinada energia na simulação e constrói-se um histograma de energia. O histograma de energia  $h(E, T')$  mede a probabilidade da energia  $E$  ocorrer na temperatura  $T_0$ , que é igual a média

## Método do histograma

---

térmica da densidade de estados, uma vez que o sistema é considerado um ensemble canônico com:

$$h(E, T') = \frac{n(E)e^{-\frac{E}{T'}}}{Z(T')} \quad (\text{B.1})$$

em que  $Z(T')$  é a função de partição na temperatura  $T'$ , que é dada por:

$$Z(T') = \sum_E n(E)e^{-\frac{E}{T'}} \quad (\text{B.2})$$

na qual,  $n(E)$  é a densidade de estados na energia  $E$  (número de conformação com energia  $E$ ).  $K_b = 1$  e  $T_0$  é a temperatura da simulação.

Rearranjando a equação B.1 obtém a densidade de estados:

$$n(E) = h(E, T')e^{-\frac{E}{T'}}Z(T') \quad (\text{B.3})$$

na qual a partição  $Z(T')$  é uma constante que deve ser calculada. Para o sistema em estudo, é possível calcular o  $Z(T')$  e obter a densidade de estados. Para isso, é necessário conhecer a multiplicidade de algum estado. A sequência estudada possui um estado não degenerado, o estado fundamental. isso significa que  $n(E_{gs}) = 1$  em que  $E_{gs}$  é a energia do estado de menor energia do sistema. Com  $Z(T')$  determinado, é possível encontrar a densidade de estados  $n(E)$  e calcular uma quantidade extensiva, como a energia livre utilizando a função de partição(Equação B.2) e a equação abaixo:

$$F(T) = -T \log(Z) \quad (\text{B.4})$$

Para o cálculo de quantidades intensivas, médias térmicas são determinadas através de:

$$\langle \vartheta \rangle(T) = \frac{\sum_E \vartheta(E)n(E)e^{-\frac{E}{T}}}{\sum_E n(E)e^{-\frac{E}{T}}} \quad (\text{B.5})$$

Rearranjando as equações B.3 e B.5 se obtém:

---


$$\langle \vartheta \rangle(T) = \frac{\sum_E \vartheta(E) h(E, T') e^{-\frac{E}{T} + \frac{E}{T'}}}{\sum_E h(E, T') e^{-\frac{E}{T} + \frac{E}{T'}}} \quad (\text{B.6})$$

A equação B.6 deve ser usada sobre certa faixa de temperatura. Em temperaturas muito maiores ou muito menores que a temperatura de simulação os erros nos cálculos da densidade de estados (Equação B.3) tornam-se significativos. O sistema é amostrado para uma dada região do espaço de fase em uma dada temperatura. Para simulações onde a temperatura é muito alta, estados de mais baixa energia não são visitados e o espaço de fase não é amostrado. Para baixas temperaturas de simulação, estados com altas energias nunca são visitados.

Dessa forma, a densidade de estados não estará correta para regiões do espaço de fase não amostrado (é zero para regiões nunca visitadas). É por isso que existe um intervalo de temperatura no qual é possível fazer extrapolações para uma simulação.