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

ANTONIO BENTO DE OLIVEIRA JUNIOR

VISUALIZAÇÃO DO FUNIL DE ENOVELAMENTO DE PROTEÍNAS

SÃO JOSÉ DO RIO PRETO - SÃO PAULO
2013

ANTONIO BENTO DE OLIVEIRA JUNIOR

VISUALIZAÇÃO DO FUNIL DE ENOVELAMENTO DE PROTEÍNAS

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

Orientador: Prof. Dr. Vitor B. Pereira Leite

São José do Rio Preto, 25 de Julho de 2013

Oliveira Junior, Antonio Bento de.
Visualização do funil de enovelamento de proteínas / Antonio Bento de Oliveira Junior. -- São José do Rio Preto, 2013
51 f. : il.

Orientador: Vitor B. Pereira Leite
Dissertação (mestrado) ï Universidade Estadual Paulista ñJúlio de Mesquita Filhoò, Instituto de Biociências, Letras e Ciências Exatas

1. Biologia molecular. 2. Biofísica. 3. Enovelamento de proteína. 4. Método de Monte Carlo. 5. Superfícies de energia potencial. I. Leite, Vitor Barbanti Pereira. II. Universidade Estadual Paulista "Júlio de Mesquita Filho". Instituto de Biociências, Letras e Ciências Exatas. III. Título.

CDU ï 577.11

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

ANTONIO BENTO DE OLIVEIRA JUNIOR

VISUALIZAÇÃO DO FUNIL DE ENOVELAMENTO DE PROTEÍNAS

Dissertação apresentada para obtenção do titulo de Mestre em Biofísica Molecular, àrea de Bíofisica 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.

BANCA EXAMINADORA

Prof. Dr. Vitor B. Pereira Leite
Professor Livre Docente
UNESP - São José do Rio Preto - SP
Orientador

Prof. Dr. Laurent Emmanuel Dardenne
Tecnologista Pleno
LNCC - Petrópolis - RJ

Prof. Dr. Sidney Jurado de Carvalho
Professor Assistente Doutor
UNESP - São José do Rio Preto - SP

São José do Rio Preto, 25 de Julho de 2013

Agradecimentos

À galera de pós-graduação, principalmente a meus amigos Alexandre, Daniel, Davi, Guilherme, Flávio, Pedro, Vinícius e Vinícius "Goias" que me proporcionaram muitos momentos maravilhosos, desde a filosofias das "cocas diárias", com discussões que passavam por: política, ciência e religião, até futebol, músicas e outras coisas... Me ensinaram que, mesmo com personalidades totalmente diferentes, pode-se contruir amizades para a vida toda.

Gostaria de agradacer a Deus, por ter me dado saúde e força para percorrer esse caminho. E me dado paciência e tranquilidade para passar por todas as dificuldades que encontrei e sabedoria pra aprender com meus erros.

Ao meu Oritentador, Prof. Dr. Vitor Barbanti Pereira 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.

À minha namorada/esposa Ana Carolina, pelo amor, carinho e por todos os momentos em que passamos juntos, pela sua paciência e ajuda nos momentos difíceis. Gostaria também de agradecer aos meus pais, Antonio e Roseli e a minha irmã Lidiana, por todo o apoio e também por me ensinarem a dar valor nas coisas importantes da vida. Muito obrigado por tudo.

Aos meus professores da graduação e pós-graduação, em especial ao professor Augusto Agostinho Neto (in memoriam) pelo seus ensinamentos, tanto dentro como fora da sala de aula, e pela boa convivência durante meus anos no departamento, que o senhor descanse em paz jovem! Agradeço aos professores Alexandre Suman de Araujo e Jorge Chahine por participarem da minha banca de Qualificação, me dando conselhos para o término deste trabalho. Agradeço também a todos os funcionarios do departamento de Física, por sempre estarem dispostos a ajudar.

Agradeço aos alunos do grupo, o Vinicius, Thiago, Laís, Paulo, pelas opniões e sugestões que contribuiram para o desenvolvimento desse trabalho

À todos vocês, meu muitíssimo obrigado!

"Let me tell you something you already know. The world ain't all sunshine and rainbows. It's a very mean and nasty place and I don't care how tough you are it will beat you to your knees and keep you there permanently if you let it. You, me, or nobody is gonna hit as hard as life. But it ain't about how hard ya hit. It's about how hard you can get hit and keep moving forward. How much you can take and keep moving forward. That's how winning is done!"

Sylvester Stallone, Rocky Balboa

Resumo

O enovelamento de proteínas é um problema fundamental em Biofísica Molecular. A teoria aceita, conhecida como "energy landscape", utiliza o funil de energia potencial como conceito fundamental para o entendimento do enovelamento de proteínas. Este funil ocorre em uma superfície multidimensional de difícil visualização. A investigação de métodos para analisar quantitativamente a estrutura desse funil é importante para o completo entendimento do problema. Neste trabalho são apresentados meios de fazer a visualização desses funis de enovelamento de proteínas para o modelo de rede cúbica $3 \times 3 \times 3$. A partir de simulações do enovelamento de proteínas são calculados as distâncias entre mínimos locais por meio de uma métrica efetiva, onde considera-se os contatos não covalentes feitos em cada conformação. Esta análise é restrita para conformações próximas ao estado nativo. Técnicas de visualização e minimização são usadas para mapear o processo do enovelamento em um espaço de fase de menor dimensionalidade. Por meio desta visualização é possível analisar o enovelamento com detalhes, como a conectividade entre conformações, os diferentes caminhos para se atingir o estado nativo e regiões onde a proteína pode ficar armadilhada. Para este trabalho, utilizou-se cinco proteínas distintas, sendo duas altamente estáveis, duas que possuem baixa estabilidade e uma quinta que tem o estado nativo degenerado. A visualização dos funis se mostraram bastantes distintas, sendo possível notar um padrão para cada proteína mesmo quando variado alguns parâmetros. Tais resultados são consistentes com as ideias associadas à teoria do funil de enovelamento de proteínas.

Palavras-chave: Enovelamento de Proteína, Modelo de Rede Cúbica, Redução Multidimensional.

Abstract

The protein folding is a fundamental problem in molecular biophysics. The accepted theory, known as energy landscape, uses the funnel potential energy as a fundamental concept to understand the protein folding problem. The energy funnel occurs in a multi-dimensional surface, which is difficult to be visualized. The investigation of methods for a quantitative analysis of the funnel structure is important for complete understanding of the problem. In this work, ways for visualize the protein folding funnels in a $3 \times 3 \times 3$ lattice models are presented. Protein folding simulations are carried out. Distances between conformations are determined by the non-covalent contacts and defined by effective metric of the structural configuration. The analysis is restricted to conformations close to the native state, i.e., beyond the transition state. Computer minimization and visualization techniques were used to map the dynamics of the folding process into a lower dimensionality phase space, and then represent the folding funnel in two and three-dimensional surface. These techniques are applied to five distinct sequences, which two are highly stable, two marginally stable and the last has a native degenerated state. Their folding funnels are very distinct, where each sequence has a signature even when some parameters varied. These results are consistent with the ideas of the theory of protein folding funnel.

Keywords: Protein Folding, Lattice Model, Multidimensional Reduction

Lista de Figuras

1.1 Esquema do funil de enovelamento de proteínas. 15

2.1 Modelo de rede cúbica $3 \times 3 \times 3$ com 27 monômeros. 18

2.2 Movimentos possíveis da cadeia. 20

2.3 Movimentos possíveis da cadeia. 21

2.4 Gráfico esquemático da Energia livre *versus* Q 25

2.5 Representação esquemática os dos intervalos de tempo. 26

3.1 Inicialização do sistema para a projeção em duas dimensões. 31

4.1 Visualização em duas dimensões da sequência 0012 para o intervalo de 100 MCs. 34

4.2 Visualização em duas dimensões, da sequência 0012 para todos os intervalos de tempo considerados neste trabalho. *i)* 30 MCs; *ii)* 100 Mcs; *iii)* 300 MCs; *iv)* 1000 MCs. 35

4.3 Visualização em duas dimensões, da sequência 0012f para todos os intervalos de tempo considerados neste trabalho. *i)* 30 MCs; *ii)* 100 Mcs; *iii)* 300 MCs; *iv)* 1000 MCs. 36

4.4 Visualização em duas dimensões, da sequência 2221 para todos os intervalos de tempo considerados neste trabalho. *i)* 30 MCs; *ii)* 100 Mcs; *iii)* 300 MCs; *iv)* 1000 MCs. 36

4.5 Visualização em duas dimensões, da sequência 43157 para todos os intervalos de tempo considerados neste trabalho. *i)* 30 MCs; *ii)* 100 Mcs; *iii)* 300 MCs; *iv)* 1000 MCs. 37

4.6 Visualização, em duas dimensões, da sequência 45568D para o intervalode tempo de 100 MCs. 38

4.7 Rotas de enovelamento da sequência 0012. 39

4.8 Histograma das distâncias percorridas pelas rotas da sequência 0012. . . . 40

4.9 Gráfico da distribuição *versus* $log_{10}(x)$. 41

4.10 Visualização, em três dimensões, da sequência 0012. 42

4.11 Visualização, em três dimensões, da sequência 2221. 43

4.12 Visualização, em três dimensões, da sequência 45568D. 44

4.13 Projeção, em duas dimensões, da sequência 0012 e sua mutação 0012f. . . . 46

Lista de Tabelas

4.1 Proteínas exploradas nesse trabalho. A sequência 0012 projetada por Socci *et.al* [30] é uma sequência estável (de acordo com seu Z_{Score}), utilizada em diversos trabalhos anteriores [33, 38, 40]; A sequência 00012 é uma pequena mutação da sequencia 0012, tal sequência apresenta 3 contatos frustrados em sua estrutura nativa [38]; As sequências 43157 e 2221 foram escolhidas devido à sua estabilidade (sendo a 43157 estável e a 2221 pouco estável (de acordo com o seu Z_{Score})); A 45568D tem por característica a degenerescência do estado nativo igual a dois, ou seja, existem duas conformações com a menor energia. Estas duas conformações diferem-se entre si por 5 contatos. As cores representa os tipos de monômeros. 32

Sumário

1 Introdução **12**

1.1 O funil de enovelamento de proteínas 14

1.2 Motivação e Objetivos 15

2 Modelo **17**

2.1 Modelos minimalistas 17

2.2 Modelo de rede cúbica 18

2.2.1 Parâmetros de hidrofobicidade 19

2.2.2 Movimentos da cadeia 21

2.3 Método de Monte Carlo 21

2.3.1 Método de Monte Carlo aplicado ao modelo 23

2.4 Matriz de contatos, parâmetro de ordem Q e Z_{score} 23

2.5 Simulação computacional 24

3 Visualização **27**

4 Resultados e Discussões **32**

4.1 Visualização em duas dimensões 33

4.1.1 Rotas de enovelamento 38

4.2 Visualização em três dimensões 42

4.2.1 Sequência 0012 42

4.2.2 Sequência 2221 43

4.2.3 Sequência 45568D 44

4.3 Análise de uma mutação 45

5 Conclusões **47**

Referências Bibliográficas **49**

A Método do histograma **53**

Capítulo 1

Introdução

As proteínas pertencem à classe de macromoléculas que desempenham papel fundamental na fisiologia dos seres vivos. Elas são definidas como cadeias peptídicas compostas, em geral, por vinte aminoácidos distintos disponíveis na natureza. O correto desempenho de suas funções está diretamente relacionado com sua estrutura terciária e/ou quaternária que, por sua vez, é dependente da sua estrutura primária, ou seja, da sequência de aminoácidos que a compõe [1]. O processo que leva uma proteína a alcançar sua configuração nativa, tornando-a apta a desempenhar o seu papel biológico é conhecido como enovelamento e ainda não é completamente entendido. Este fato proporciona grande interesse quanto ao entendimento dos processos envolvidos, sendo um dos problemas fundamentais na Biofísica Molecular [2].

Os estudos de Anfinsen [3] 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 enovela, 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 nível 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 os mecanismos envolvidos neste processo biológico e desvendar quais as forças que governam o enovelamento.

Em 1968, Cyrus Levinthal [4] argumentou sobre a forma que a proteína encontra o estado nativo, e mostrou por meio de um exemplo que esta procura não poderia ser de forma aleatória. O argumento de Levinthal foi a 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 picosegundo (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 $1,4.10^{10}$ 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 microsegundos. 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 especifico, 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 inicio 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. Uma das contribuições mais importantes da Física para compreensão deste fenômeno foi, precisamente, a de reconciliar as perspectivas de Anfinsen e Levinthal no quadro de uma teoria unificada, Esta teoria unificada se baseia na natureza estatística do processo conhecido como funil de energia, introduzida por Onuchic, Wolynes e colaboradores [5–7], no qual se observa a superfície de energia global da proteína (*energy landscape theory*) [8].

1.1 O funil de enovelamento de proteínas

O estudo das superfícies de energia revelam os princípios gerais que regem o processo de enovelamento. Esta teoria tem fornecido um plano de fundo para a interpretação de experimentos sobre o processo de enovelamento, tanto qualitativo como quantitativamente [9–20].

A superfície de energia, quando vista sob a perspectiva de uma coordenada de reação, apresenta uma topologia afunilada com um gradiente de energia direcionado para a região do estado enovelado, Além disso, existem muitos caminhos que levam ao estado nativo, como se pode verificar na Figura 1.1.

As ideias sobre o funil de enovelamento são utilizadas para o entendimento do mecanismo cinético de rearranjo das estruturas das proteínas durantes o enovelamento. Para isso, considera-se as seguintes propriedades:

- Proteínas se enovelam à partir de um estado aleatório;
- O enovelamento ocorre de maneira difusiva e segue uma queda de altas energias para baixas energias conformacionais;

Sendo assim, Leopold et al.[5] define o funil de enovelamento como uma coleção de estruturas geometricamente similares, uma das quais é termodinamicamente estável em relação as outras. Para uma superfície de energia desse tipo, existem muitas estruturas com altas energias e poucas com baixas energias. Quanto mais próximo o sistema se

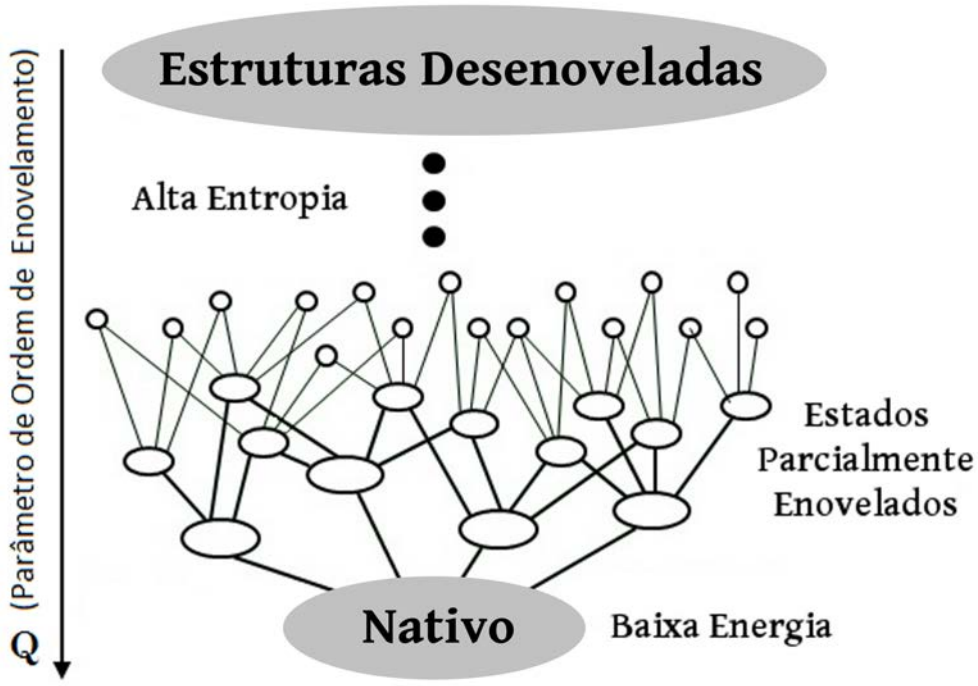


Figura 1.1: Esquema do funil de enovelamento de proteínas. A coordenada de reação Q está associada a proximidade do estado nativo, ou seja, para pequenos valores de Q, a proteína está desenovelada, enquanto que o estado nativo é encontrado no maior valor de Q.

encontra do estado nativo, menor é a energia dessas estruturas. Uma proteína em um estado desenovelado encontra seu estado nativo dinamicamente por meio dos diversos caminhos presentes neste funil de energia.

1.2 Motivação e Objetivos

Os funis de enovelamento teorizados por Onichic, Wolynes e colaboradores apresentam características importantes do processo de enovelamento, conforme indicado na seção anterior, sendo de grande valia a sua visualização. No entanto, a grande dificuldade encontrada para realizar a visualização desses funis é 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 de métodos que permitam a análise quantitativa da estrutura desses mínimos e a compreensão total do problema, podendo-se, a partir dessa análise, estudar características do enovelamento de macromoléculas. Muitos trabalhos vêm sendo desenvolvidos afim de

caracterizar a superfície de energia. Técnicas como PCA (Principal Compoent Analisys) [21], hierarchical model of networks [22], Markov State Models [23–25] e disconnectivity graphs [26, 27] buscam caracterizar a superficie de energia e obter informações sobre os processo de enovelamento.

O modelo de proteínas em rede objetiva reduzir a grande complexidade do problema do enovelamento ao maior grau possível de simplificação, mantendo, entretanto, os aspectos relevantes para a compreensão dos mecanismos do enovelamento. Apesar da simplificação, foi a partir desse modelo que surgiram as primeiras ideias para a teoria da superfície de energia [5].

Visando reduzir a complexidade do problema, neste trabalho é utilizado um modelo de rede cúbica de 27 monômeros. O potencial utilizado é o de interação de contato entre monômeros vizinhos próximos na rede. Para obter o mapeamento do funil de enovelamento é necessário um mapa de conexões entre as diferentes configurações de mínimos locais. Esse mapeamento é obtido por meio de técnicas de visualização e projeção multidimensional.

Os objetivos gerais do trabalho são:

- Implementações no código responsável por simular uma proteína em modelo de rede cúbica para obtenção de dados necessários para realizar oa projeção;
- Realizar o mapeamento da conectividade entre mínimos locais;
- Desenvolver uma métrica que relacione as semelhanças e as diferenças entre conformações;
- Aplicar um método para efetuar a projeção dos dados multidimensionais em duas e em três dimensões.

Capítulo 2

Modelo

2.1 Modelos minimalistas

Uma das grandes dificuldades no tratamento científico do problema do enovelamento de proteínas reside no fato da irredutibilidade dos sistemas que descrevem a cadeia polipeptídica. De fato, múltiplos ingredientes estão envolvidos neste cenário, como as interações químicas entre os átomos que compõe a cadeia, as interações estéricas decorrentes das diferentes formas e dos tamanhos distintos dos aminoácidos e as questões de unicidade conformacional, uma vez que, no seu estado nativo, a proteína não possui somente uma estrutura, mas uma coleção de estruturas bem definidas, com alto grau de similaridade em equilíbrio dinâmico.

Devido à complexidade do problema do enovelamento proteico e do custo computacional necessário para fazer as simulações, o uso de modelos minimalistas tem sido largamente empregado [28–30]. Tais modelos são capazes de reproduzir aspectos característicos do enovelamento como, por exemplo o tempo de enovelamento, a identificação dos caminhos que levam à conformação nativa, e a descriç ao de propriedades cinéticas e termodinâmicas do enovelamento [31–36].

2.2 Modelo de rede cúbica

No modelo de rede cúbica, uma proteína globular é reduzida para um heteropolímero simplificado composto por 27 monômeros dispostos em uma rede cúbica tridimensional. Neste trabalho, os monômeros representam os aminoácidos das proteínas e estão ligados covalentemente ao longo da cadeia, onde cada monômero pode ocupar somente um sítio na rede (condição de volume excluso)[5].

A estrutura maximamente compacta (enovelada) é um cubo de tamanho $3 \times 3 \times 3$, com um número máximo de contatos igual a 28. Um contato é definido quando dois monômeros estão na distância de primeiro vizinho e não estão ligados covalentemente, como ilustrado pelas setas da Figura 2.1.

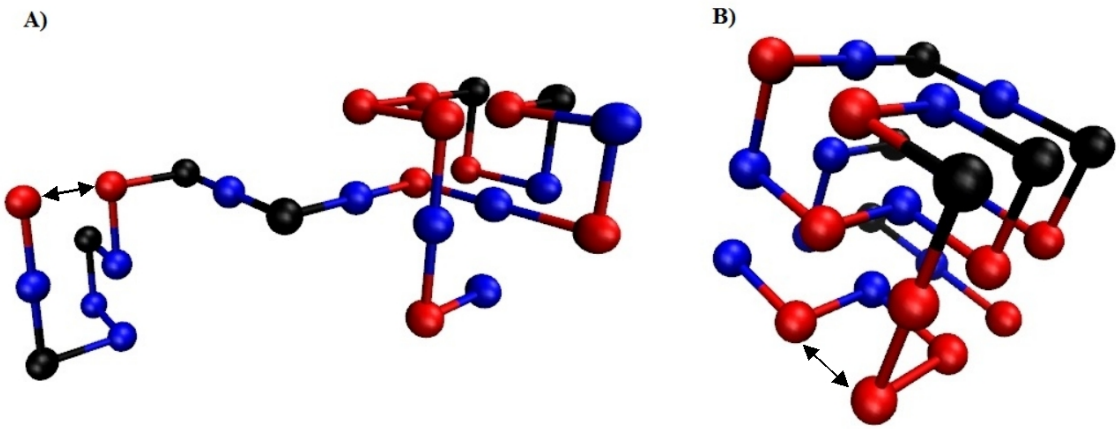


Figura 2.1: Modelo de rede cúbica $3 \times 3 \times 3$ com 27 monômeros. Cada cor representa um tipo de monômero e as setas indicam a formação de contatos não covalentes. A) Configuração inicial aleatória. B) Configuração nativa, formando os 28 contatos.

O principal efeito que governa o mecanismo do enovelamento proteico é o efeito hidrofóbico, decorrente das interações entre diferentes cadeias laterais hidrofóbicas das proteínas e entre essas cadeias laterais e o solvente. Esse efeito faz com que a cadeia sofra um colapso, causando a formação de um núcleo hidrofóbico.

Para esse trabalho, esse efeito foi modelado escolhendo um potencial que favoreça a formação de contatos entre quaisquer dois monômeros da cadeia. Com a finalidade de assegurar que a proteína tenha um único estado de menor energia, adicionou-se um termo que depende do tipo de monômero que está interagindo, ou seja, um termo de energia que distingue se dois monômeros em contato são do mesmo tipo ou não. Assim, a energia é dada por:

$$E = \sum_{<i,j>|i-j\neq 1|} H_{t_i t_j} \tag{2.1}$$

onde o somatório é efetuado sobre todos os pares de monômeros vizinhos na rede, excluindo-se os pares ligados covalentemente.

A questão central do enovelamento é determinar a estrutura da proteína a partir da sua sequência de aminoácidos. Neste trabalho, para atingir essa finalidade é necessária a presença de pelo menos dois tipos de monômeros no modelo. Para verificar que tipo de contato está sendo feito entre os monômeros, foi inserida a seguinte matriz de interação:

$$H_{t_i t_j} = \begin{matrix} A \\ B \\ C \end{matrix} \begin{pmatrix} A & B & C \\ E_l & E_u & E_u \\ E_u & E_l & E_u \\ E_u & E_u & E_l \end{pmatrix} \tag{2.2}$$

onde E_l é a energia de contato entre dois monômeros do mesmo tipo e E_u é a energia entre dois monômeros diferentes. A, B e C estão representando três tipos diferentes de monômeros. Neste trabalho serão utilizadas proteínas com três a cinco tipos de monômeros diferentes.

2.2.1 Parâmetros de hidrofobicidade

Devido à interação da proteína com a água, o núcleo de uma proteína enovelada é constituído, em sua grande maioria, de resíduos hidrofóbicos, enquanto sua superfície é composta prioritariamente por resíduos hidrofílicos. Esse efeito é considerado na descrição do potencial definido na equação 2.1, sendo que pode-se reescrevê-la da seguinte forma:

$$E = n_l E_l + n_u E_u \tag{2.3}$$

onde n_l representa o número de contatos não covalentes entre monômeros do mesmo tipo e n_u é o número de contatos não covalentes entre monômeros de tipos diferentes.

O comportamento do sistema é analisado por meio dos parâmetros de hidrofobicidade

[9]:

$$\bar{E} = \frac{1}{2}(E_l + E_u) \tag{2.4}$$

$$E_{het} = (E_u - E_l) \tag{2.5}$$

onde $\bar{E}$ simula o efeito hidrofóbico, ou seja, dirige a formação dos contatos e é responsável pela compactação da cadeia. Já E_{het} determina a rugosidade da superfície de energia, a heterogeneidade dos diferentes resíduos. Se $E_{het} < 0$, a formação de contatos é favorecida e a cadeia colapsa veja figura 2.2.

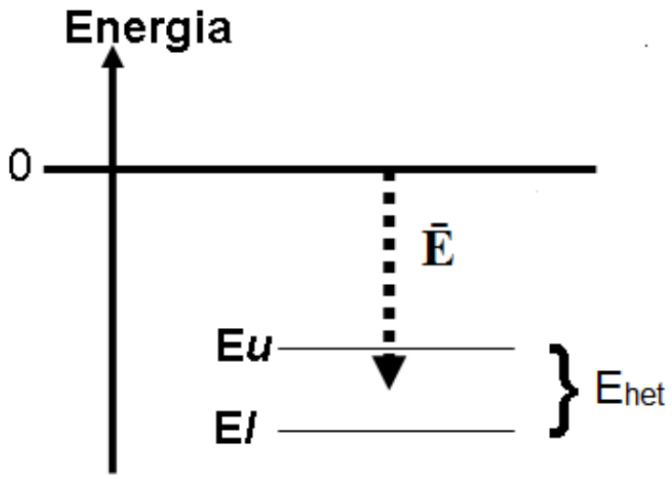


Figura 2.2: Diagrama dos níveis de energia do modelo. No diagrama está representada a energia de interação do mesmo tipo (E_l) e de tipos diferentes (E_u). $\bar{E}$ dirige a cadeia à formação dos contatos. E_{het} determina a diferença entre os níveis de energia.

O grau de colapso (k) é um parâmetro que auxilia na compreensão da influência da mudança de força que dirige o heteropolímero ao colapso, e é definido por:

$$k = -\frac{\bar{E}}{E_{het}} \tag{2.6}$$

Ajustando k, modifica-se o grau de compactação, ou seja, a hidrofobicidade do sistema. No limite de alta hidrofobicidade tem-se $k = 1$; em contrapartida, no o limite de baixa hidrofobicidade, tem-se $k = 0$. No enovelamento em regime de alta hidrofobicidade, a cadeia passa por um rápido colapso seguido de um lento enovelamento até atingir o estado nativo. No regime de baixa hidrofobicidade, a compactação e o enovelamento ocorrem simultaneamente até atingir o estado nativo [33].

2.2.2 Movimentos da cadeia

No modelo de rede, a dinâmica do enovelamento dependerá dos tipos de movimentos adotados para a cadeia [30]. Diferentes tipos de movimentos implicam em diferentes comportamentos cinéticos no enovelamento.

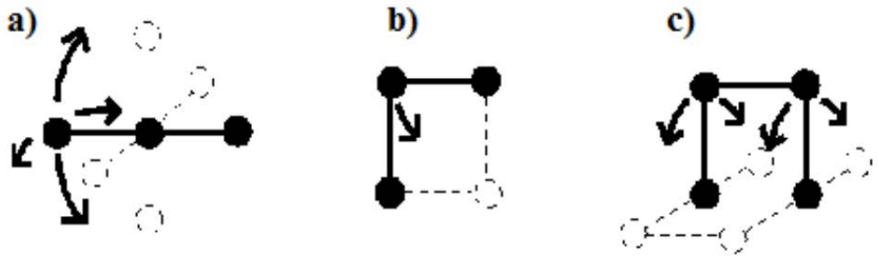


Figura 2.3: Tipos de movimentos possíveis baseados na dinâmica da Física de polímeros. a) Movimento de fim de cadeia (end move); b) Movimento de canto (corn move); c) Movimento de manivela (crankshaft move)

O conjunto de movimentos escolhidos para este trabalho (Figura 2.3) é baseado na Física de polímeros. Esse conjunto de movimentos consiste em deslocamentos locais que preservam a ligação covalente e a propriedade de volume excluso.

Esses movimentos são suficientes para responder apenas a questão cinética do tempo de enovelamento. Para a obtenção de outros dados correspondentes como, por exemplo, à dinâmica de formação de estruturas secundárias e terciárias, seria necessária a inclusão de movimentos mais complexos e globais.

A partir de uma estrutura aleatória crescida na rede, são efetuados movimentos locais, com monômeros escolhidos aleatoriamente, até que a proteína se enovele, ou até um tempo máximo estabelecido. Esta dinâmica será explicada, rm mais detalhes, nas seções seguintes.

2.3 Método de Monte Carlo

O Método de Monte Carlo é um poderoso método estatístico de simulação estocástica com diversas aplicações destinadas, principalmente, para o estudo de sistemas em estado de equilíbrio termodinâmico. O algoritmo de Metropolis[37] é uma versão particular que propõe encontrar o macroestado de equilíbrio de um sistema físico a uma dada tempera-

tura.

A eficiência do algoritmo está diretamente ligada ao fato de não levar em conta a probabilidade de uma dada conformação, mas sim uma razão, pois a razão entre as probabilidades de duas dadas configurações pode ser determinada independentemente das outras. Assim, dada duas configurações "n" e "m" quaisquer, a razão entre as probabilidades é dada por:

$$\frac{P_n}{P_m} = \frac{exp(-\frac{U_n}{K_bT})}{exp(-\frac{U_m}{K_bT})} = exp(\frac{U_m - U_n}{K_bT}) \tag{2.7}$$

onde P_n, é a probabilidade do estado n; U a energia e K_b é a constante de Boltzmann.

A partir dessa expressão, o algoritmo de Metropolis pode ser implementado por meio do seguinte conjunto de regras:

i Gera-se uma configuração aleatória, ou seja, com valores aleatórios para todos os graus de liberdade do sistema, respeitando as suas restrições. Atribui-se o índice n a essa configuração, que é aceita como configuração inicial na amostra.

ii Gera-se uma nova "configuração-tentativa", de índice m, que é resultado de pequenas alterações nas coordenadas da configuração n.

iii Se a energia da configuração m for menor que a energia da configuração n, inclui-se a configuração m na amostra, e atribui-se a ela o índice n a partir desse momento.

Caso a energia de m seja maior, realiza-se os passos descritos abaixo:

a) Gera-se um numero aleatório entre 0 e 1;

b) Se esse numero for aleatório for menor que P_n/P_m, aceita-se na amostra a configuração m, atribuindo-se a ela o índice n. Caso contrario, o índice n permanece designando a configuração original.

iv Repetem-se os passos *ii* e *iii* até que algum critério de parada seja satisfeito. Cada uma dessas repetições é denominada passo de Monte Carlo - *Monte Carlo steps* - (MCs).

2.3.1 Método de Monte Carlo aplicado ao modelo

A partir de uma estrutura crescida aleatoriamente no espaço da rede, os movimentos locais são realizados em um monômero escolhido de maneira aleatória. Se este for um monômero do final da cadeia, então um sítio vizinho na rede também é escolhido aleatoriamente para realizar o movimento (Figura 2.3.a). Caso não seja um monômero de fim de cadeia, então é possível executar o movimento de canto ou de manivela (Figura 2.3b e 2.3c), dependendo da disposição dos monômeros vizinhos.

Em todos os movimentos possíveis, se o monômero violar o princípio de volume excluso, ou seja, tentar ocupar um sítio que já está ocupado por outro monômero, a antiga configuração é restaurada e é contado como um passo de Monte Carlo. É contado, caso o movimento seja permitido, a energia da nova configuração é calculada e comparada com a energia da configuração antiga, seguindo, desse modo, o algoritmo de Metropolis.

2.4 Matriz de contatos, parâmetro de ordem Q e $\mathbf{Z}_{score}$

O processo de enovelamento envolve uma gradual compactação do sistema e à medida que o sistema atinge uma relativa compactação, pode-se reprentar cada configuração do sistema por uma matriz.

Assim, pode-se definir uma matriz de contatos para uma dada conformação, como uma matriz triangular 27×27, onde cada elemento pode assumir o valor 0 (zero) ou 1 (um). O valor 1 (um) aparecerá quando dois monômeros, definidos pelas colunas horizontais e verticais, estabelecerem contato e não estiverem ligado covalentemente. O valor 0 (zero) aparecerá quando dois monômeros não formarem contato.

O parâmetro Q representa a medida do grau de similaridade de uma dada conformação com a estrutura nativa, ou seja, a quantidade de contatos que estão presentes na forma nativa. Para o calculo desse parâmetro é utilizado o produto da sobreposição de duas matrizes de contato, definido como:

$$Q = N \cap C = \sum_{<i,j>} N_{ij} C_{ij} \tag{2.8}$$

onde N é a matriz em que estão presentes todos os contatos da conformação nativa e C é uma matriz que representa uma configuração qualquer. O resultado é um escalar que define a proximidade entre uma dada conformação e a conformação nativa. No modelo de rede, Q pode variar desde 0 (zero), o que corresponde a nenhum contato nativo, até 28, que é o número total de contatos existente na estrutura totalmente compactada.

Dentre os parâmetros conhecidos para o modelo teórico na rede, utilizou-se neste estudo o Z_{Score} [31, 38], por ser o mais apropriado para o cálculo da estabilidade termodinâmica e da acessibilidade cinética. Em termos matemáticos, o Z_{Score} é dado por:

$$Z_{score} = \frac{\langle E \rangle - E_{nativo}}{\sqrt{\langle E^2 \rangle - \langle E \rangle^2}} \tag{2.9}$$

Valendo-se da equação 2.9, verifica-se que o Z_{Score} mede a diferença de energia entre o estado nativo e a energia média de todas as conformações de uma mesma proteína, divido pelo valor do desvio padrão. Em outras palavras, o Z_{Score} é correlacionado com o grau de estabilidade da proteína, quanto maior o Z_{Score}, mais facilidade a protéina tem para enovelar.

2.5 Simulação computacional

Para realizar a visualização do funil de enovelamento da proteína é necessário um conjunto de informações que defina a unicidade de cada configuração como, por exemplo, a energia, a estrutura espacial, o parâmetro de ordem Q, o grau de compactação, entre outros.

Neste trabalho, o conjunto de informações foi escolhido de forma que cada configuração seja definida pela sua energia e pela quantidade de contatos formados. Esta informação é dada pela matriz de contato das conformações que a proteína assumiu durante toda cinética do enovelamento.

No entanto, quando a proteína está desenovelado, ou seja, com um valor de Q próximo de zero, o número de conformações é na ordem de milhares, sendo inviável computacionalmente analisar estas informações. A maneira encontrada para contornar esse problema foi analisar o enovelamento da proteína quando ela está parcialmente enovelada, ou seja,

um pouco antes do seu estado de transição ($Q_t - 1$), como apontado na Figura 2.4.

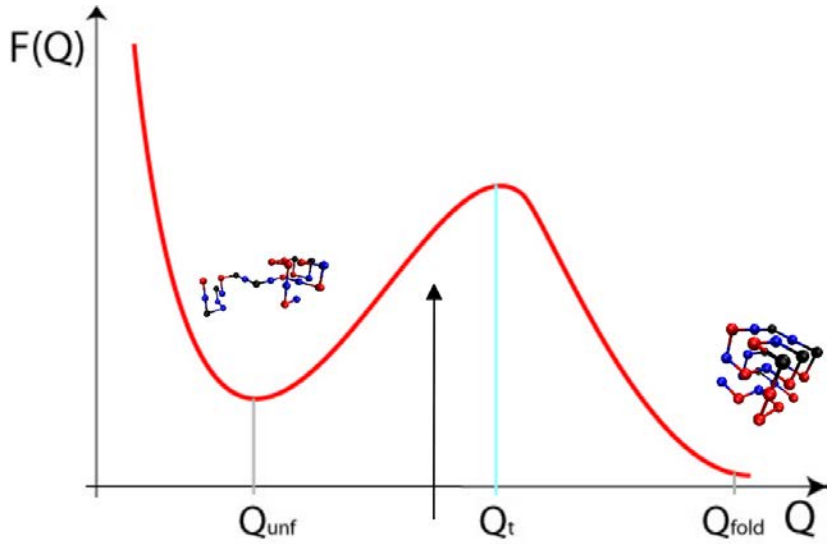


Figura 2.4: Exemplo de um gráfico de Energia livre *versus* Q, onde cada vale corresponde a um estado da proteína. Q_{unf} representa os estados desenovelados, Q_{fold} é o estado nativo e Q_t, o estado de transição. A seta indica a região onde se inicia a análise das conformações nesse estudo;;

Para se obter o comportamento da energia livre e indicar onde se encontra a barreira de transição, Q_t, fez-se o uso do método dos histogramas, que está descrito no Apêndice A.

A partir de $Q_t - 1$ é definido intervalos de tempo (em MC Steps) onde será obtida a matriz de contato da conformação. A escolha da conformação é baseado na sua energia, na qual a configuração de menor energia está contida em cada intervalo. Para analisar a importância dos intervalos de tempo na visualização, escolheram-se quatro intervalos distintos de tempo: 30, 100, 300 e 1000 passos de MC. Durante todo o tempo de simulação são obtidos milhares de minimos locais, nos quais podem se repetir. Na Figura 2.5 está ilustrado um esquema de como são coletados esses dados. O intervalo de tempo define o refinamento da obtenção desses mínimos locais, ao passo que, quanto menor for o intervalo, melhor será distribuição de energia.

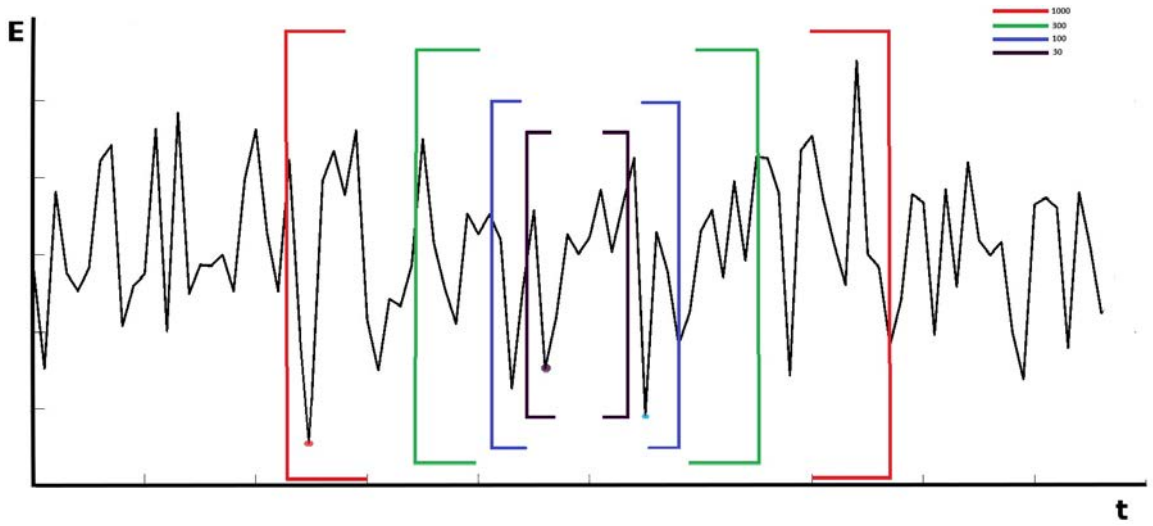


Figura 2.5: Representação esquemática os dos intervalos de tempo. Os colchetes representam os intervalos e cada cor corresponde a um intervalo de tempo. Em detalhe, os mínimos locais para cada intervalo. É possivel observar, que em um intervalo pequeno (30 MCs), a existência de um minimo local. Em contrapartida, ao observar a mesma região num intervalo maior (1000 MCs) verifica-se que aquela conformação é apenas uma flutação da energia.

Capítulo 3

Visualização

Uma vez realizada a simulação computacional, obtém-se uma base de dados com milhares de instâncias, onde cada instância uma representa uma conformação do enovelamento. Cada instância possui atributos que identificam a estrutura da conformação a qual representam. Pode-se definir o trabalho a ser realizado como a extração de informações de uma densa base de dados, na qual as informações de interesse são relacionadas a padrões e tendências no processo de enovelamento de proteínas.

O processo para realizar a visualização é constituído de quatro partes: *i*) Processamento de dados; *ii*) criação da matriz de distância; *iii*) Projeção multidimensional em duas dimensões; e *iv*) Criaçao da superfície em três dimensões.

Para o processamento de dados, mencionou-se anteriormente que cada instância da base de dados, ou seja, cada conformação, era representada por uma matriz binária 27×27 (seção 2.4). Para efeito de cálculo, cada conformação será representada por um vetor binário de 729 posições. Assim, pode-se considerar que cada instância de dados tem 729 atributos, ou seja, os dados se encontram originalmente em um espaço m-dimensional, com m equivalente a 729.

Para cada conformação distinta da base de dados é dada uma identificação. Cria-se, então, o mapa de conectividade entre as conformações, ou seja, para cada par de conformação é calculado o caminho mínimo para ir de uma à outra. Calcula-se também a incidência em que cada conformação foi visitada durante a simulação.

O próximo passo para a construção da projeção é compreender o que é medido. Buscar

a relação que melhor reflita a diferença entre os objetos em determinado domínio. No contexto atual, a distância deve ser capaz e suficiente para distinguir duas conformações.

O conceito de contatos, explicado em detalhes na seção 2.4, é essencial para a elaboração de uma medida de distância entre as conformações, pois diferentes combinações de contatos resultam em diferentes estruturas e, consequentemente, em diferentes conformações. Como mencionado anteriormente, cada instância da base de dados possui 729 atributos, e são justamente esses atributos que indicam os contatos realizados por uma conformação.

De acordo com o modelo proposto, uma conformação faz no máximo 28 contatos. Logo, a maioria dos atributos observados em uma conformação são nulos. Este fato deve ser levado em conta durante a elaboração de uma medida de distância, isto é, para comparar duas conformações, deve-se observar apenas os contatos existentes em uma delas ou em ambas, descartando-se desse modo, informações sobre a inexistência de um determinado contato em ambas as estruturas.

De maneira mais formal, a medida para a comparação da estrutura entre duas conformações pode ser escrita como:

$$M_e(x_i, x_j) = \frac{D_{ij}}{C_{ij} + 1} \tag{3.1}$$

onde M_e é uma matriz denominada de Medida Estrutural; $C_{i,j}$ representa os contatos comuns às conformações x_i e x_j, isto é, valores iguais a "1"em x_i e x_j, e $D_{i,j}$ representa os contatos distintos encontrados em x_i e x_j.

Como pode ser observado na formulação dessa medida, o valor calculado nunca será negativo. Outra característica que pode ser notada é que, quanto mais próximo de zero, maior será a semelhança entre as duas conformações, de modo que tem-se $M_e(x_i, x_j) = 0$ para $x_i = x_j$. Além disso, uma propriedade importante dessa medida é a simetria, ou seja, $M_e(x_i, x_j) = M_e(x_j, x_i)\ \forall\ x_i,\ x_j \in \mathrm{X}$.

Decidiu-se incluir uma segunda característica para diferenciar duas conformações, pois *a priori* considerou-se a comparação entre as estruturas de duas conformações insuficiente para diferenciá-las durante o processo de enovelamento. Uma questão importante que deve ser considerada é a dinâmica do sistema, ou seja, como ocorrem as transições entre as conformações. Duas conformações podem ser estruturalmente semelhantes, porém

para transitar de uma conformação para outra, pode ser necessário passar por diversas conformações intermediárias. A Medida Dinâmica, responsável por avaliar essas transições entre conformações, pode ser descrita como:

$$M_d(x_i, x_j) = n \tag{3.2}$$

onde n é o número mínimo de conformações intermediárias necessárias para transitar de x_i para x_j em qualquer sentido, este calculo é obtido pelo mapa de conectividade feito na primeira etapa de visualização, o processamento de dados. Quando a transição for direta, o valor da medida será zero. Ao término do cálculo, divide-se todas as medidas calculadas pela maior distância encontrada, normalizando, assim, os valores entre 0 e 1.

Assim, com as duas medidas definidas (equações 3.1 e 3.3), calcula-se $\delta(x_i,x_j)$ como:

$$\delta(x_i, x_j) = M_e(x_i, x_j) + M_e(x_i, x_j)M_d(x_i, x_j) \tag{3.3}$$

onde M_e e M_d representam, respectivamente, a medida estrutural e a medida dinâmica entre as conformações x_i e x_j. Observa-se que a medida dinâmica funciona como um ajuste à medida estrutural, uma vez que seu valor esta normalizado entre 0 e 1.

O termo à esquerda do sinal de igualdade na expressão 3.4 ($\delta(x_i,x_j)$) é chamado de matriz de distâncias. É uma matriz triangular $m \times m$, onde m é o número de conformações existente no espaço de fase.

A etapa mais complexa para a criação deste modelo é a redução do espaço dimensional 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 similaridade entre os dados [39] .

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 $\mid \delta(x_i, x_j)$–$d(x_i, x_j) \mid$ 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 δ. Intuitivamente, pontos distantes representam objetos com pouca relação, também de acordo com δ. 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 (δ) são calculadas.

Dentre as várias técnicas de projeção que podem ser utilizadas, adotou-se, neste trabalho, a técnica Force Scheme [39], 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 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 3.1, onde são definidos círculos de energia para cada valor de energia encontrado durante o processo de enovelamento. As conformações são espalhadas uniformemente em seus respectivos círculos. Energias menores são representadas por círculos contidos dentro de círculos de maior energia.

Após o posicionamento inicial dos pontos, a técnica Force Scheme realiza iterações para aproximar as distâncias entre os objetos projetados das distâncias δ 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 $\vec{v}$, uma fração de Δ. Com Δ dado por:

$$\Delta = \left\{ \left[1 + (2j+1)^{\frac{1}{2j+1}} \right] \right\} - 1 \tag{3.4}$$

onde j é o número de iterações realizadas até aquele momento.

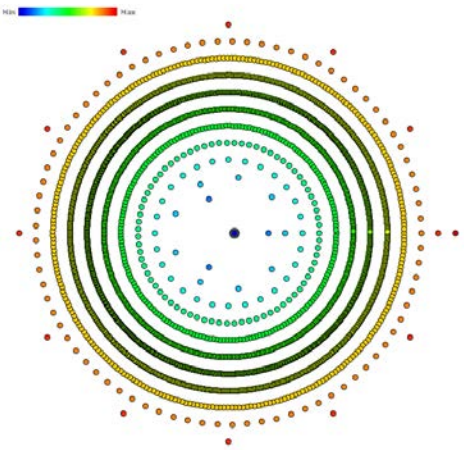


Figura 3.1: Inicialização do sistema para a projeção em duas dimensões. Neste caso, a disposições das conformações foram baseada em energias. No centro encontra-se o estado nativo (menor energia).

Ao término da iteração, cada objeto sofreu um deslocamento na direção de cada outro objeto, aproximando a distância entre eles da distância calculada entre as conformações que eles representam. As iterações são repetidas sucessivamente até o sistema atingir o equilibrio.

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 comparado 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.

Ao se obter uma projeção final bi-dimensional, prossegue-se para a última etapa da construção do modelo visual. Esta etapa usa as energias das conformações para deslocar os pontos projetados em um eixo perpendicular ao plano de projeção. Logo, obtêm-se uma estrutura tridimensional, onde conformações de baixa energia são representadas por pontos em níveis mais baixos e conformações de maior energia por pontos em níveis mais altos nesse novo eixo.

Capítulo 4

Resultados e Discussões

Os resultados foram obtidos a partir de um código computacional responsável por simular o enovelamento de proteínas, composta por 27 monômeros em uma rede cúbica $3\times3\times3$, obedecendo ao modelo descrito no capítulo 2. Para realizar as simulações, foram escolhidas proteínas com características diferentes, tais como o tipo dos monômeros, o valor do Z_{Score} e o número de degenerescência. As proteínas utilizadas foram:

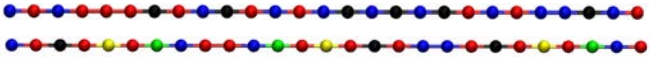
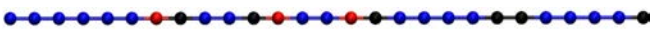
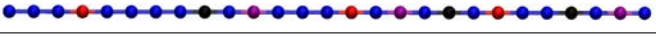
Sequência	**Z-score**	**Tipos**	**Representação**	$\mathbf{T}_f$
0012	6.75	3		1.89
0012f	5.91	3		1.23
43157	8.58	5		1.90
2221	5.90	3		1.95
45568D	6.27	4		1.73

Tabela 4.1: Proteínas exploradas nesse trabalho. A sequência 0012 projetada por Socci *et.al* [30] é uma sequência estável (de acordo com seu Z_{Score}), utilizada em diversos trabalhos anteriores [33, 38, 40]; A sequência 00012 é uma pequena mutação da sequencia 0012, tal sequência apresenta 3 contatos frustrados em sua estrutura nativa [38]; As sequências 43157 e 2221 foram escolhidas devido à sua estabilidade (sendo a 43157 estável e a 2221 pouco estável (de acordo com o seu Z_{Score})); A 45568D tem por característica a degenerescência do estado nativo igual a dois, ou seja, existem duas conformações com a menor energia. Estas duas conformações diferem-se entre si por 5 contatos. As cores representa os tipos de monômeros.

Um parâmetro comum a todas as simulações é o grau de hidrofobicidade, que foi definido de forma que o sistema tenha uma baixa hidrofobicidade. Assim, estabeleceu-se $E_m = 0$ e $E_{het} = 6$, resultando, de acordo com as equações 2.4 e 2.5, em $E_l = -3$ e $E_u = 3$. Optou-se simular em baixa hidrofobicidade, pois o enovelamento ocorre de

maneira continua, sem ficar preso em mínimos locais, uma vez que as barreiras de transição são menores do que no regime de alta hidrofobicidade [9, 38].

As unidades de temperatura e energia foram escolhidas convenientemente de forma que $K_b = 1$ e que a temperatura de vidro do sistema fosse $T_g = 1$ para todas as estruturas estudadas. Para a sequência 0012f, o estado nativo apresenta 3 contatos desfavoráveis (*unlike*) e, de acordo com a função potencial descrita na equação 2.3, a energia do estado nativo será -66, pois $n_l = 25$, E_l = -3, $n_u = 3$ e $E_u = +3$. Para todas as outras proteínas, não existem contatos desfavoráveis na sua estrutura nativa. Logo, a energia do estado fundamental será -84, pois $n_l = 28$ e $n_l = -3$.

Para as sequências 0012, 0012f, 43157 e 2221, que são definidas por terem um único estado de menor energia (*protein-like*), executou-se a simulação em intervalos de tempo iguais a 30, 100, 300 e 1000 MCs. Para a proteína 45568D, definida por ter o estado nativo degenerado, executou a simulação em um único intervalo de tempo de 1000 MCs . Para todas as sequências, a quantidade de passos de Monte Carlo foi escolhida de modo que, em média, são obtidos $1,0.10^7$ matrizes de contato.

4.1 Visualização em duas dimensões

Utilizando a metodologia apresentada nos capítulos 2 e 3, gerou-se a projeção em duas dimensões para a sequência 0012 (Figura 4.1). Nesta figura, cada ponto representa uma conformação (mínimo local) e a cor está associada à energia daquela conformação, onde a cor azul representa a conformação de menor energia e a cor vermelho ilustra a conformação de maior energia. A distância entre quaisquer dois pontos representa a similaridade entre as conformações, ou seja, quanto mais próximos estão os pontos, mais similares são as conformações, em outras palavras, as conformações apresentam os mesmos contatos. Na Figura 4.1 é indicado três regiões sendo, cada um delas representadas por uma conformação . Nota-se que as três conformações do exemplo apresentam diferenças estruturais evidentes, o que nos colocar em posição de acreditar que a projeção dos dados esta sendo feita de maneira coerente, representando fielmente os dados multidimensionais.

Como a projeção mostrou-se eficiente, objetivou-se analisar o comportamento da visualização para diferentes intervalos de tempo. Como foi explicado na seção 2.5, para um dado intervalo de tempo é armazenado a conformação de menor energia logo, a mudança do intervalo de tempo representa um refinamento desses mínimos locais.

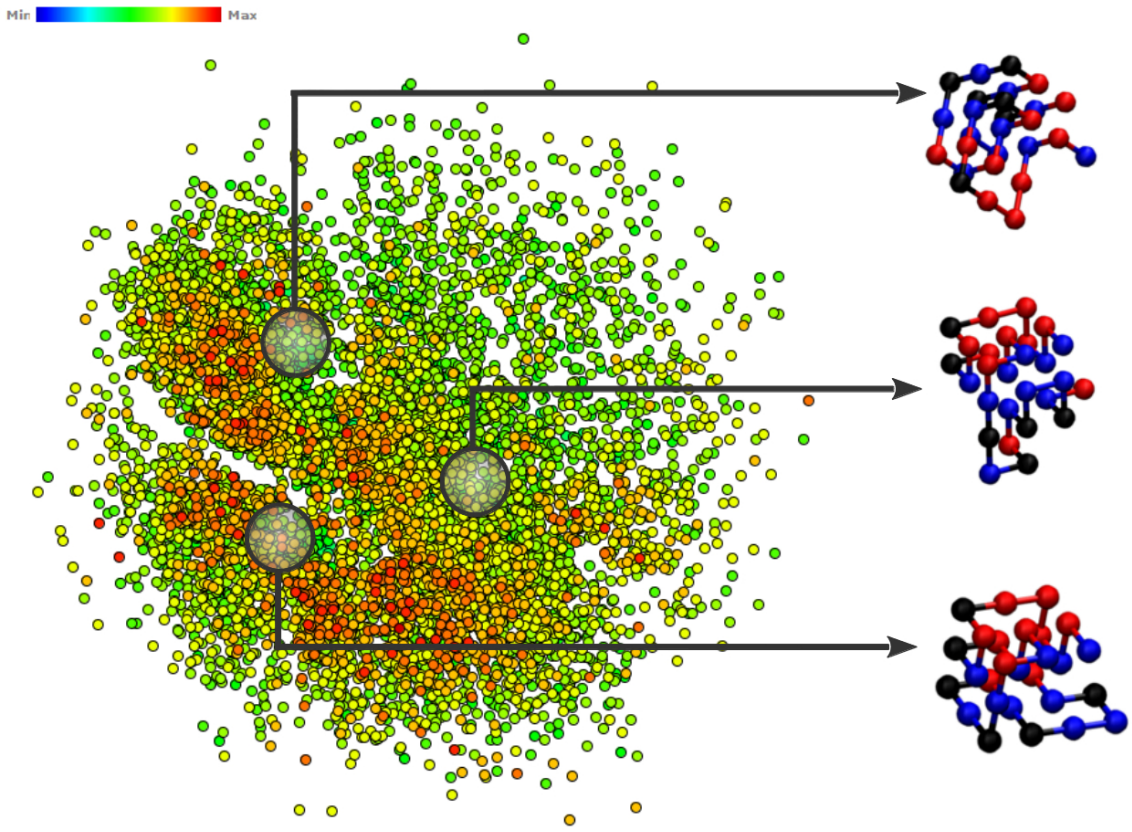


Figura 4.1: Visualização em duas dimensões da sequência 0012 para o intervalo de 100 MCs. Em detalhe, três regiões da projeção que apresentam estruturas diferentes.

Em seguida, construiu-se a projeção, em duas dimensões, para os quatro intervalos de tempo para todas as proteínas tratadas nesse estudo (Figuras 4.2 a 4.6).

De posse da Figura 4.2, que ilustra a projeção, em duas dimensões, da sequência 0012 é possivel afirmar que para todos os intervalos de tempo, o padrão da visualização permanece inalterado, ou seja, a distribuição das conformações continua semelhante. Tal fato garante que cada proteína analisada tem uma assinatura independente da janela temporal observada, exibindo aglomerados de conformações que ficam melhor definidos nos intervalos de 30, 100 e 300 MCs. Esses aglomerados indicam regiões conexas, onde a proteína pode ficar armadilhada. (Essas regiões serão analisadas, em maiores detalhes, mais a frente).

No intervalo de tempo igual a 1000 MCs é possível observar, com maior facilidade, a distribuição energética. Consequentemente, nesse intervalo, o mínimo local é bastante refinado, causando uma diminuição nas conformações de alta energia que, nos outros intervalos, são consideradas mínimos locais.

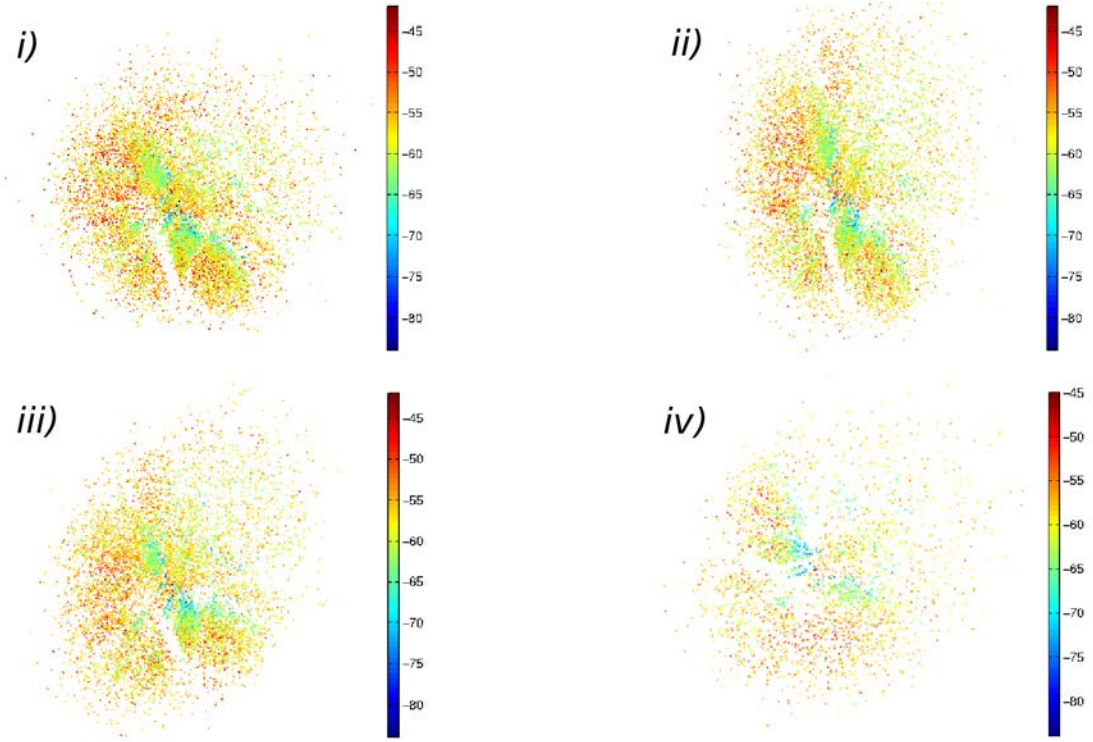


Figura 4.2: Visualização em duas dimensões, da sequência 0012 para todos os intervalos de tempo considerados neste trabalho. *i)* 30 MCs; *ii)* 100 Mcs; *iii)* 300 MCs; *iv)* 1000 MCs.

Para as outras proteínas estudadas, observaram-se os mesmos padrões para os diferentes intervalos de tempo, seguindo, portanto a mesma discussão feita para a sequência 0012. Isto pode ser constatado nas Figuras 4.3, 4.4 e 4.5, nas quais são apresentadas as visualizações, das sequências 0012f, 2221 e 43157, respectivamente.

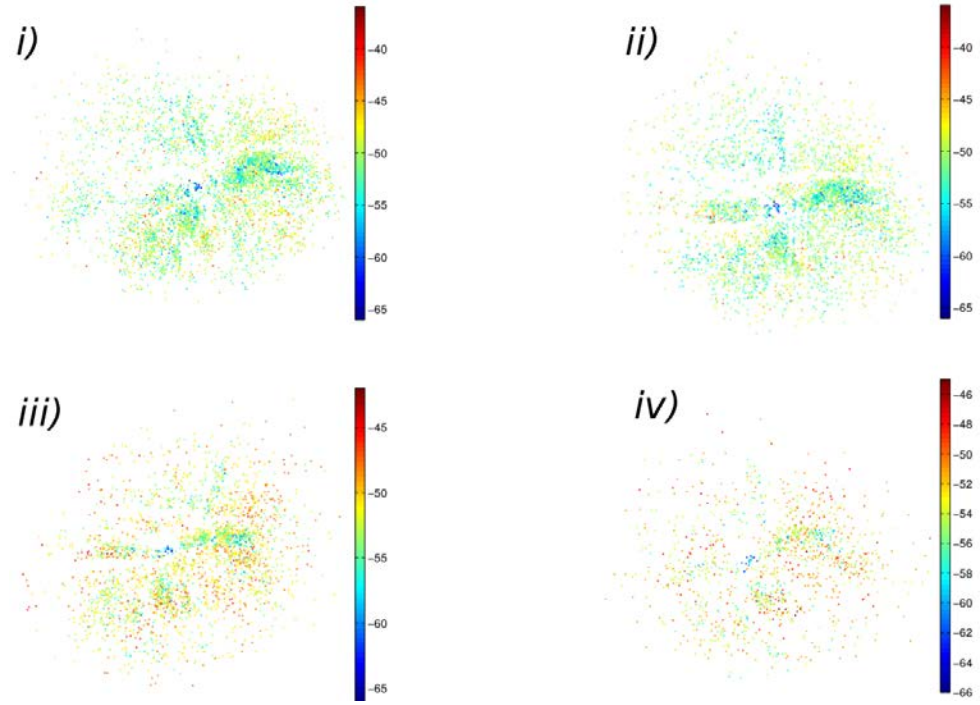


Figura 4.3: Visualização em duas dimensões, da sequência 0012f para todos os intervalos de tempo considerados neste trabalho. *i)* 30 MCs; *ii)* 100 Mcs; *iii)* 300 MCs; *iv)* 1000 MCs.

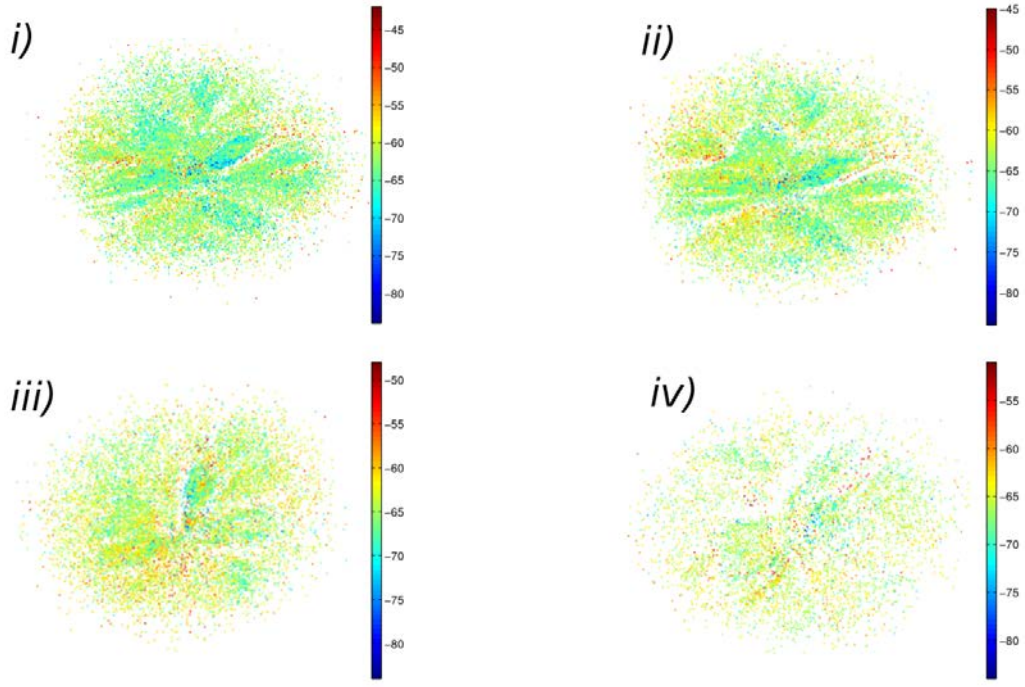


Figura 4.4: Visualização em duas dimensões, da sequência 2221 para todos os intervalos de tempo considerados neste trabalho. *i)* 30 MCs; *ii)* 100 Mcs; *iii)* 300 MCs; *iv)* 1000 MCs.

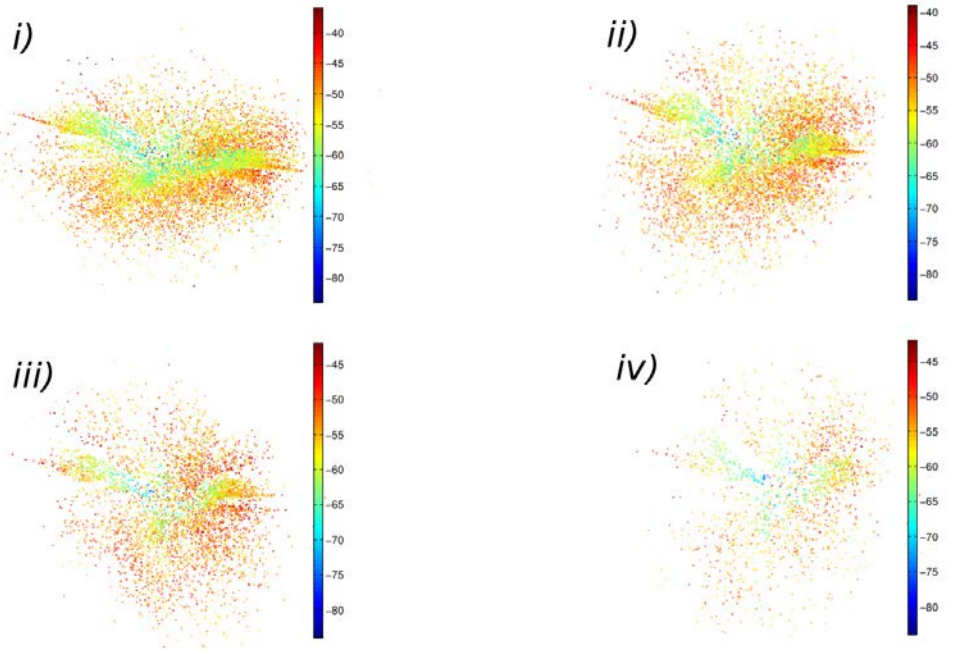


Figura 4.5: Visualização em duas dimensões, da sequência 43157 para todos os intervalos de tempo considerados neste trabalho. *i)* 30 MCs; *ii)* 100 Mcs; *iii)* 300 MCs; *iv)* 1000 MCs.

Na Figura 4.6 está ilustrada a projeção, em duas dimensões, da a sequência 45568D. Esta sequência tem por característica a degenerescência do estado nativo igual a dois, ou seja, existem duas conformações com menor energia. Estas duas conformações diferem-se entre si por 5 contatos. Nesta proteína, pode-se observar a presença de dois aglomerados bem definidos, representando as regiões onde estão as conformações de menor energia (os estados nativos).

Estes aglomerados podem ser interpretados como regiões na qual a proteína fica armadilhada próxima a uma das conformações nativas e, para conseguir atingir o outro estado nativo, ela deve se desenovelar, ou seja, quebrar contatos e formar novos contatos direcionados à próxima estrutura nativa.

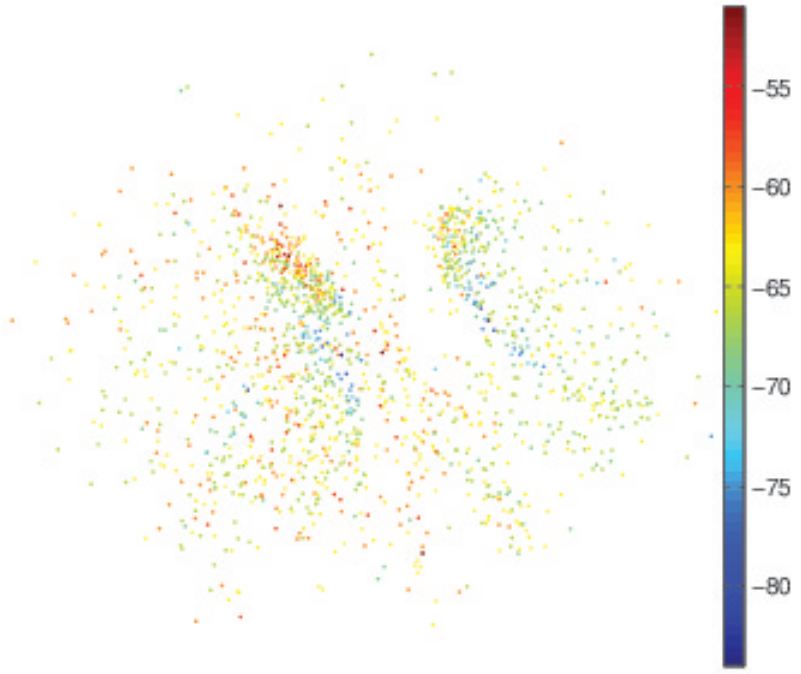


Figura 4.6: Visualização, em duas dimensões, da sequência 45568D para o intervalode tempo de 100 MCs.

4.1.1 Rotas de enovelamento

Para entender o significado desses aglomerados, que estão presentes em todas as visualizações em duas dimensões, propôs-se analisar as rotas de enovelamento. A partir de uma conformação desenovelada aleatória, observou-se o caminho percorrido pela proteína até atingir o estado nativo pela primeira vez. Na Figura 4.7 pode-se acompanhar algumas dessas rotas para a sequência 0012.

Cada rota analisada se originou de uma conformação aleatória e, em geral, quando a proteína entra em uma região que apresenta aglomerados, ela fica se interconvertendo em conformações próximas, direcionando-se para o estado nativo.

Outro aspecto observado na Figura 4.7 foi que as rotas visualizadas durante o enovelamento não atravessam as regiões que apresentam um vale (regiões onde não existe nenhuma conformação). Desse modo, pode-se inferir que cada aglomerado representa um caminho para o enovelamento e, para se atingir um outro aglomerado, a proteína deve romper vários contados (direcionados para a periferia da visualização) e eftuar novos

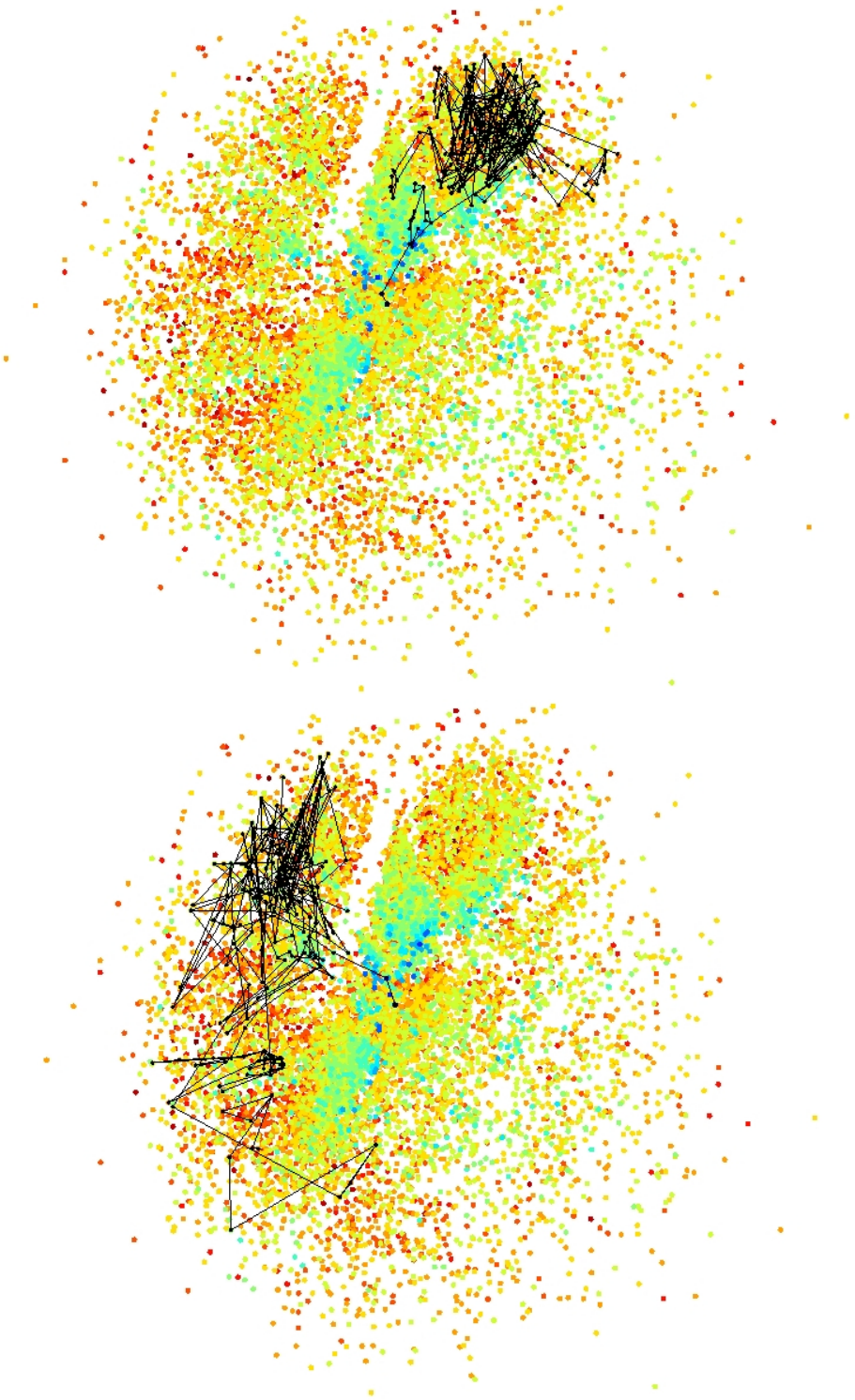


Figura 4.7: Rotas de enovelamento da sequência 0012. Pode se observar que cada rota começa em uma conformação aleatória diferente, traçando caminhos distintos até atingir o estado nativo.

contatos direcionados a um outro aglomerado.

Uma segunda forma de analisar esses aglomerados e garantir que a projeção está mos-

trando um resultado coerente, foi simular uma grande quantidade de rotas, calculando a distância entre cada conformação e sua subsequente, ou seja, a distância entre a conformação i e a conformação *i+1*, até atingir o estado nativo. Para este cálculo, geraram-se 10^4 rotas aleatórias para a sequência 0012. O resultado pode ser visto na Figura 4.8.

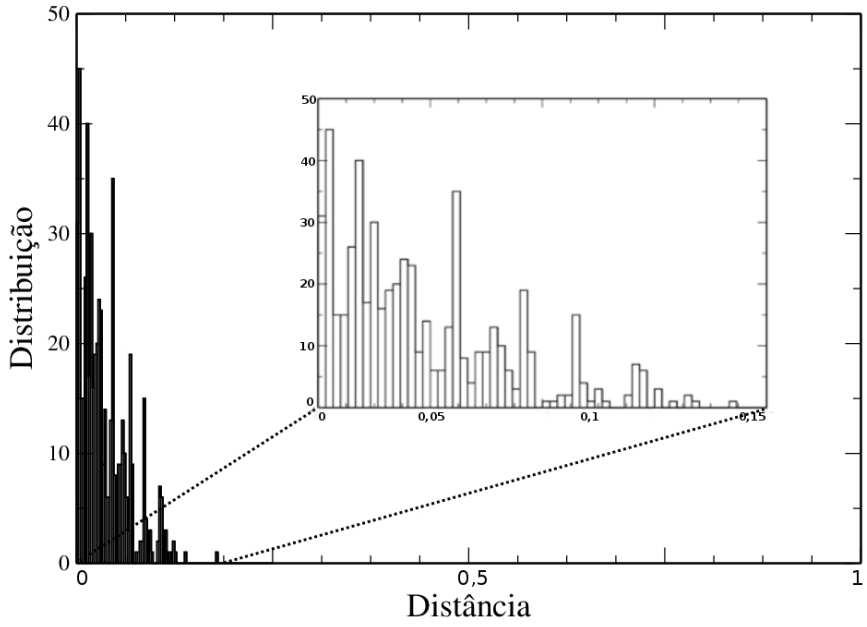


Figura 4.8: Histograma das distâncias percorridas pelas rotas referentes a sequência 0012. A distância (eixo x) está normalizada, sendo $x = 1$ a maior distância encontrada (421 em unidades arbitrárias). No detalhe, pode se observar uma ampliação da região de maior incidência.

O resultado da Figura 4.8 demonstra que distâncias pequenas ocorrem com maior frequência. Tal fato demonstra que a projeção em duas dimensões está sendo representada de maneira eficiente, uma vez que, ao se acompanhar as rotas passo a passo, espera-se que, em média, a proteína não mude drasticamente a sua conformação. Assim, de acordo com a definição da distância efetiva (equação 3.4), para pequenas diferenças conformacionais, o valor do $\delta(x_i, x_j)$ deve ser baixo.

Como ultima etapa, analisou-se a quantidade de conformações necessárias para que a proteína atinja, partindo de uma estrutura aleatória, o seu estado nativo. Este resultado pode ser visto na Figura 4.9, na qual foram geradas 10^4 trajetórias para cada proteína (*protein-like).*

Pode-se observar que as duas proteínas com alto valor de Z_{score} (sequências 0012 e

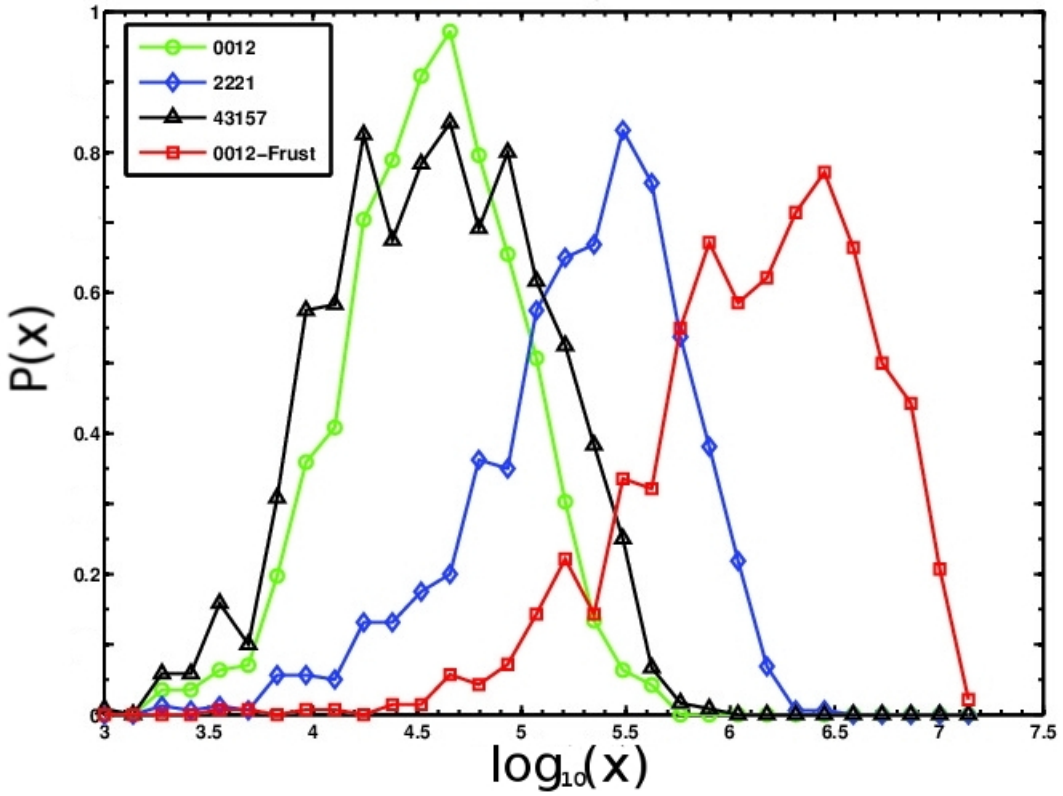


Figura 4.9: Gráfico da distribuição *versus* $log_{10}(x)$, onde x é o número de conformações para a proteína, a partir de um estado aleatório, atingir estado nativo pela primeira vez.

43157) enovelam-se, em média, mais rapidamente. Já nas sequências com baixa estabilidade, (2221 e 0012f), o número médio de conformações necessárias para se atingir o estado nativo é muito maior. Isto se deve ao fato que essas proteínas ficam armadilhadas em regiões com mínimos locais, demorando, assim, mais tempo para conseguir se enovelar.

Este fato correlaciona-se com a quantidade de conformações com energias intermediárias das sequências com Z_{score} baixo (vide Figura 4.4 e 4.3). Elas apresentam um número grande de conformações na faixa de energia do azul claro/verde, regi ao em que as proteínas ficam armadilhadas em muitos mínimos locais, demorando um tempo maior para conseguir atingir o estado nativo.

4.2 Visualização em três dimensões

4.2.1 Sequência 0012

Para construir a visualização em três dimensões, utilizou-se a projeção em duas dimensões como eixos x e y e, para o eixo z, usou-se a distribuição energética, na qual, o ponto mais baixo é o estado nativo. A Figura 4.10 apresenta a visualização da sequência 0012.

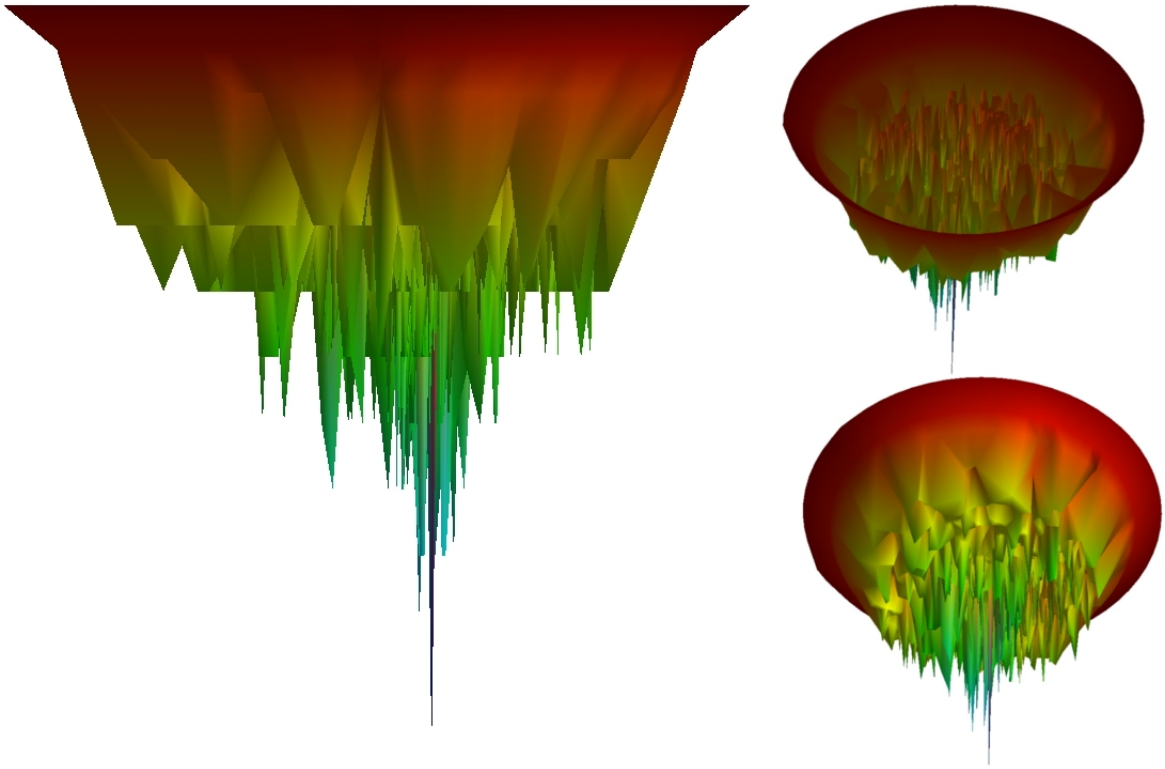


Figura 4.10: Visualização, em três dimensões, da sequência 0012. À esquerda, tem-se a vista de perfil do funil; à direita, tem-se, em detalhes, a parte interna e externa do funil.

Na visualização em três dimensões, observa-se uma distribuição energética discreta (faixas de energia). Tal fato é consequência da forma como é definida a energia no modelo de rede (equação 2.3). Também é possível observar a forma afunilada desta superfície, mostrando que o número de conformações diminui conforme a energia também decai, até se atingir o estado de menor energia, que é quando a proteína está enovelada.

4.2.2 Sequência 2221

Para proteínas diferentes, a forma do superfície pode variar, sendo observado que, quanto maior o valor do Z_{score}, mais a superfície se aproxima de um funil. A seguir, mostra-se a visualização do funil para sequência 2221.

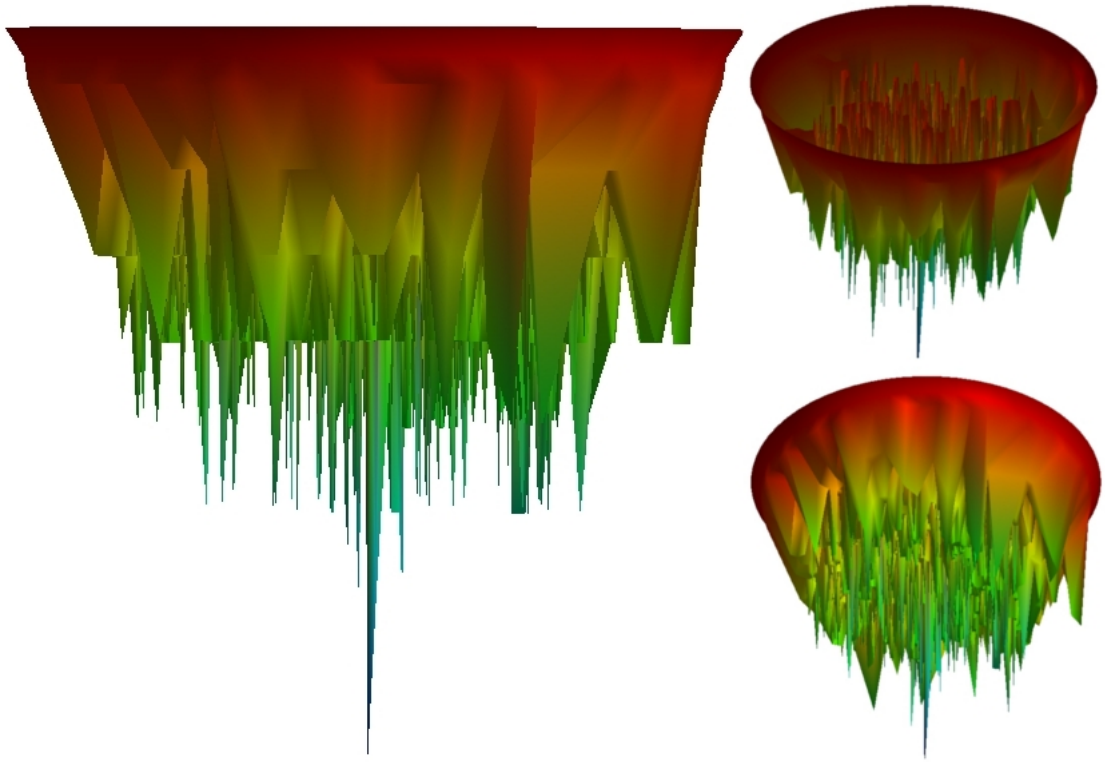


Figura 4.11: Visualização, em três dimensões, da sequência 2221. À esquerda, tem-se a vista de perfil do funil; à direita, tem-se, em detalhes, a parte interna e externa do funil.

Para proteínas com baixo Z_{score}, a forma da superfície pode variar, como cosequência desse fato, que devido ao número de conformações com energias intermediárias a superfície fica deformada, ou seja, afasta-se da forma afunilada. Este resultado fica evidente ao se observar a Figura 4.11, pois nota-se que a distribuição das conformações é mais esparsa, indicando regiões onde a proteína pode ficar armadilhada, em regiões afastadas da estrutura nativa (regiões de mínimos locais).

4.2.3 Sequência 45568D

A Figura 4.12 ilustra a visualização, em três dimensões, para a proteína degenerada 45568D:

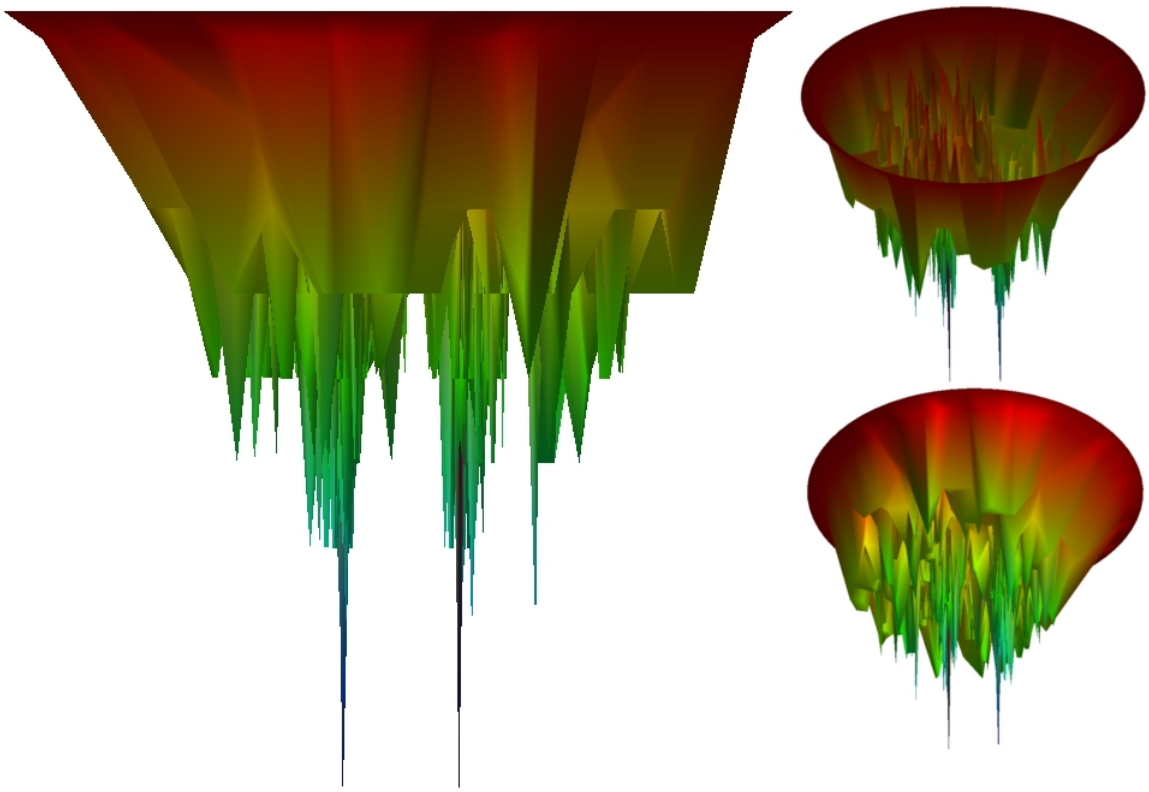


Figura 4.12: Visualização, em três dimensões, da sequência 45568D. À esquerda, tem-se a vista de perfil do funil; à direita, tem-se, em detalhes, a parte interna e externa do funil.

A visualização em três dimensões da sequência 45568D se torna interessante devido à presença do estado nativo degenerado. Este fato permite a verificação da consistência da projeção, pois devido a esta característica, esperava-se um funil com dois mínimos bem definidos (funil duplo).

Na Figura 4.12 fica evidente a diferença entre os estados nativos, ao passo que, para sair de um estado nativo e ir para o outro, há a necessidade de que a proteína ganhe energia (para subir no funil), e fazer outros contatos (para que assim, alcance o outro poço).

4.3 Análise de uma mutação

A projeção em 2D foi aplicada para explorar uma pequena mutação na sequência 0012. Esta mutação consistiu em permutar dois monômeros, tornando a sequênca mutada menos estável (sequência 0012f). Para avaliar os efeitos da mutação, criou-se uma única projeção das duas conformações (0012 e 0012f). Este resultado pode ser visto na Figura 4.13.

De posse da Figura 4.13, verifica-se na projeção que, devido a mutação, perde-se toda a região esquerda do enovelamento, ou seja, um conjunto de conformações foi severamente desfavorecido energéticamente, o que nos leva a inferir que, ao realizar a mutação, a proteína, em geral, fica desestabilizada. Tal fato se reflete na redução de caminhos possíveis para atingir o estado nativo, evidenciando os altos tempos para essa sequência se enovelar (vide figura 4.9 para sequência 0012f).

Foi calculado também 10^4 rotas de enovelamento para ambas sequências(Figura 4.13b e 4.13c). Para a sequência 0012, a análise foi apresentada no seção 4.1.1. Para a sequência mutada, em geral, as rotas percorrem praticamente todo o espaço conformacional antes de atingir o estado nativo, sendo que 95% das rotas ocorrem na região esquerda da projeção.

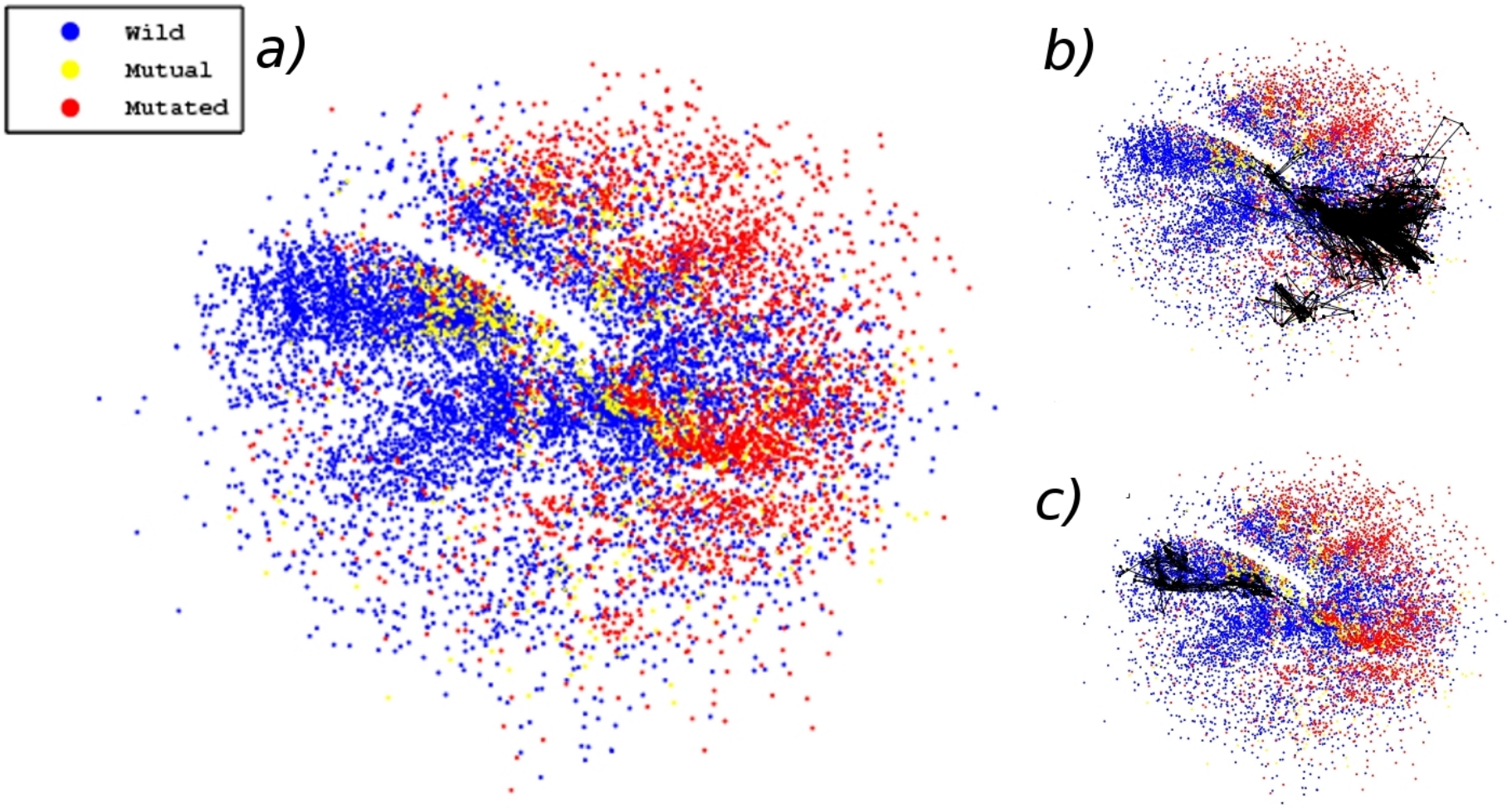


Figura 4.13: Projeção em, em duas dimensões, da sequência 0012 e sua mutação 0012f. Os pontos em azul representam a sequência 0012; os pontos em vermelho representam a sequência mutada 0012f e os pontos em amarelo são conformações comum a ambas sequências. Na direita é apresentado um exemplo de rota para cada sequência, *b)*: Rota para sequência 0012f; *c)*: Rota para sequência 0012.

Capítulo 5

Conclusões

A ideia geral do funil de enovelamento é definida como um vale profundo em uma superfície multidimensional, centrado no mínimo global de energia do sistema que, por sua vez, corresponde ao estado nativo da proteína. O formato típico deste funil é largo no "topo", representando o alto número de conformações (regiões de altas energias e entropias) e vai gradualmente se estreitando até atingir o seu mínimo global, que representa a estrutura nativa da proteína.

Como as superfícies de energia são caracterizadas por um grande número de mínimos locais, espera-se que que esses funis de energia sejam "rugosos", isto é, apresentem muitos mínimos que variem amplamente em suas profundidades e nas alturas das barreiras de energia. No processo de enovelamento, em condições fisiológicas, e considerando-se que a proteína esteja no "topo"do funil, ela segue através de um grande número de caminhos até atingir a estrutura nativa. A teoria do funil possibilita a compreensão destes caminhos que permitem que a proteína alcance seu estado nativo em tempo hábil de desempenhar seu papel biológico.

A partir de simulações computacionais, aliadas ao modelo minimalista de rede cúbica, foi possível fazer uma abordagem qualitativa e quantitativa desses funis, analisando-se o comportamento conformacional do enovelamento de proteínas diversificadas e classificando-se não só suas estruturas conformacionais, mas também capaz de avaliar as similaridades entre conformações por meio de uma métrica efetiva. A partir desses dados multidimensionais, utilizou-se a técnica de projeção Force-Scheme que possibilitou a visualização em duas e três dimensões do funil dessas proteínas.

As visualizações em duas dimensões possibilitaram a verificação da existência de padrões para cada proteína, mesmo variando-se alguns parâmetros, como o intervalo de tempo. Posteriormente, analisou-se a robustez das projeções, verificando se, realmente as conformações parecidas estruturalmente se encontravam próximas na projeção. De acordo com os resultados obtidos na figura 4.8, pode-se afirmar que a projeção está consistente.

Finalmente, aplicou-se o método descrito nesse trabalho para verificar a influência de uma pequena mutaçao sobre sua projeção. Este resultado apontou que, devido à mutação, o mapa conformacional foi modificado, como se pode observar na Figura 4.13a, no qual é possivel verificar que toda uma região foi drasticamente desfavorecida. Este resultado sugere que a visualização do funil de enovelamento pode ser aplicado à modelos mais complexos, que se baseam em estruturas de proteínas reais, possibilitando, desse modo, a constatação de regiões que são inibidas ou favorecidas devido a uma determinada mutação. Realizando-se tal tarefa, espera-se poder inferir sobre possíveis problemas causados por "mal enovelamento" da proteína conhecidos como *misfold*.

A partir dos resultados aqui apresentados, conclui-se que a visualização do funil de enovelamento de proteínas em modelo de rede se mostrou satisfatório, possibilitando a extração de informações importantes sobre a distribuição conformacional da proteína e a conectividade entre as sua sconformações.

Referências Bibliográficas

[1] David L. Nelson and Michael M. Cox. *Lehninger Principles of Biochemistry*. W. H. Freeman, 5th edition, February 2008.

[2] Donald B. Wetlaufer. *The Protein folding problem*. Published by Westview Press for the American Association for the Advancement of Science, 1984.

[3] C B Anfinsen. Principles that govern the folding of protein chains. *Science (New York, N.Y.)*, 181(4096):223–230, July 1973. PMID: 4124164.

[4] Cyrus Levinthal. Are there pathways for protein folding? *Extrait du Journal de Chimie Physique*, 65(1), 1968.

[5] 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.

[6] P. E Leopold and E. I Shakhnovich. Protein folding kinetics in the dense phase. In *Proceeding of the Twenty-Sixth Hawaii International Conference on System Sciences, 1993*, volume i, pages 726– 735 vol.1. IEEE, January 1993.

[7] 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.

[8] 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.

[9] N D Socci, J N Onuchic, and P G Wolynes. Protein folding mechanisms and the multidimensional folding funnel. *Proteins*, 32(2):136–158, 1998.

[10] Hugh Nymeyer, Angel E. García, and José Nelson Onuchic. Folding funnels and frustration in off-lattice minimalist protein landscapes. *Proceedings of the National Academy of Sciences*, 95(11):5921–5928, May 1998. PMID: 9600893.

[11] V. I. Abkevich, A. M. Gutin, and E. I. Shakhnovich. Free energy landscape for protein folding kinetics: Intermediates, traps, and multiple pathways in theory and lattice model simulations. *The Journal of Chemical Physics*, 101(7):6052–6062, October 1994.

[12] Ken A. Dill and Hue Sun Chan. From levinthal to pathways to funnels. *Nature Structural & Molecular Biology*, 4(1):10–19, January 1997.

[13] D. K. Klimov and D. Thirumalai. Linking rates of folding in lattice models of proteins with underlying thermodynamic characteristics. arXiv e-print cond-mat/9805061, May 1998.

[14] J Sabelko, J Ervin, and M Gruebele. Observation of strange kinetics in protein folding. *Proceedings of the National Academy of Sciences of the United States of America*, 96(11):6031–6036, May 1999. PMID: 10339536.

[15] Jin Wang, Jose Onuchic, and Peter Wolynes. Statistics of kinetic pathways on biased rough energy landscapes with applications to protein folding. *Physical Review Letters*, 76(25):4861–4864, June 1996.

[16] Charles L. Brooks, José N. Onuchic, and David J. Wales. Taking a walk on a landscape. *Science*, 293(5530):612–613, July 2001. PMID: 11474087.

[17] J N Onuchic, H. Nymeyer, A E García, J. Chahine, and N D Socci. The energy landscape theory of protein folding: insights into folding mechanisms and scenarios. *Advances in Protein Chemistry*, 53:87–152, 2000.

[18] L S Itzhaki, D E Otzen, and A R Fersht. The structure of the transition state for folding of chymotrypsin inhibitor 2 analysed by protein engineering methods: evidence for a nucleation-condensation mechanism for protein folding. *Journal of molecular biology*, 254(2):260–288, November 1995. PMID: 7490748.

[19] 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.

[20] Hue Sun Chan and Ken A. Dill. Transition states and folding dynamics of proteins and heteropolymers. *The Journal of Chemical Physics*, 100(12):9238–9257, June 1994.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] Francesco Rao and Amedeo Caflisch. The protein folding network. *Journal of molecular biology*, 342(1):299–306, September 2004. PMID: 15313625.

[28] H S Chan and K A Dill. Compact polymers. *Macromolecules*, 22(12):4559–4573, 1989.

[29] Hue Sun Chan and Ken A Dill. The effects of internal constraints on the configurations of chain molecules. *The Journal of Chemical Physics*, 92(5):3118–3135, 1990.

[30] Nicholas D Socci and Jose Nelson Onuchic. Folding kinetics of protein like heteropolymers. *Journal of Chemical Physics*, 101(2):23, 1994.

[31] Ruxandra I Dima, Jayanth R Banavar, Marek Cieplak, and Amos Maritan. Statistical mechanics of protein-like heteropolymers. *Proceedings of the National Academy of Sciences*, 96(9):4904–4907, April 1999.

[32] K A Dill, S. Bromberg, K. Yue, K M Fiebig, D P Yee, P D Thomas, and H S Chan. Principles of protein folding–a perspective from simple exact models. *Protein Science*, 4(4):561–602, 1995.

[33] 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.

[34] Henry Cejtin, Jan Edler, Allan Gottlieb, Robert Helling, Hao Li, James Philbin, Ned Wingreen, and Chao Tang. Fast tree search for enumeration of a lattice model of protein folding. *The Journal of Chemical Physics*, 116(1):352–359, January 2002.

[35] D K Klimov and D Thirumalai. Cooperativity in protein folding: from lattice models with sidechains to real proteins. *Folding & Design*, 3(2):127–139, 1998. PMID: 9565757.

[36] Hao Li, Robert Helling, Chao Tang, and Ned Wingreen. Emergence of preferred structures in a simple model of protein folding. *Science*, 273(5275):666–669, August 1996.

[37] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, June 1953.

[38] Leandro C Oliveira, Ricardo T H Silva, Vitor B P Leite, and Jorge Chahine. Frustration and hydrophobicity interplay in protein folding and protein evolution. *The Journal of chemical physics*, 125(8):084904, August 2006. PMID: 16965054.

[39] 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.

[40] Eugene Shakhnovich and Alexander Gutin. Enumeration of all compact conformations of copolymers with random sequence of links. *The Journal of Chemical Physics*, 93(8):5967, 1990.

[41] Alan M. Ferrenberg and Robert H. Swendsen. New monte carlo technique for studying phase transitions. *Physical Review Letters*, 61(23):2635–2638, December 1988.

[42] Alan M. Ferrenberg and Robert H. Swendsen. Optimized monte carlo data analysis. *Physical Review Letters*, 63(12):1195–1198, September 1989.

Apêndice A

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 [41, 42], 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 térmica da densidade

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

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

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

$$Z(T^{'}) = \sum_{E} n(E)e^{-\frac{E}{T'}} \tag{A.2}$$

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 A.1 obtém a densidade de estados:

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

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 A.2) e a equação abaixo:

$$F(T) = -Tlog(Z) \tag{A.4}$$

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

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

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

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

A equação A.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 A.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, estado 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.

No modelo de rede cúbica, para uma sequência de 27 monômeros, o método do histograma simples satisfaz o intervalo de temperatura de interesse. Isso porque, a largura do histograma de energia é $\frac{1}{\sqrt{N}}$, onde N é o tamanho do sistema. como o sistema é pequeno o suficiente para que o histograma amostre uma grande região do espaço de fase numa temperatura, o método do histograma fornece resultados satisfatórios.

Autorizo a reprodução xerográfica para fins de pesquisa.

São José do Rio Preto, 25/08/2013

Antonio B. Oliveira Junior

Assinatura