



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

Instituto de Biociências, Letras e Ciências Exatas

Dissertação de Mestrado

Murilo Nogueira Sanches

Departamento de Física
Instituto de Biociências, Letras e Ciências Exatas (IBILCE)
Universidade Estadual Paulista (UNESP)
São José do Rio Preto – SP

2021

Murilo Nogueira Sanches

**Representação detalhada do relevo de energia de monômeros de
 β -amilóide e de proteínas de repetições**

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

Agência financiadora: CNPq – Proc. 130147/2020-6

Orientador: Prof. Dr. Vitor Barbanti Pereira Leite

São José do Rio Preto - SP

2021

S211r	<p>Sanches, Murilo Nogueira</p> <p>Representação detalhada do relevo de energia de -amiloide e de proteínas de repetições / Murilo Nogueira Sanches. -- São José do Rio Preto, 2021</p> <p>50 f.</p> <p>Dissertação (mestrado) - Universidade Estadual Paulista (Unesp), Instituto de Biociências Letras e Ciências Exatas, São José do Rio Preto</p> <p>Orientador: Vitor Barbanti Pereira Leite</p> <p>1. Superfícies de energia. 2. -amiloides. 3. Proteínas de Repetições.</p> <p>I. Título.</p>
-------	--

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca do Instituto de Biociências Letras e Ciências Exatas, São José do Rio Preto. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

Murilo Nogueira Sanches

**Representação detalhada do relevo de energia de monômeros de
 β -amilóide e de proteínas de repetições**

Comissão examinadora

Prof. Dr. Vitor Barbanti Pereira Leite

Instituto de Biociência, Letras e Ciências Exatas, UNESP – São José do Rio Preto

Orientador

Profa. Dra. Thereza Amélia Soares da Silva

Instituto de Física, USP – São Paulo

Prof. Dr. Sidney Jurado de Carvalho

Instituto de Biociência, Letras e Ciências Exatas, UNESP – São José do Rio Preto

São José do Rio Preto - SP

15 de dezembro de 2021

2021

Agradecimentos

Agradeço primeiramente à minha família, especialmente meus pais, que me proveram ótimas condições de estudo e sempre me auxiliaram incentivando meus estudos. Agradeço também à minha namorada Isabela Brunozi de Oliveira (Unesp), por sempre me apoiar e incentivar, por todo o carinho, zelo, paciência, compreensão, auxílio por sempre estar presente quando precisei, e tornando meus dias mais alegres.

Agradeço ao meu orientador Dr. Vitor Barbanti Pereira Leite (Unesp) por me aceitar em seu grupo de pesquisa e sempre auxiliar em meu projeto, assim como agradeço pelos ensinamentos transmitidos e pela confiança em mim depositada. Agradeço também a todo o grupo de pesquisa pela parceria e apoio.

Por fim, agradeço ao apoio do Instituto de Biociências, Letras e Ciências Exatas (IBILCE - Unesp) pela infraestrutura fornecida e agradeço ao CNPq que permitiu que o projeto fosse desenvolvido com todo suporte e apoio financeiro oferecido.

Resumo

O processo pelo qual uma proteína assume sua configuração funcional é denominado enovelamento e corresponde a um problema que intriga os pesquisadores desde a década de 50. Devido à importância que essa classe de macromoléculas possui e às doenças que sua má atividade acarretam, compreender a formação de sua estrutura tridimensional é uma questão fundamental. Contudo, devido à sua complexidade, o processo ocorre em um espaço de fase multidimensional, de forma que sua análise não é trivial. Esse problema se torna ainda mais complicado devido ao fato de existirem classes especiais de proteínas que apresentam certas particularidades, como é o caso das proteínas intrinsecamente desordenadas e das proteínas de repetições. Enquanto a primeira classe apresenta alta flexibilidade e a ausência de um estado nativo de referência, a segunda classe corresponde à estruturas terciárias compostas pela repetição de um mesmo domínio, e que não enovelam de forma independente mas sim por meio de um mecanismo de cooperação. Neste trabalho foram analisados dois sistemas distintos. O primeiro consiste nas dinâmicas moleculares de monômeros de β -amilóides, visando compreender suas diferentes taxas de agregações. O segundo sistema corresponde às simulações de proteínas da família de repetições de anquirina, buscando a compreensão do mecanismo de formação dessas repetições. As análises foram realizadas por meio do *Energy Landscape Visualization Method* que, usando uma métrica baseada na distância interna entre os resíduos de cada conformação, é capaz de projetar a superfície de energia em espaços 2D e 3D. Os resultados demonstraram que a propensão de agregação das diferentes espécies de β -amilóide está contida no relevo de energia livre dos monômeros, enquanto que as repetições de anquirina apresentaram diferentes rotas de enovelamento com distintas probabilidades de ocorrência.

Palavras-chaves: Superfícies de energia; β -amiloide; proteínas de repetições.

Abstract

The process by which a protein assumes its functional configuration is called folding and corresponds to a problem that has intrigued researchers since the 1950s. Due to the importance of this class of macromolecules and the diseases that their harmful activity entail, understanding the formation of its three-dimensional structure is a fundamental question. However, due to its complexity, the process takes place in a multidimensional phase space, so its analysis is not trivial. This problem becomes even more complicated due to the fact that there are special classes of proteins that have certain particularities, such as intrinsically disordered proteins and repeat proteins. While the first class presents high flexibility and the absence of a native state of reference, the second class corresponds to tertiary structures composed by the repetition of the same domain, and which does not fold independently but through a cooperation mechanism. In this work, two distinct systems were analyzed. The first consists of the molecular dynamics of β -amyloid monomers, aiming to understand their different aggregation rates. The second system corresponds to the simulations of proteins from the ankyrin repeat family, in order to understand the mechanism of formation of these repeats. The analyzes were performed using the Energy Landscape Visualization Method. This method uses a metric based on the internal distance between the residues of each conformation, and it is able to represent the ensemble of molecular conformations in an effective 2D and 3D phase space. The results show that the aggregation propensities of different species of amyloid- β are encoded in the free energy spectrum of its monomers, while the ankyrin repetitions presented different folding routes with distinct occurrence probability.

Keywords: Energy landscape; amyloid- β ; repeat proteins.

Lista de Figuras

1	Esquema do funil de enovelamento. A coordenada de reação ou parâmetro de ordem, Q , é correlacionada com a energia e indica o grau de similaridade dos estados com o estado nativo (Figura adaptada de [1]).	13
2	Representação da estrutura terciária das proteínas de repetições de anquirina, com a 3ANK correspondendo ao PDB 1N0QA, 4ANK ao PDB 1N0R e 6ANK ao PDB 1NFIE.	14
3	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 1.	18
4	Espaço de fase conformacional das estruturas obtidas por dinâmica molecular visualizado através do ELViM com a) distinção das espécies presentes na projeção com A β 40 em azul, A β 42 em vermelho e as mutações da A β 40 em verde; b) coloração em função do raio de giro, com as conformações significativas de cada região exibidas ao redor.	26
5	Projeção bidimensional do ELViM como função da coordenada q com o monômero de fibra como referência (estrutura em vermelho, PDB 2LMN). A linha vermelha indica a divisão entre as regiões de alto e baixo q_{fiber} na projeção.	27
6	(a) Perfil de energia livre em função do q_{fiber} para a A β 40 e A β 42. (b) Perfil de energia livre em função do q_{fiber} para os mutantes da A β 40.	27
7	(a) Distribuição de energia livre das estruturas da A β 40 (em cima) e da A β 42 (em baixo) na projeção do ELViM, com a região de menor energia indicada pelo G em vermelho e a região de alto q_{fiber} indicada pela elipse com o F em vermelho. (b) Distribuição média do q_{fiber} da Figura 5.	29
8	Distribuição de energia livre das mutações de A β 40 com o caminho de menor energia livre destacado em amarelo.	30
9	Diferença entre os mapas da média de contato para as regiões de menor energia livre com as regiões de alto q_{fiber} para as estruturas de A β 40, A β 42 e E22Q.	30

10	Diferença entre os mapas da média de contato para as regiões de menor energia livre com as regiões de alto q_{fiber} para as estruturas das mutações de A β 40.	31
11	Análises das trajetórias enviesadas da anquirina com três repetições. A) Perfil de energia livre em função da coordenada Qo . B) Grau de enovelamento de cada resíduo em função de Qo com a cor indicando o Qo médio local para um resíduo em relação ao conjunto de estruturas do Qo global. C) Espaço de fase conformacional projetado através do ELViM em função de Qo . D) Projeção tridimensional das trajetórias enviesadas da 3ANK em função da energia total da dinâmica com as estruturas não enviesadas destacadas em preto e o eixo Z correspondendo à coordenada Qo	33
12	Projeção bidimensional do espaço de fase conformacional da 3ANK em (A) função do raio de giro e (B) em função da coordenada Qw . (C) Projeção destacando as estruturas não enviesadas em vermelho com as conformações enviesadas em cinza.	33
13	Mapa de contato médio para as estruturas da 3ANK que correspondem as diferentes regiões da projeção do ELViM (destacadas em vermelho).	34
14	Mapa de contato médio para os estados de transição da 3ANK com os mesmos valores de Qo global. A diferença entre essas conformações combinada com o fato de que possuem valores similares de Qo é a razão por trás do aparente <i>backtracking</i> observado na Figura 11B.	34
15	A) Perfil de energia livre em função da coordenada Qo para a 4ANK. B) Grau de enovelamento de cada resíduo em função de Qo com a cor indicando o Qo médio local para um resíduo em relação ao conjunto de estruturas do Qo global. C) Projeção bidimensional do espaço de fase conformacional da 4ANK em função da coordenada Qo . D) Projeção 3D das trajetórias enviesadas em função da energia total da dinâmica com as trajetórias não enviesadas destacadas em preto.	36
16	Projeção bidimensional do espaço de fase conformacional da 4ANK em A) função do raio de giro e B) em função da coordenada Qw . C) Projeção amostrando as estruturas obtidas com os dois tipos de simulações: as conformações não enviesadas em vermelho e as enviesadas em cinza.	37

17	Mapa de contato médio das estruturas de 4ANK correspondendo às diferentes regiões da projeção do ELViM (destacadas em vermelho).	37
18	Mapa de contato médio dos estados de transição da 4ANK com valores semelhantes de Qo global. A diferença entre essas conformações combinada com o fato de que apresentam valores próximos de Qo é o motivo por trás do aparente <i>backtracking</i>	38
19	Taxa de enovelamento para a região I e região II da 4ANK para cada temperatura diferente.	39
20	Análises das trajetórias da 6ANK. A) Perfil de energia livre em função da coordenada Qo . B) Grau de enovelamento para cada resíduo em função de Qo , com a cor indicando a média do Qo local para um resíduo sobre o conjunto de estruturas de um dado Qo global. C) Espaço de fase conformacional da 6ANK visualizado através do ELViM em função de Qo . D) Projeção tridimensional das trajetórias em função da energia total da dinâmica, com o eixo Z correspondendo à coordenada Qo	40
21	Mapa de contato médio das estruturas da 6ANK correspondendo às diferentes regiões da projeção do ELViM (destacadas em vermelho).	41
22	Mapa de contato médio dos estados de transição da 6ANK com mesmos valores de Qo global. A diferença entre essas conformações combinada com o fato de que apresentam valores próximos de Qo é o motivo por trás do aparente <i>backtracking</i>	41
23	A) Grau de enovelamento de cada resíduo em função de Q . B) Projeções bidimensionais para as trajetórias enviesadas da 3ANK em função de Q^{Rn} para cada repetição. C) Projeções bidimensionais para as trajetórias não enviesadas da 3ANK em função de Q^{Rn} para cada repetição.	42
24	A) Grau de enovelamento de cada resíduo em função de Qo . B) Projeções bidimensionais para as trajetórias enviesadas da 4ANK em função de Q^{Rn} para cada repetição. C) Projeções bidimensionais para as trajetórias não enviesadas da 4ANK em função de Q^{Rn} para cada repetição.	42
25	A) Grau de enovelamento de cada resíduo em função de Qo . B) Projeções bidimensionais para as trajetórias enviesadas da 6ANK em função de Q^{Rn} para cada repetição.	43

Sumário

Trabalhos apresentados e publicações	12
1 Introdução	12
1.1 β -amilóides	13
1.2 Proteínas de repetições	14
2 Objetivos	16
3 Metodologia	17
3.1 ELViM: Energy Landscape Visualization Method	17
3.2 Método dos Múltiplos Histogramas	20
3.3 Modelo AWSEM	20
3.3.1 Dinâmicas com β -Amilóide	22
3.3.2 Dinâmicas com Proteínas de repetições	22
4 Resultados e discussões	25
4.1 β -Amilóide	25
4.2 Proteínas de repetições	29
4.2.1 3 repetições	30
4.2.2 4 repetições	35
4.2.3 6 repetições	37
5 Conclusões	43
Referências	45
A Descrição da representação <i>coarse-grained</i>	49
B Campo de força do AWSEM	49

Trabalhos apresentados e publicações

Apresentações em painel

- "Comparação de potenciais não específicos com potenciais baseados em estrutura no estudo de enovelamento de proteínas" - 9^a Escola de Modelagem Molecular em Sistemas Biológicos, Petrópolis de 20 a 24/08/2018.

- "Análise de enovelamento de proteínas por diferentes potenciais" - 1^a fase do 30^o Congresso de Iniciação Científica da UNESP, São José do Rio Preto de 30 a 31 de outubro de 2018.

- "Análise de enovelamento de proteínas por diferentes potenciais" - 2^a fase do 30^o Congresso de Iniciação Científica da UNESP, São Paulo de 21 a 22 de novembro de 2018.

- "Estudo energético das conformações de β -amilóide" - 1^a fase do 31^o Congresso de Iniciação Científica da UNESP, São José do Rio Preto, de 3 a 5 de setembro de 2019.

- "Estudo energético das conformações de β -amilóide" - 2^a fase do 31^o Congresso de Iniciação Científica da UNESP, Sorocaba, de 25 a 26 de novembro de 2019.

- "Probing propensities of amyloid- β fiber formation through analysis of its energy landscape" - Encontro de Outono da Sociedade Brasileira de Física, de 21 a 25 de junho de 2021.

Apresentações orais

- "Caracterização conformacional e visualização do perfil energético em estudos de proteínas intrinsecamente desordenadas" - Semana da Física, São José do Rio Preto, de 4 a 11 de novembro de 2019.

Artigo publicado

B. da Silva, F.; M. de Oliveira, V.; Sanches, M. N.; Contessoto, V. G.; Leite, V. B. "Rational Design of Chymotrypsin Inhibitor 2 by Optimizing Non-Native Interactions" Journal of Chemical Information and Modeling 2019, 60, 982–988.

Artigos submetidos

Sanches, M. N.; Knapp, K.; de Oliveira, A. B.; Wolynes, P. G.; Onuchic, J. N.; Leite, V. B. "Examining the Ensembles of Amyloid- β Monomer Variants and their Propensities to Form Fibers Using an Energy Landscape Visualization Method" Journal of Physical Chemistry B.

Sanches, M. N.; Parra, R. G.; Viegas, R. G.; de Oliveira, A. B.; Wolynes, P. G.; Ferreira, D. U.; Leite, V. B.; "Resolving the Fine Structure in the Energy Landscapes of Repeat Proteins" Quarterly Reviews of Biophysics Discovery.

1 Introdução

Compreender os mecanismos biológicos é um dos principais objetivos e desafios da ciência moderna. Nesse intuito, a principal classe de macromoléculas a ser estudadas são as proteínas. Responsáveis por inúmeras funções essenciais para a manutenção da vida, grande parte de seus comportamentos permanece sem uma solução definitiva. O processo no qual a cadeia de aminoácidos assume uma estrutura funcional é denominado enovelamento, e é o principal alvo de estudos. Essas pesquisas procuram compreender como a informação presente em sua sequência de resíduos é capaz de guiar as proteínas ao seu estado nativo. Entender esse processo é de extrema importância, uma vez que diversas doenças estão diretamente relacionadas ao mal enovelamento de algumas proteínas, como a doença de Alzheimer [2]. Estudos revelaram que a superfície de energia dos estados assumidos pela proteína tem um papel fundamental para determinar o caminho até seu estado nativo [1]. Na década de 90, Onuchic propôs um modelo capaz de explicar o enovelamento de uma proteína por meio de sua superfície de energia [3]. Através de um mapeamento estatístico das possíveis conformações de uma proteína, a superfície apresenta o estado nativo como o mínimo global de energia, de forma que sua topologia seja afunilada, com múltiplos caminhos que levam a proteína ao estado estrutural nativo. Devido a esses caminhos, o funil apresenta rugosidades em sua superfícies, que correspondem a mínimos locais de energia, responsáveis por armadilhar essas macromoléculas em um estado parcialmente enovelado [4, 5].

Todavia, existem diversas famílias de proteínas que apresentam especificidades que tornam seu enovelamento de difícil compreensão. Seja por não possuírem um estado nativo ou por enovelarem através de interações específicas entre determinadas regiões, suas superfícies de energia são demasiadamente complexas. É nessa classe que se encontram dois tipos de moléculas de grande interesse: as proteínas intrinsecamente desordenadas (IDPs) e as proteínas de repetições. Enquanto que as IDPs não possuem um estado nativo ou de referência, as proteínas de repetições apresentam a união de domínios idênticos que se formam através de um processo cooperativo.

Para analisar essas superfícies de forma convencional, é necessário o uso de inúmeras coordenadas de reação, o que acaba por dificultar ou até mesmo inviabilizar os estudos. É nesse contexto que o *Energy Landscape Visualization Method* (ELViM) surge como uma alternativa que possibilita a análise e o entendimento do modelo estudado. Nesse trabalho, foram estudadas as mudanças conformacionais nos monômeros de β -amilóides e

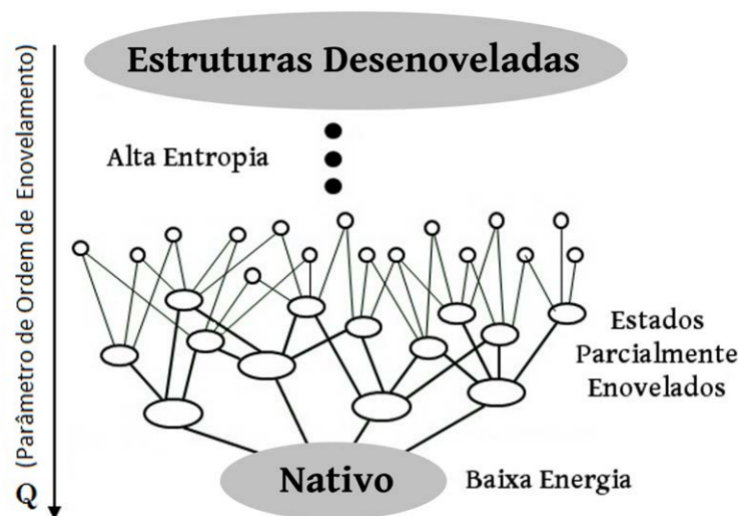


Figura 1: Esquema do funil de enovelamento. A coordenada de reação ou parâmetro de ordem, Q , é correlacionada com a energia e indica o grau de similaridade dos estados com o estado nativo (Figura adaptada de [1]).

o enovelamento das repetições de anquirina. Esse estudo foi realizado em colaboração com a *Rice University*, o *Instituto de Química Biológica de la Facultad de Ciencias Exactas y Naturales*, e o *European Molecular Biology Laboratory*.

1.1 β -amilóides

Nos últimos 20 anos, o entendimento do processo de agregação das proteínas que causam a doença de Alzheimer tem sido um grande foco de estudo. Essa doença possui uma relação direta com a proteína β -amilóide ($A\beta$), uma IDP que surge de forma mais abundante como um polímero de 40 ou 42 resíduos através da clivagem da proteína amilóide precursora [6, 7]. A relação dessa IDP com a doença de Alzheimer reside na tendência de agregação, formando placas que podem bloquear a sinalização entre as células nas sinapses [2]. Contudo, estudos recentes sugerem que essa tendência pode estar contida no espectro energético das formas monoméricas [8]. Um dos indícios está no fato de que, apesar de diferirem em apenas dois aminoácidos, as duas sequências apresentam agregações distintas. Enquanto que a abundância natural da $A\beta_{40}$ é de 10/1 em relação à $A\beta_{42}$, ela somente está presente em 1/3 das placas formadas, de modo que a agregação da $A\beta_{42}$ ocorre de forma muito mais rápida e espontânea [9, 10]. Esse resultado corrobora com estudos anteriores que indicam que a taxa de agregação da $A\beta_{42}$ está relacionada com o

conteúdo de folhas- β presente em sua forma monomérica [11]. Junto a esse estudo também foram identificadas mutações na sequência de 40 aminoácidos que apresentam um aumento na taxa de agregação [12–15].

O ELViM foi utilizado nesse trabalho para analisar as mudanças conformacionais dos monômeros de A β 40, A β 42 e de seis mutações da A β 40: A21G, D7N, D23N, E22G, E22K e E22Q, a fim de compreender suas taxas de agregação. Essa técnica de visualização da superfície de energia já se provou capaz de discernir essas superfícies em IDPs, como demonstrado em trabalho com a PAGE4 [16].

1.2 Proteínas de repetições

Em geral, um domínio proteico estrutural é um módulo relacionado à estabilidade e ao processo de enovelamento de uma proteína. Contudo, existe uma classe especial que apresenta repetições de um mesmo domínio que não enovelam de forma independente, mas sim através de uma cooperação que leva ao estado nativo por um caminho específico [17]. Dentre essas proteínas, destacam-se as chamadas de solenoides, compostas em geral por 20 a 40 trechos de aminoácidos semelhantes que se enovelam em estruturas alongadas [18]. A Figura 2 ilustra três exemplos de solenoides para a família de repetições de anquirina.



Figura 2: Representação da estrutura terciária das proteínas de repetições de anquirina, com a 3ANK correspondendo ao PDB 1N0QA, 4ANK ao PDB 1N0R e 6ANK ao PDB 1NFIE.

Uma das particularidades dessas proteínas é o fato de que essas matrizes são estabilizadas apenas por interações dentro de cada repetição ou entre repetições adjacentes, sem a interação entre os contatos de resíduos muito distantes na cadeia [19, 20]. Essa arquitetura contrasta com a típica de proteínas globulares, nos quais a cadeia se dobra para formar interações entre resíduos que se encontram distantes na sequência. Dessa

forma, o processo de enovelamento, bem como a superfície de energia, das proteínas de repetições difere do enovelamento de proteínas globulares, uma vez que as interações dos resíduos distantes em sequência frequentemente desempenham um papel central em definir a topologia da proteína. [21].

Nesse trabalho foi analisado o enovelamento das matrizes de repetições por meio do ELViM [22], utilizado para explorar as superfícies de energia da proteína anquirina de três, quatro e seis repetições (3ANK, 4ANK e 6ANK), assim como demonstrar a cooperação entre as diferentes repetições e a forma como essas proteínas varrem o espaço conformacional.

2 Objetivos

Esse trabalho tem por objetivo geral analisar as superfícies de energia dos monômeros de β -amilóide e da família de repetições de anquirina, em especial as de três, quatro e seis repetições, para compreender o processo pelo qual as $A\beta$ s iniciam sua agregação e as proteínas de repetições assumem seu estado enovelado nativo.

3 Metodologia

3.1 ELViM: Energy Landscape Visualization Method

O enovelamento de proteínas é usualmente representado em termos de uma coordenada de reação efetiva associada ao estado nativo único. No entanto, dependendo do sistema estudado nem sempre essa abordagem é possível. Para conseguir visualizar esse processo uma alternativa é empregar o ELViM, um método que utiliza uma métrica baseada na distância interna entre aminoácidos para descrever as diferenças entre duas conformações [23].

Matriz de dissimilaridade

O primeiro passo para a construção da visualização é definir uma métrica que descreva de maneira robusta a informação obtida na simulação. As posições dos aminoácidos são essenciais para a elaboração da métrica, cujas coordenadas internas caracterizam as diferentes conformações. De maneira formal, a medida de comparação entre duas conformações (k, l) pode ser escrita como:

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

onde $r_{i,j}^{k(l)}$ é a distância entre os aminoácidos i e j da conformação $k(l)$, 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 σ_0 e ϵ valores ajustáveis. Essa equação foi apresentada primeiramente por Wolynes e colaboradores, no qual foi utilizada para a predição da estrutura de proteínas [24]. No ELViM, essa medida é calculada para todos os pares de estruturas k e l de uma dada trajetória, resultando em uma matriz simétrica $n \times n$, onde n é o número total de conformações analisadas. Na Figura 3 é apresentado um esquema de como as distâncias são calculada para duas conformações diferentes (no caso, k e l).

Nos estudos presentes nesse trabalho, essa coordenada foi utilizada não só como métrica no ELViM mas também como coordenada de reação para realizar e analisar as simulações. No caso das β -amilóides, utilizou-se o monômero de uma fibra de A β 40 (PDB: 2LMN) como a estrutura de referência, definindo a coordenada de reação q_{fiber} , enquanto que para as proteínas de repetições a estrutura de referência foi o próprio estado nativo da trajetória, definindo a coordenada de reação global Qw .

Definido isso, para encontrar o valor de dissimilaridade e montar a matriz, é utilizada a expressão:

$$\delta^{k,l} = 1 - q^{k,l}, \quad (2)$$

sendo $\delta^{k,l}$ um valor compreendido entre 0 e 1, no qual valores próximos de 0 significam que as duas conformações são muito similares, e valores próximos a 1 significam que elas são estruturalmente bem diferentes.

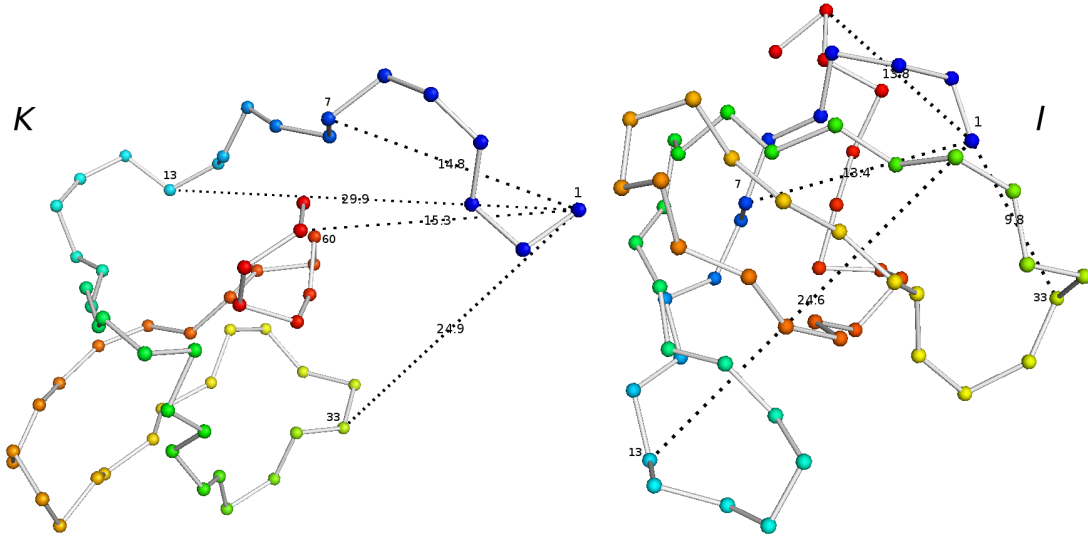


Figura 3: Exemplo de duas estruturas (k e l) onde são indicadas as distâncias internas ($r_{i,j}^{k(l)}$) entre os aminoácidos (1, 7, 13, 33 e 60). As diferenças estruturais estão relacionadas com as variações destas distâncias entre todos os aminoácidos, de acordo com a Equação 1.

Projeção multidimensional

A etapa mais complexa para a criação deste modelo é a redução do espaço multidimensional original 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 [25].

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 , 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 $|\delta(x_i, x_j) - d(f(x_i), f(x_j))|$ 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 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 método, a técnica *Force Scheme* [25], 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. Após a inicialização, a técnica *Force Scheme* realiza iterações para aproximar as distâncias entre os objetos projetados, a partir 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 $\vec{v}$, acrescentando-se a fração de Δ , onde Δ é dado por:

$$\Delta = \left\{ \left[1 + (2j + 1)^{\frac{1}{2j+1}} \right] \right\} - 1, \quad (3)$$

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.

3.2 Método dos Múltiplos Histogramas

Para calcular a energia livre de todos os sistemas, utilizou-se o Método dos Múltiplos Histogramas (WHAM) [26,27]. O método consiste em determinar a densidade de estados Ω em uma dada temperatura T , como a soma ponderada de densidades de estados estimadas Ω_i , obtidas por meio de diversas simulações em temperaturas T_i , diferentes de T . Os pesos, usados para o cálculo da soma ponderada, são obtidos pela minimização dos erros estatísticos, $\delta^2\Omega$, por meio da sobreposição das distribuições de probabilidade, para cada simulação executada em uma temperatura T_i .

Para o caso das β -amilóides, utilizou-se o WHAM especificamente para repesar as conformações e descrever a superfície do ELViM através da energia livre. Para todas as trajetórias e em todas as temperaturas (200K, 250K, 300K, 350K e 400K) foram obtidas as densidades de estado $\Omega(E)$.

A probabilidade de ocorrência de um estado com energia i na temperatura inversa β é dado por:

$$P(E_i, \beta) = \frac{\Omega(E_i)e^{-\beta E_i}}{\Omega_T} \quad (4)$$

onde

$$\Omega_T = \sum_i \Omega(E_i) \quad (5)$$

Os estados foram então divididos em *bins*; em 1D para os perfis de energia livre em função da coordenada q_{fiber} , e em 2D para a projeção do ELViM, formando uma representação em grade. Considerando todos os n estados dentro de um determinado *bin*, a energia livre desse conjunto será:

$$F = -\frac{1}{\beta} \ln \left(\sum_{i=1}^n \frac{\Omega(E_i)e^{-\beta E_i}}{\Omega_T} \right) \quad (6)$$

3.3 Modelo AWSEM

Todas as dinâmicas presentes nesse trabalho foram realizadas utilizando o AWSEM (*Associative memory, Water mediated, Structure and Energy Model*), um pacote de

simulação de proteínas que emprega uma representação *coarse-grained*, no qual a posição e orientação de cada aminoácido é descrita pela posição de seus átomos de carbono alfa (C_α), carbono beta (C_β) e oxigênio (O). A posição dos demais átomos da cadeia é calculada assumindo uma geometria ideal. A descrição completa do modelo estrutural e do campo de força estão presentes nos Apêndices A e B.

A função de energia total do modelo utilizado é dada por:

$$V_{total} = V_{backbone} + V_{contact} + V_{burial} + V_{helical} \quad (7)$$

$V_{backbone}$ é o termo responsável por manter a geometria da cadeia principal. A forma completa desse potencial de cadeia é expressa por:

$$V_{backbone} = V_{con} + V_{chain} + V_\chi + V_{rama} + V_{excl} \quad (8)$$

V_{con} é o potencial que garante a conectividade da cadeia através de uma série de potenciais harmônicos (ver Apêndice para maiores informações). A forma correta dos ângulos de ligações são determinadas através do potencial V_{chain} . V_χ , V_{rama} e V_{excl} são responsáveis pela quiralidade do átomo de C_α , a distribuição correta dos ângulos diedrais e as interações de volume excluído, respectivamente.

$V_{contact}$, V_{burial} , and $V_{helical}$ são potenciais baseados em diferentes aspectos da física de proteínas. $V_{contact}$ é um termo dependente do tipo de interações terciárias que age entre os pares de resíduos que estão separados por dez ou mais aminoácidos, de forma que a força desse potencial também depende dessa distância de separação e da densidade local. O termo de hidrofobicidade (V_{burial}) representa a tendência de um tipo de aminoácido ser encontrado na região interna ou na superfície da proteína. Por fim, $V_{helical}$ é um termo explícito de ligação de hidrogênio que atua entre o oxigênio carbonílico do resíduo i e o hidrogênio amídico do resíduo $i + 4$. A força de interação depende da propensão de formação de hélice de ambos os resíduos participantes na interação [28].

As simulações com as β -amilóides foram realizadas pela aluna de PhD Kaitlin Knapp, da *Rice University*, e as simulações com as proteínas de repetições foram executadas pelo Dr. Gonzalo R. Parra, do *European Molecular Biology Laboratory*. Os potenciais do pacote, bem como os detalhes das simulações, estão descritos a seguir para cada proteína estudada.

3.3.1 Dinâmicas com β -Amilóide

Para as dinâmicas realizadas com as β -amilóides foram utilizados os potenciais padrões do AWSEM. Foi utilizado uma abordagem de solvente implícito e as características de exposição ao solvente de cada resíduo de aminoácido. As simulações de A β 42 foram iniciadas utilizando uma estrutura de solução determinada por ressonância magnética nuclear (PDB 1Z0Q). No caso da A β 40 utilizou-se a mesma estrutura inicial que a A β 40 porém sem os dois últimos resíduos. Cada simulação utilizou um passo de 2 femto segundos (fs), com minimização de energia realizada para os primeiros 10^5 passos seguida pela simulação de 10^6 passos. As dinâmicas foram realizadas na temperatura de 200K, 250K, 300K, 350K e 400K, com duas simulações independentes para cada temperatura. O resultado consiste em 0.4 μ s de simulação e 10^5 estruturas amostradas. Para as estruturas iniciais dos mutantes da A β 40 primeiro foi executado um *annealing*, para evitar qualquer tendência na formação de estrutura secundária, uma vez que as estruturas monoméricas dos mutantes ainda não foram solucionadas. O procedimento das simulações dos mutantes segue as mesmas configurações utilizadas para a A β 40 e A β 42.

3.3.2 Dinâmicas com Proteínas de repetições

Para as simulações com as proteínas de repetições, além dos potenciais padrões, também foi utilizado o potencial AMH-G $\bar{o}$ que tem a particularidade de incluir um termo de não aditividade que permite simular de forma mais realista a cooperatividade entre interações nativas, especialmente ao amostrar estados de transição [29]. A estrutura geral do Hamiltoniano desse modelo é composta por um termo que modela a geometria do *backbone* da cadeia polipeptídica e outro termo que modela as interações entre os resíduos de forma não aditiva.

$$V_{amh} = V_{backbone} - V_{na} \quad (9)$$

no qual $V_{backbone}$ é um dos potenciais básicos do AWSEM, mencionado anteriormente e descrito em detalhes no Apêndice B, enquanto que a componente energética que corresponde ao potencial V_{na} depende de termos de interação do tipo gaussiano para interações nativas e possui a seguinte forma:

$$V_{na} = -\frac{1}{2} \sum_i \sum_j |\epsilon_{ij}(r_{ij})|^p \quad (10)$$

onde

$$\epsilon_{ij}(r_{ij}) = -\left|\frac{\epsilon}{a}\right|^{1/p} \theta(r_c - r_{ij}^N) \gamma_{ij} \exp\left[-\frac{(r_{ij} - r_{ij}^N)^2}{2\sigma_{ij}^2}\right] \quad (11)$$

Os índices i e j indicam todos os átomos de C_α e C_β respectivamente, r_{ij} é a distância entre os átomos i e j , r_c é o parâmetro de *cut-off*, σ_{ij} é o mesmo parâmetro presente na Equação 1, p é a potência desejada para a não-aditividade e γ_{ij} é o peso desejado para as interações. Todos esses parâmetros foram obtidos por resultados experimentais e estão tabelados no trabalho de Wolynes et. al. [29]. Por fim, a é o fator de normalização e é dado por:

$$a = \frac{1}{8N} \sum_i \left| \sum_j \gamma_{ij} \theta(r_c - r_{ij}^N) \right|^p \quad (12)$$

Para iniciar as dinâmicas primeiro foi realizado um *annealing* para estimar o T_f (temperatura de *melting*). As variações de temperatura foram feitas em mais de 10^6 passos entre os quais o intervalo de tempo simulado foi de 3 fs. Uma vez que o T_f foi estimado, as simulações foram realizadas a esta temperatura utilizando o método de amostragem de *Umbrella Sampling*, com Qw como a coordenada coletiva que assume valores entre 0 (totalmente desenovelado) e 1 (totalmente enovelado), formando as trajetórias enviesadas. Esse método de amostragem funciona ancorando um potencial a um determinado valor de Qw , o que permite melhor explorar aquelas estruturas que possuem um Qw próximo a esse valor. O intervalo da coordenada Qw foi dividido em 40 janelas igualmente espaçadas que simularam 10^6 passos a 3 fs. Ao final das dinâmicas as 40 simulações foram integradas utilizando o WHAM [27] para obter os parâmetros termodinâmicos, como a energia livre.

Após as análises iniciais, o autor desse trabalho realizou novas simulações no T_f para a proteína de três e quatro repetições, com o mesmo potencial AMH-G $\bar{o}$ que as anteriores, porém sem o método de *Umbrella Sampling*, de modo a permitir que a proteína variasse suas conformações sem a restrição das janelas, formando um novo conjunto de trajetórias não enviesadas. Para a 3ANK foram executadas três réplicas partindo da estrutura nativa em 300K e indo até 700K em 3×10^9 passos. Para a 4ANK as configurações foram as mesmas utilizadas para a 3ANK.

Por fim, para analisar as trajetórias com e sem viés, foi escolhida uma outra coordenada que já foi bem estabelecida e frequentemente utilizada como coordenada de reação: a fração de contatos nativos, aqui denominada de Q_o [30] e definida como:

$$Q_o(X) = \frac{1}{S} \sum_{i,j \in S} \frac{1}{1 + \exp [\beta^0(r_{ij}(X) - \lambda r_{ij}^0)]} \quad (13)$$

onde $r_{ij}(X)$ é a distância entre os resíduos i e j em uma conformação X , r_{ij}^0 é a distância entre os aminoácidos no par correspondente no estado nativo, S é o conjunto de todos os pares de contatos nativos (i, j) da estrutura nativa, β^0 é um parâmetro de amortecimento e λ é um fator que leva em consideração as oscilações dos contatos. Os aminoácidos que realizam contatos no estado nativo são definidos como separados por quatro ou mais resíduos na estrutura primária e no qual $r_{ij}^0 < 1.2$ nm. Os parâmetros da equação 13 são $\beta^0 = 50 \text{ nm}^{-1}$ e $\lambda = 1.2$ nm, como sugerido para o modelo AWSEM [31]. Essa nova coordenada de reação foi escolhida devido à sensibilidade do Q_w , que super-amostrou caminhos na superfície de energia do ELViM durante os resultados iniciais.

4 Resultados e discussões

4.1 β -Amilóide

O primeiro trabalho consiste na análise das mudanças conformacionais dos monômeros β -amilóide, sendo a continuação de um estudo iniciado durante a iniciação científica. Por se tratar de uma proteína intrinsecamente desordenada, sua análise requer o uso de múltiplas coordenadas de reações. Contudo, utilizando o ELViM foi possível comparar as superfícies energéticas das espécies selecionadas. Primeiro criou-se uma única e grande projeção contendo as estruturas da A β 40, A β 42 (formando o espaço conformacional de referência) e as seis mutações de A β 40: A21G, D7N, D23N, E22G, E22K e E22Q. Essa projeção incluiu conformações de A β 40 e A β 42 amostradas em 250K, 300K e 350K, para garantir que a amostragem do espaço de fase conformacional fosse o mais detalhada quanto possível. Todavia, o ELViM possui uma limitação quanto ao número de estruturas, por isso as temperaturas de 200K e 400K não foram incluídas na projeção, mas foram utilizadas na obtenção da densidade de estados, como será explicado mais a frente. Desejando examinar os mutantes na mesma análise, restringimos suas conformações escolhidas àquelas obtidas em 300K, que são mais significativas e melhor ponderadas. A Figura 4A apresenta essa projeção com as diferentes espécies separadas pela cor enquanto que a Figura 4B exibe a projeção completa em função do raio de giro. Através desse resultado é constatado que a maior parte dos pontos da A β 40 estão localizados no lado direito da projeção, enquanto que os mutantes e a A β 42 são mais amplamente espaçados. As estruturas ao redor da projeção em (b) exemplificam as conformações de cada região, o que demonstra que as regiões mais populadas por A β 40 correspondem a clusters de baixo raio de giro, compostos por estruturas mais compactas e que apresentam formação de hélice- α , enquanto que as regiões de maior R_g e estruturas mais abertas correspondem aos pontos da A β 42 e das mutações.

Esse resultado está de acordo com a Figura 5, que agora exibe a mesma projeção em função da coordenada q_{fiber} , calculada utilizando o monômero de uma fibra de A β 40 como referência (para essa análise foram desconsiderados os dois últimos resíduos das estruturas de A β 42). Nota-se que as áreas populadas por A β 40 correspondem às regiões de baixo q_{fiber} enquanto que grande parte dos pontos de A β 42 e dos mutantes estão presentes em zonas de alto q_{fiber} , o que pode explicar o porquê deles apresentarem maior propensão para formar as fibras e o motivo da A β 40 necessitar de uma configuração pré-fibrilar para

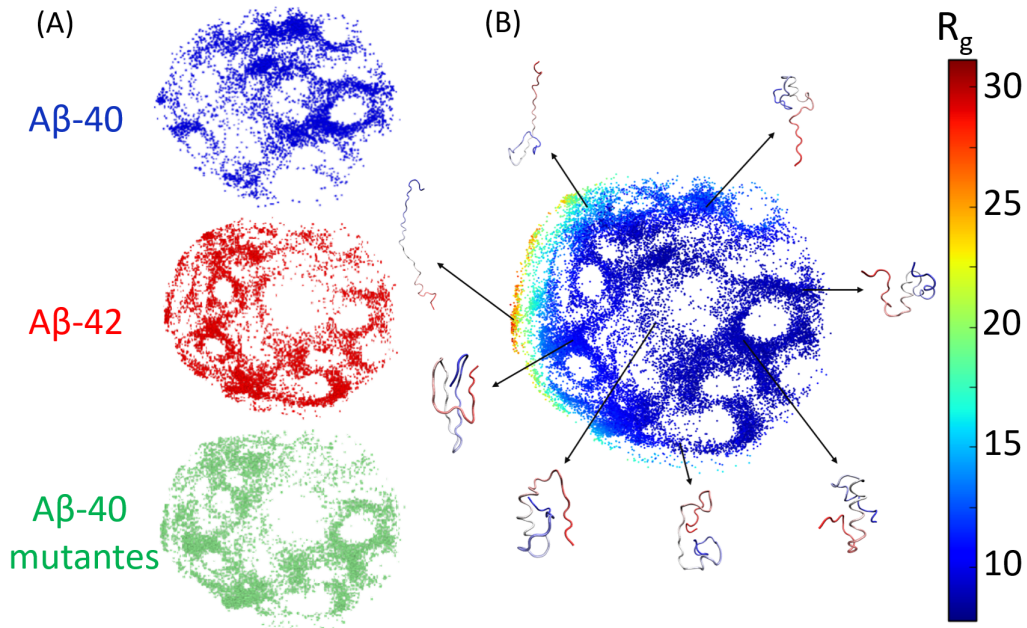


Figura 4: Espaço de fase conformacional das estruturas obtidas por dinâmica molecular visualizado através do ELViM com a) distinção das espécies presentes na projeção com $A\beta 40$ em azul, $A\beta 42$ em vermelho e as mutações da $A\beta 40$ em verde; b) coloração em função do raio de giro, com as conformações significativas de cada região exibidas ao redor.

iniciar sua agregação.

Os valores de energia livre foram então calculados para confirmar essa ideia. Utilizando o WHAM construiu-se os perfis de energia livre presentes na Figura 6 em função do q_{fiber} de cada espécie, o que demonstra que a barreira energética da $A\beta 40$ é muito maior do que a da $A\beta 42$ e as das mutações. A Tabela 1 apresenta os valores dos perfis de energia livre, com o $\Delta F^\ddagger$ indicando o valor da barreira energética, enquanto que o ΔF indica a posição do estado fundamental. Apesar da alta barreira, a mutação D23N apresenta um comportamento semelhante ao da $A\beta 42$, como será discutido a seguir.

O próximo passo para analisar a distribuição das estruturas na projeção do ELViM consistiu em repesar as conformações utilizando o microestado $\Omega(E)$ obtido através do WHAM. A Figura 7 exibe o resultado em uma projeção em grade no qual a cor de cada bin indica a energia livre do conjunto de estruturas contido dentro dele. Com isso foi possível confirmar que a região de hélice- α da Figura 4 corresponde ao estado fundamental (menor energia livre) da $A\beta 40$, como indicado pelo G em vermelho na parte superior da Figura 7A, enquanto que a região das estruturas semelhantes à fibra (identificada por meio da Fig.

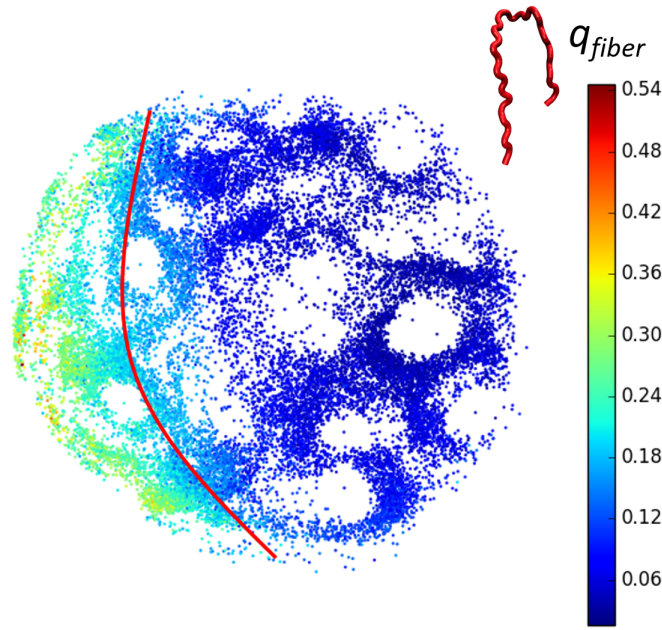


Figura 5: Projeção bidimensional do ELViM como função da coordenada q com o monômero de fibra como referência (estrutura em vermelho, PDB 2LMN). A linha vermelha indica a divisão entre a regiões de alto e baixo q_{fiber} na projeção.

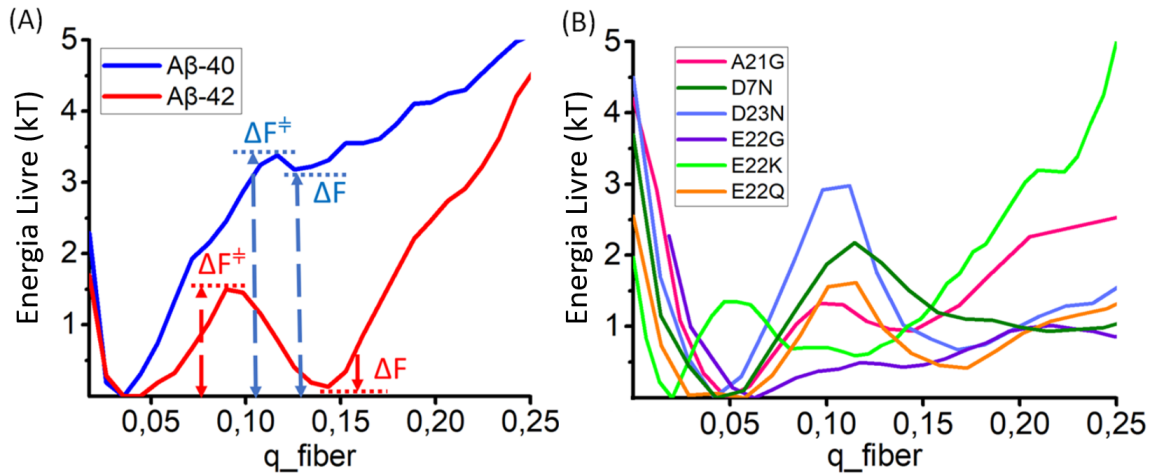


Figura 6: (a) Perfil de energia livre em função do q_{fiber} para a A β 40 e A β 42. (b) Perfil de energia livre em função do q_{fiber} para os mutantes da A β 40.

7B) está destacada com a elipse e com o F em vermelho. As linhas em vermelho e amarelo da Fig. 7A superior indicam as rotas mais prováveis (o caminho de menor energia livre) que a A β 40 utiliza para assumir as conformações semelhantes à fibra. A parte inferior da Figura 7A exibe a projeção em grade para as estruturas da A β 42, na qual a energia livre indica que o estado fundamental está muito mais próximo da região de fibra do que no

Espécie	$\Delta F(\text{kT})$	$\Delta F^\ddagger(\text{kT})$
A β 40	3.2	3.7
A β 42	0.2	1.6
A21G	0.9	1.3
D7N	0.9	2.1
D23N	0.9	2.9
E22G	0.8	0.9
E22K	1.1	2.5
E22Q	0.4	1.5

Tabela 1: Energia Livre

caso da A β 40, com o caminho (linha amarela) sendo mais curto e simples, o que implica que as conformações mais estáveis já possuem alguns dos contatos necessários para que a proteína adote sua configuração presente na fibra. A localização dos estados fundamentais para ambos os peptídeos confirma que a diferença na energia livre entre esses estados e os estados propensos à agregação é menor no caso da A β 42 do que no da A β 40, o que está de acordo com a Figura 6 e com o indício de que a propensão de agregação está codificada no espectro de energia livre do monômero [8]. O mesmo é observado para as mutações na Figura 8, com todos os estados fundamentais localizados na mesma região que o da A β 42 e, como mostram os caminhos em amarelo, seguem um comportamento semelhante ao indicado pelos resultados anteriores.

Com base neste resultado, foram calculadas as diferenças nos mapas de contato feitos com as estruturas localizadas no bin de menor energia livre e as localizadas na região da fibra para cada espécie. Para essa análise os dois últimos resíduos da A β 42 foram removidos para que a comparação entre os mapas fosse possível. A Figura 9 apresenta essas diferenças para a A β 40, A β 42 e a mutação E22Q, no qual fica claro que A β 40 precisa primeiro desfazer a hélice- α para então iniciar a formação dos primeiros contatos de folha- β ,

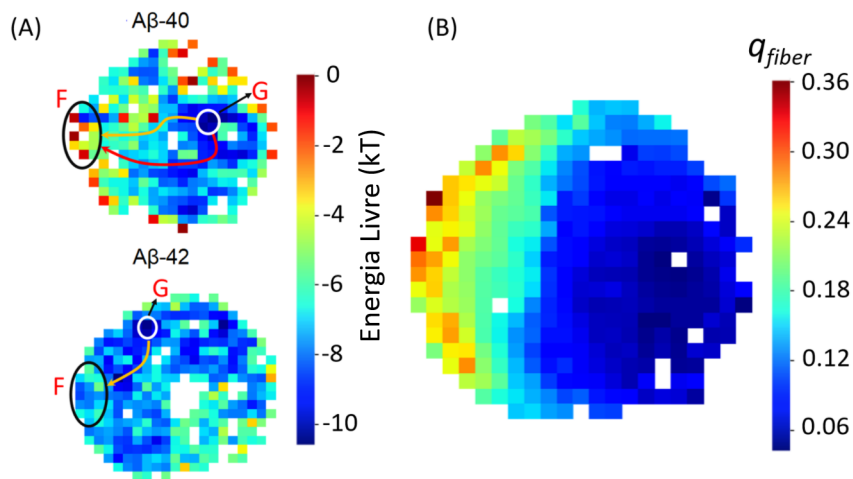


Figura 7: (a) Distribuição de energia livre das estruturas da A β 40 (em cima) e da A β 42 (em baixo) na projeção do ELViM, com a região de menor energia indicada pelo G em vermelho e a região de alto q_{fiber} indicada pela elipse com o F em vermelho. (b) Distribuição média do q_{fiber} da Figura 5.

enquanto que para a A β 42 ocorre apenas um "deslizamento" na formação da fibra, visível nos contatos paralelos entre os resíduos 15 a 30, de forma que a maioria dos contatos finais realmente já se encontram formados. Por último há a mutação E22Q, que apesar de não possuir suas folhas- β totalmente formadas em seu estado fundamental, também não apresenta contatos de hélice- α , de forma que para assumir a conformação de fibra não é necessário superar uma alta barreira energética, como visto na Figura 6. Os mapas de contato, junto às projeções do ELViM, confirmam que a A β 42 possui uma maior propensão em formar folhas- β do que a A β 40 e seus mutantes, algo que já havia sido notado em trabalhos anteriores com monômeros de amilóides [32]. A Figura 10 exibe a diferença dos mapas de contato para as demais mutações, nos quais os resultados são semelhantes aos observados para a E22Q.

4.2 Proteínas de repetições

Os resultados com as proteínas de repetições foram divididos de acordo com o grau de complexidade do sistema estudado, isto é, de acordo com o número de repetições presente em cada proteína.

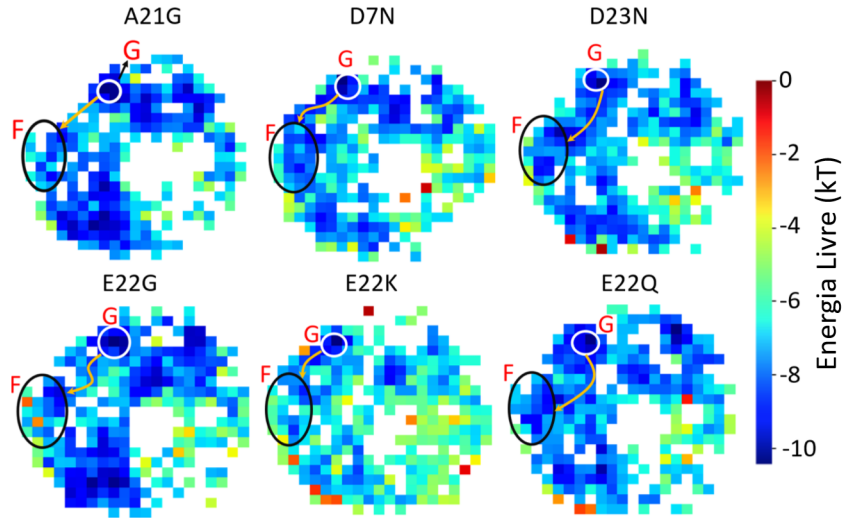


Figura 8: Distribuição de energia livre das mutações de A β 40 com o caminho de menor energia livre destacado em amarelo.

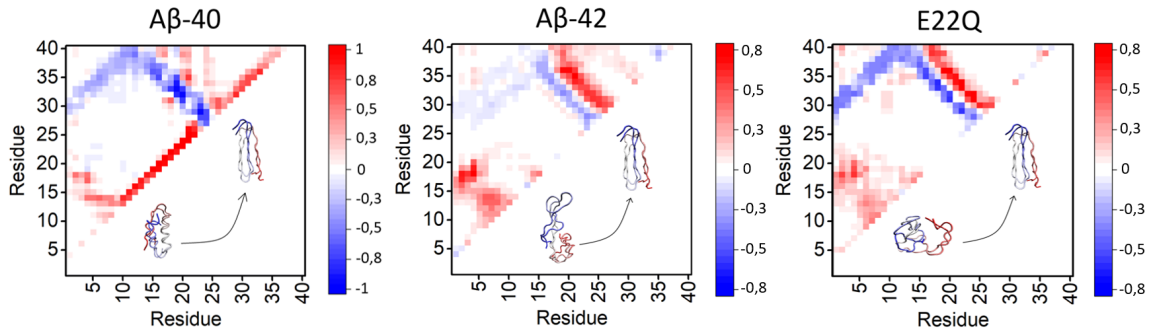


Figura 9: Diferença entre os mapas da média de contato para as regiões de menor energia livre com as regiões de alto q_{fiber} para as estruturas de A β 40, A β 42 e E22Q.

4.2.1 3 repetições

O estudo do enovelamento da anquirina de três repetições (PDB 1N0QA) se iniciou a partir das dinâmicas moleculares realizadas por Gonzalo R. Parra. A Figura 11A apresenta um comportamento típico de uma proteína de dois estados, com os estados enovelado e desenovelado separados por uma barreira igual a 19.3 kT, localizada por volta de $Q_o \approx 0.35$, com o estado enovelado em $Q_o \approx 0.85$. A fim de investigar o mecanismo de enovelamento, foi analisado como o valor de Q_o local evolui para cada resíduo em função da coordenada da reação Q_o global (Figura 11B). Nessa figura, o eixo y corresponde ao Q_o global, enquanto o mapa de cores corresponde ao Q_o local, que é calculado pela média do Q_o para cada resíduo sobre todo o conjunto de conformações com um dado valor de Q_o .

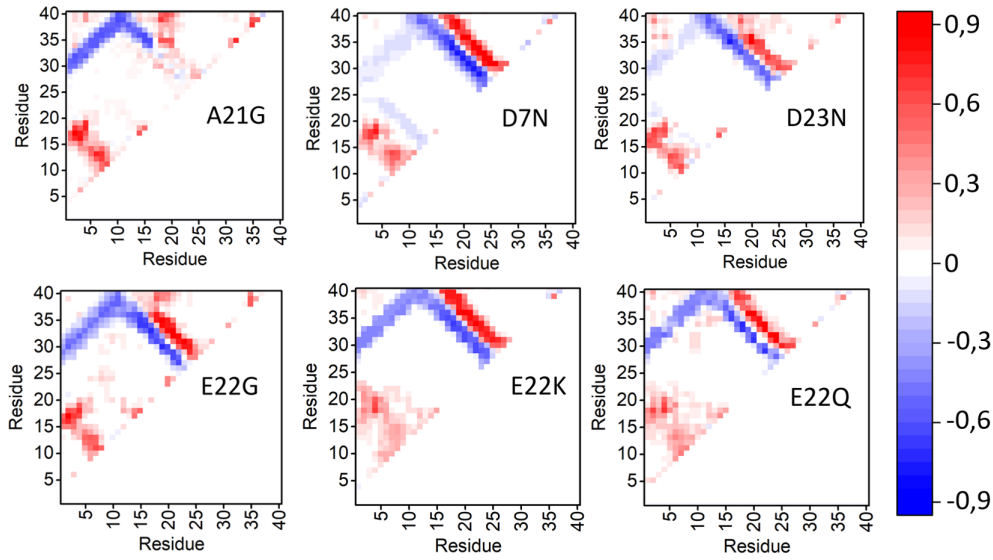


Figura 10: Diferença entre os mapas da média de contato para as regiões de menor energia livre com as regiões de alto q_{fiber} para as estruturas das mutações de A β 40.

global. É possível perceber que as interações nativas não são distribuídas simetricamente. Os contatos nativos começam a se formar através de alguns resíduos nas repetições R1 e R2, conforme Qo se aproxima do estado de transição, por volta de 0.3. Conforme o Qo global aumenta, a repetição R3 começa a enovelar em $Qo \approx 0.6$. As Figuras 11A e 11B demonstram que o enovelamento é altamente cooperativo, com a nucleação iniciando em algumas regiões de R2 e em algumas partes da região do N-terminal. Contudo, é importante notar que o Qo local de alguns resíduos em R1 diminui ligeiramente quando o Qo global varia na faixa de 0.6 a 0.8. Diferentes mecanismos podem explicar essa observação, como múltiplos caminhos paralelos, mecanismo de *backtracking* ou armadilhas cinéticas. Entretanto, não é possível distinguir entre esses mecanismos somente a partir do perfil de energia livre em relação à Qo . As superfícies de energia dessa proteína analisada por meio do ELViM resulta em um espaço de fase efetivo 2D, como demonstrado na Figura 11C. Nesta projeção cada ponto foi colorido de acordo com seu valor de Qo global. É possível notar que as conformações desenoveladas populam a região superior esquerda da projeção (azul escuro), enquanto que a região enovelada é composta de configurações aglomeradas na parte vermelha central (região I). De acordo com o perfil de energia livre na Figura 11A, os estados de transição possuem um valor de Qo igual a 0.3 e são representados pelos pontos em azul claro. Essas estruturas de transição então dividem-se em dois caminhos distintos que ligam a área enovelada à desenovelada, pela parte inferior (região II) ou

alternativamente pela parte superior (região III) da projeção. Podemos ver que há uma grande densidade de estados na região II, sugerindo que essa é a rota de enovelamento mais favorável. Para distinguir entre essas regiões calculou-se o mapa de contato médio para as conformações selecionadas que pertencem a cada uma dessas regiões (Figuras 13 e 14).

Esses mapas de contato demonstram que a região III é composta de estruturas no qual as repetições 2 e 3 estão parcialmente enoveladas, enquanto que a região II é composta de estruturas no qual as repetições 1 e 2 estão parcialmente enoveladas. Essa análise também revela que esses estados de transição possuem quase os mesmos valores de Qo global apesar de serem bastante diferentes estruturalmente.

Em seguida, foi examinado se a região II é de fato o caminho de enovelamento cineticamente mais favorável simulando um conjunto de dinâmicas não enviesadas com o AWSEM-MD e reconstruindo a projeção 2D contendo estruturas de ambas as dinâmicas, com e sem viés (Figura 12C). Apesar de as simulações com o *Umbrella Sampling* apresentarem uma ampla exploração do espaço de fase conformacional, a amostragem não enviesada mostra claramente que há um caminho preferencial em direção ao estado nativo pela região II. Além disso, foi construída uma representação 3D desse espaço de fase efetivo com a inserção de um terceiro eixo, correspondendo ao Qo global. Na Figura 11D, as cores dos pontos correspondem à energia total de cada conformação, com as estruturas da simulação sem o *Umbrella* destacadas em preto e o estado nativo no mínimo global. Essa projeção 3D possui a forma de um funil e é exibida em três diferentes ângulos. Com ela é possível distinguir a rota principal que leva ao estado nativo, que é composta pelas estruturas em preto e as da região II agrupadas em seu meio.

É importante destacar que as conformações da região III (Fig. 11C) não foram amostradas na simulação não enviesada, o que sugere que são estados de transição de alta energia. Esse resultado é consistente com a discussão de Choe et. al. [33]. O fato de que as conformações dessas duas rotas paralelas possuem Qo global semelhante, mas podem nuclear a partir do C ou N-terminal leva ao aparecimento do *backtracking* que foi observado ao ordenar as estruturas apenas por seu Qo global, como mostrado na Figura 11B. Assim, o EIViM demonstra que utilizar apenas uma análise unidimensional nem sempre é o suficiente para distinguir entre estes dois conjuntos de rotas paralelas.

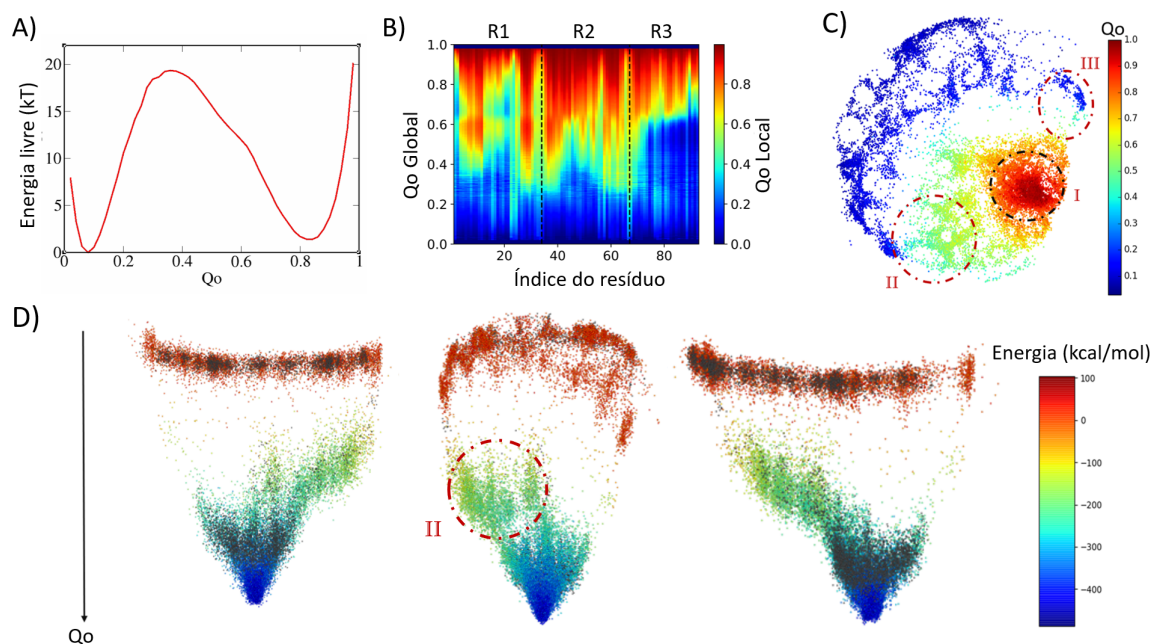


Figura 11: Análises das trajetórias enviesadas da anquirina com três repetições. A) Perfil de energia livre em função da coordenada Q_o . B) Grau de enovelamento de cada resíduo em função de Q_o com a cor indicando o Q_o médio local para um resíduo em relação ao conjunto de estruturas do Q_o global. C) Espaço de fase conformacional projetado através do ELViM em função de Q_o . D) Projeção tridimensional das trajetórias enviesadas da 3ANK em função da energia total da dinâmica com as estruturas não enviesadas destacadas em preto e o eixo Z correspondendo à coordenada Q_o .

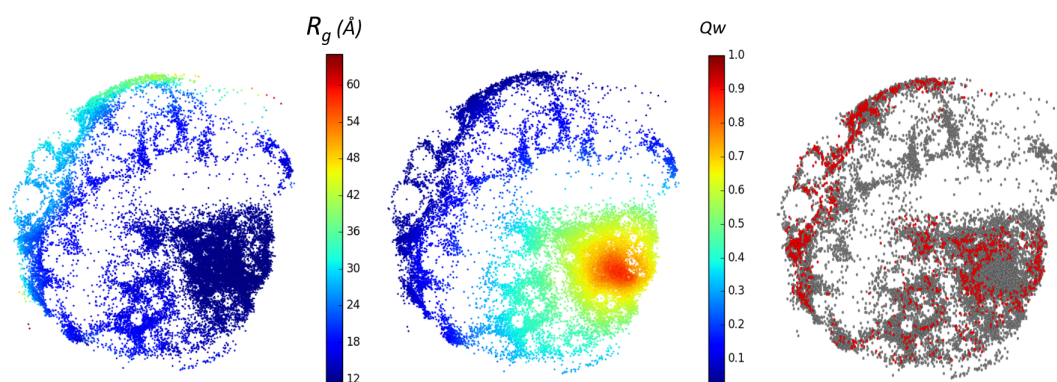


Figura 12: Projeção bidimensional do espaço de fase conformacional da 3ANK em (A) função do raio de giro e (B) em função da coordenada Q_w . (C) Projeção destacando as estruturas não enviesadas em vermelho com as conformações enviesadas em cinza.

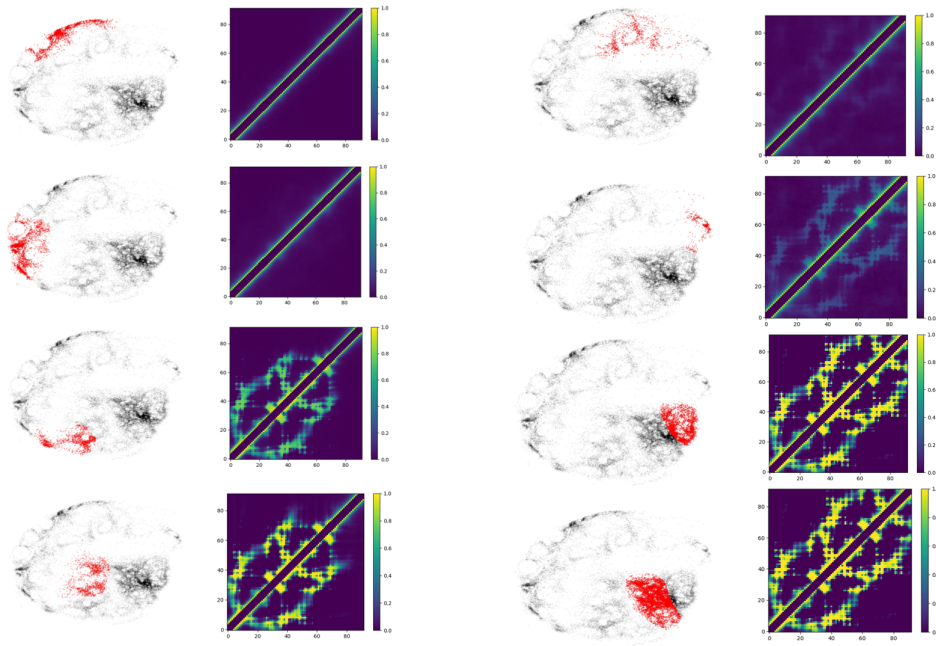


Figura 13: Mapa de contato médio para as estruturas da 3ANK que correspondem as diferentes regiões da projeção do ELViM (destacadas em vermelho).

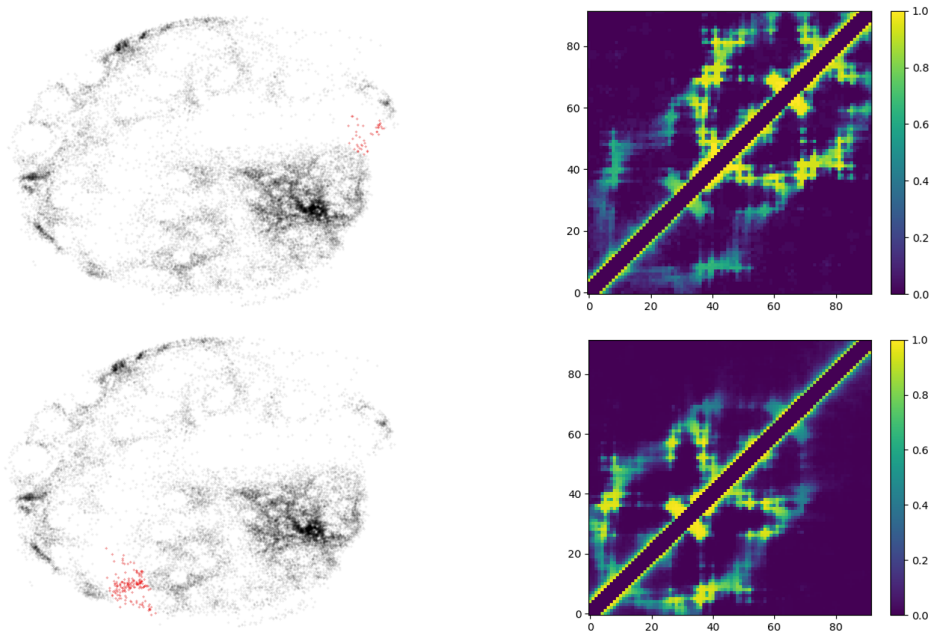


Figura 14: Mapa de contato médio para os estados de transição da 3ANK com os mesmos valores de Q_o global. A diferença entre essas conformações combinada com o fato de que possuem valores similares de Q_o é a razão por trás do aparente *backtracking* observado na Figura 11B.

4.2.2 4 repetições

Partindo novamente das dinâmicas e dos resultados obtidos por Gonzalo, vemos que a anquirina de quatro repetições apresenta um comportamento semelhante à de três, com um perfil de energia livre típico de um sistema de dois estados com uma barreira de 19.5 kT em $Qo \approx 0.4$ e com o estado enovelado em $Qo \approx 0.8$ como visto na Figura 15A. Analizando o Qo local em função do Qo global (Figura 15B) é possível ver que a nucleação também é assimétrica, iniciando próxima ao C-terminal e se propagando até o N-terminal de forma cooperativa. No estado de transição ($Qo \approx 0.3$), R3 e R4 já estão completamente enovelados. No momento em que alguns resíduos de R2 se enovelam ($Qo \approx 0.4$), R4 e alguns resíduos de R3 parecem passar por um processo de desenovelamento que será revertido novamente quando $Qo > 0.6$. Esse fenômeno é semelhante ao que foi observado com R1 na 3ANK e mais uma vez é explicado pela análise dos mapas de contatos (Figuras 17 e 18) como uma sobreposição de famílias distintas de estruturas com o mesmo Qo global. A projeção 2D do ELViM para a 4ANK (Figura 15C) se assemelha com a da 3ANK, com o Qo variando suavemente desde os estados desenovelados à esquerda até a região enovelada na direita. Uma diferença importante é que os estados intermediários (exibidos em azul claro com $Qo \approx 0.5$) não parecem conectar a região enovelada à desenovelada por um caminho denso. Ao invés disso eles formam quase que dois blocos separados: (I) um bloco que é altamente populado (acima), e (II) outro bloco que é altamente populado abaixo da região enovelada. Os mapas médios de contatos disponíveis nas Figuras 17 e 18 indicam que a região I é populada por estruturas que possuem o N-terminal parcialmente enovelado, enquanto que a região II é populada por estruturas que possuem o C-terminal parcialmente enovelado. Assim, a superfície de energia do ELViM sugere novamente dois caminhos de enovelamento paralelos e concorrentes. Como poucas estruturas parecem ligar as regiões desenoveladas às enoveladas através do caminho I, é possível que essas estruturas representem conformações de transição de alta energia. Para testar essa hipótese, foi simulado um conjunto de trajetórias não enviesadas com o AWSEM-MD e criado uma nova projeção 2D que agora contém as estruturas com e sem viés (Figura 16C). Essa nova projeção mostra que todas as trajetórias feitas sem o *Umbrella Sampling* seguem pela região II, demonstrando a existência de um caminho de enovelamento favorável.

Esse caminho pode ser melhor visto na projeção 3D da superfície de energia, exibido em três diferentes pontos de vista na Figura 15D, com a amostragem não enviesada destacada em preto. Nessa representação, o aglomerado de energia intermediária é claramente visto

quase que desconectado do mínimo da superfície (região I). Para determinar se esse *cluster* corresponde às estruturas que também são capazes de se enovelarem, foram realizadas vinte dinâmicas em cinco temperaturas diferentes, iniciadas a partir de estruturas da região I e da região II. O resultado demonstrou que as conformações da região II foram capazes de alcançar o estado nativo 76% das vezes, enquanto que as simulações que começaram a partir da região I tiveram sucesso apenas 37% das vezes. Isso indica que enquanto o enovelamento pode ocorrer de ambas as regiões, as estruturas na região II são mais propensas a alcançar o estado nativo, definindo o caminho destacado pelas trajetórias sem a tendência do *Umbrella* como sendo preferencial sobre a rota com a região I. A Figura 19 exibe a taxa de enovelamento para ambas as regiões em diferentes temperaturas.

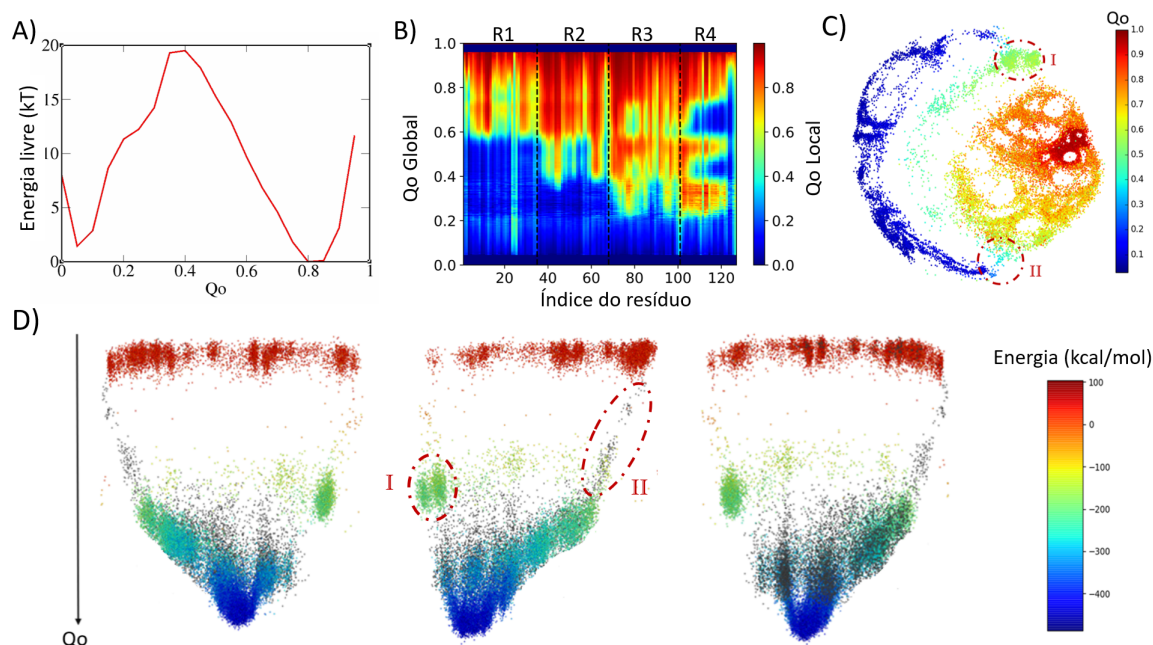


Figura 15: A) Perfil de energia livre em função da coordenada Q_o para a 4ANK. B) Grau de enovelamento de cada resíduo em função de Q_o com a cor indicando o Q_o médio local para um resíduo em relação ao conjunto de estruturas do Q_o global. C) Projeção bidimensional do espaço de fase conformacional da 4ANK em função da coordenada Q_o . D) Projeção 3D das trajetórias enviesadas em função da energia total da dinâmica com as trajetórias não enviesadas destacadas em preto.

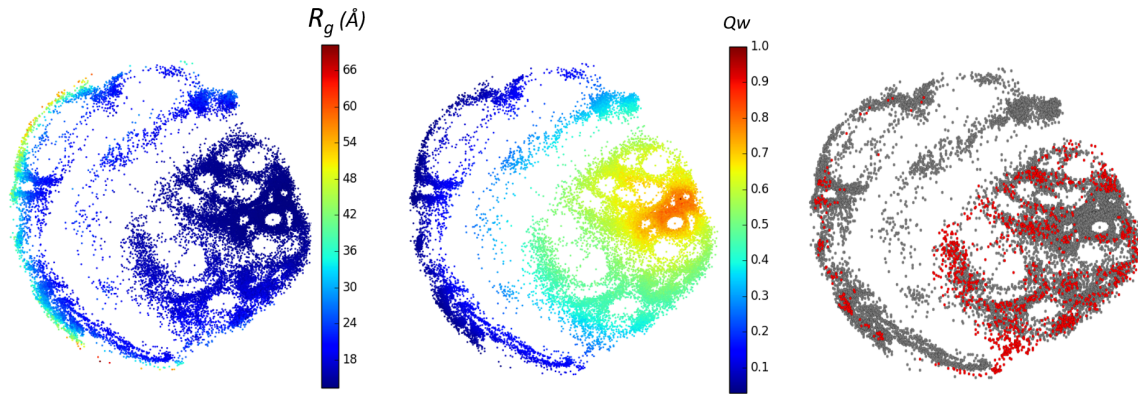


Figura 16: Projeção bidimensional do espaço de fase conformacional da 4ANK em A) função do raio de giro e B) em função da coordenada Qw . C) Projeção amostrando as estruturas obtidas com os dois tipos de simulações: as conformações não enviesadas em vermelho e as enviesadas em cinza.

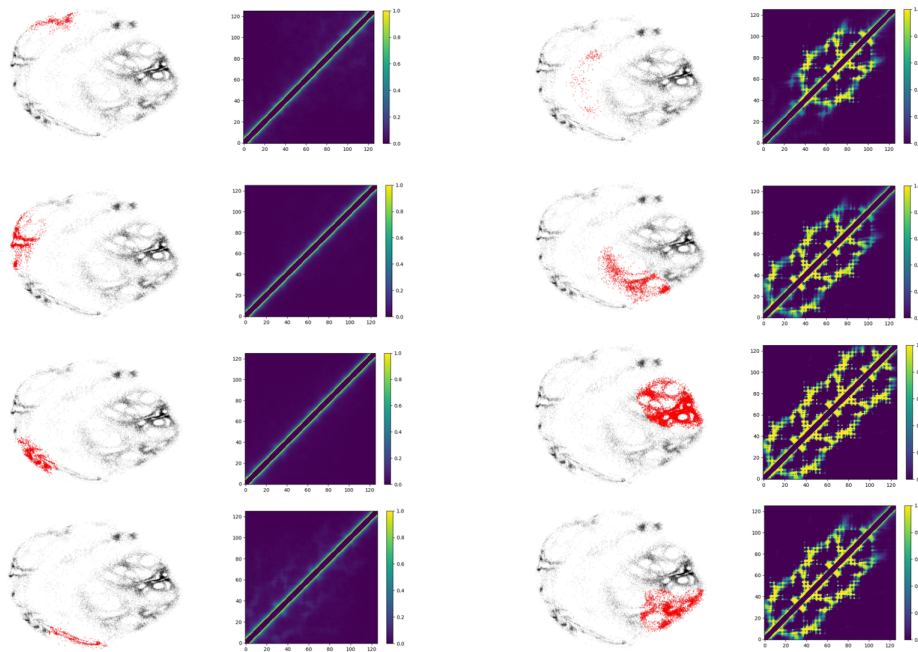


Figura 17: Mapa de contato médio das estruturas de 4ANK correspondendo às diferentes regiões da projeção do ELViM (destacadas em vermelho).

4.2.3 6 repetições

As dinâmicas da anquirina de seis repetições, que correspondem à proteína $I\kappa B\alpha$, também foram realizadas por Gonzalo. Em contraste com a 3ANK e 4ANK, que são modelos sintéticos feitos de repetições idênticas produzindo estruturas altamente simétricas,

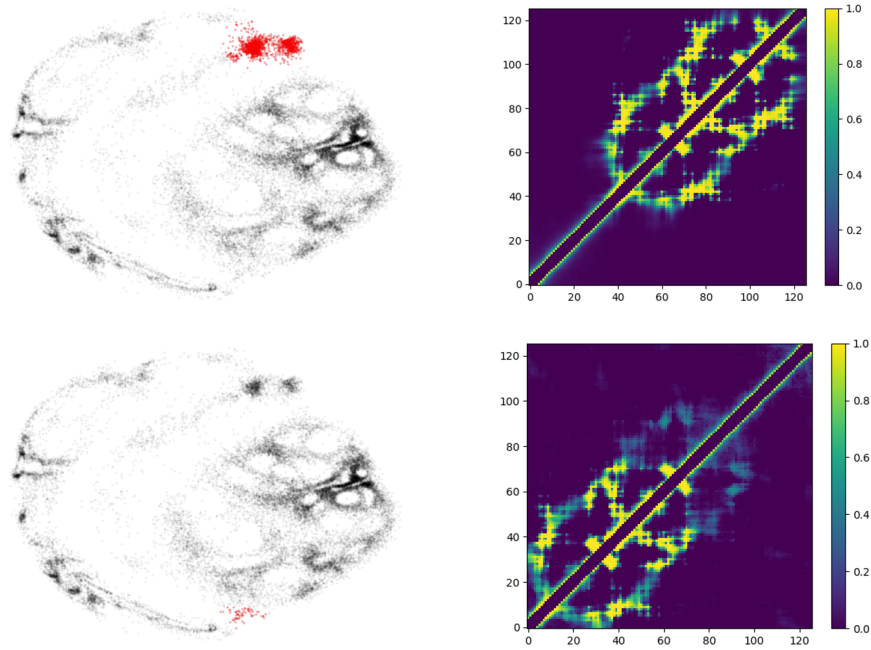


Figura 18: Mapa de contato médio dos estados de transição da 4ANK com valores semelhantes de Qo global. A diferença entre essas conformações combinada com o fato de que apresentam valores próximos de Qo é o motivo por trás do aparente *backtracking*.

a $I\kappa B\alpha$ é uma proteína que ocorre naturalmente com uma superfície de energia complexa devido à sua larga matriz de repetição e à presença de inserções e deleções dentro e entre algumas repetições. O perfil de energia livre (Figura 20A) apresenta apenas uma pequena barreira em $Qo \approx 0.6$ que separa o estado enovelado em $Qw \approx 0.8$, de um estado intermediário em $Qo \approx 0.35$, com esse estado correspondendo a duas repetições formadas (R2 e R3) e quatro não enoveladas. Mais uma vez as análises de Qo local em função de Qo global (Figura 20B) apresentam um aparente *backtracking* no enovelamento da repetição R1 entre Qo 0.6 e 0.8, que é explicado através da análise dos mapas da média de contato presentes nas Figuras 21 e 22, que demonstram que mais uma vez o aparente *backtracking* é causado por serem diferentes rotas de enovelamento com estruturas com Qo global semelhante.

A Figura 20C exibe a complexa superfície de energia que guia o enovelamento, com múltiplos caminhos com *clusters* intermediários levando ao estado nativo. É possível distinguir quatro regiões diferentes baseadas na coordenada Qo : (1) a região azul escura (entre 0 e 0.2) no fundo, correspondendo aos estados desenovelados; (2) a região azul clara (entre 0.3 e 0.4) na esquerda, correspondendo aos estados intermediários; (3) a região verde

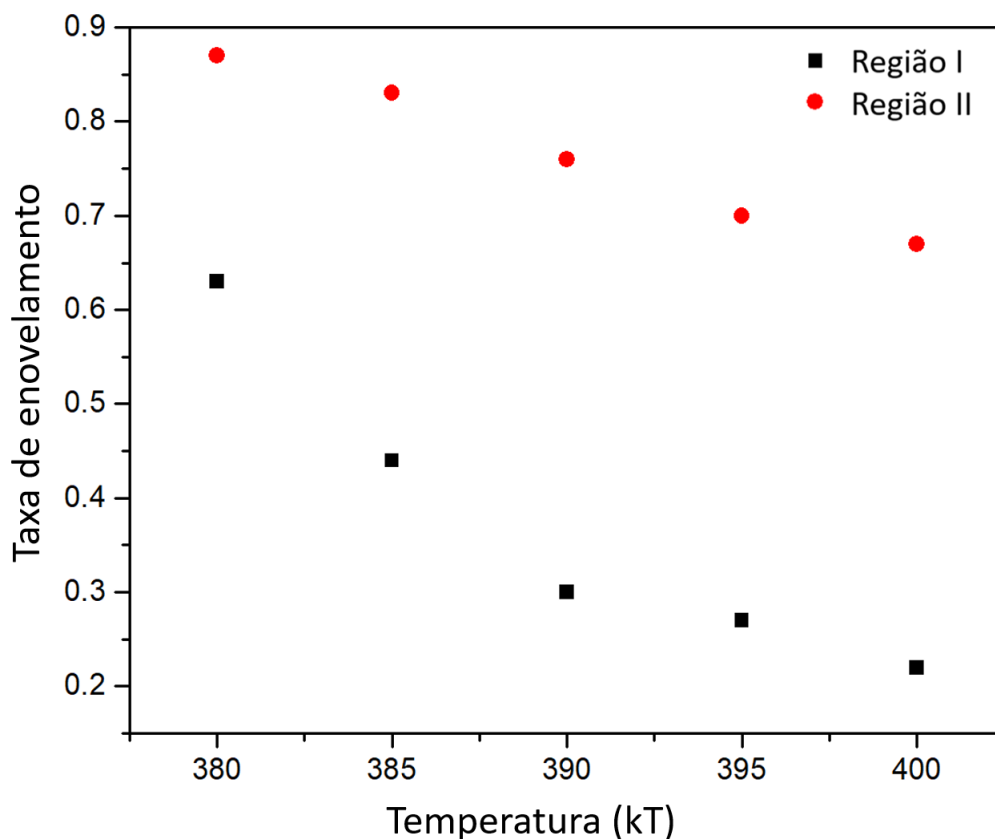


Figura 19: Taxa de enovelamento para a região I e região II da 4ANK para cada temperatura diferente.

clara (por volta de 0.5) ao meio superior, correspondendo aos estados de transição e (4) a região que vai de laranja ao vermelho (0.7 para 1.0) à direita, correspondendo do pico da barreira de energia livre à conformação totalmente enovelada, mapeando o processo pelo qual a proteína enovela. A complexa superfície de energia é melhor ilustrada pela projeção 3D na Figura 20D, na qual é possível ver que as regiões de alto Q_0 estão quase que completamente desanexadas do resto do funil, com poucos caminhos as conectando. Além disso, embora esta proteína pareça altamente simétrica, as rotas de enovelamento não são igualmente populadas, o que revela uma assimetria em seu enovelamento. É necessário destacar que a $I\kappa B\alpha$ encontra-se apenas parcialmente enovelada em seu estado livre, e se enovela completamente apenas quando interage com a $NF\kappa B$, seu parceiro ligante [34].

As projeções bidimensionais do ELViM, que permitem melhor enxergar a formação de cada repetição (para cada sistema) estão dispostas nas Figuras 23, 24 e 25 respectivamente.

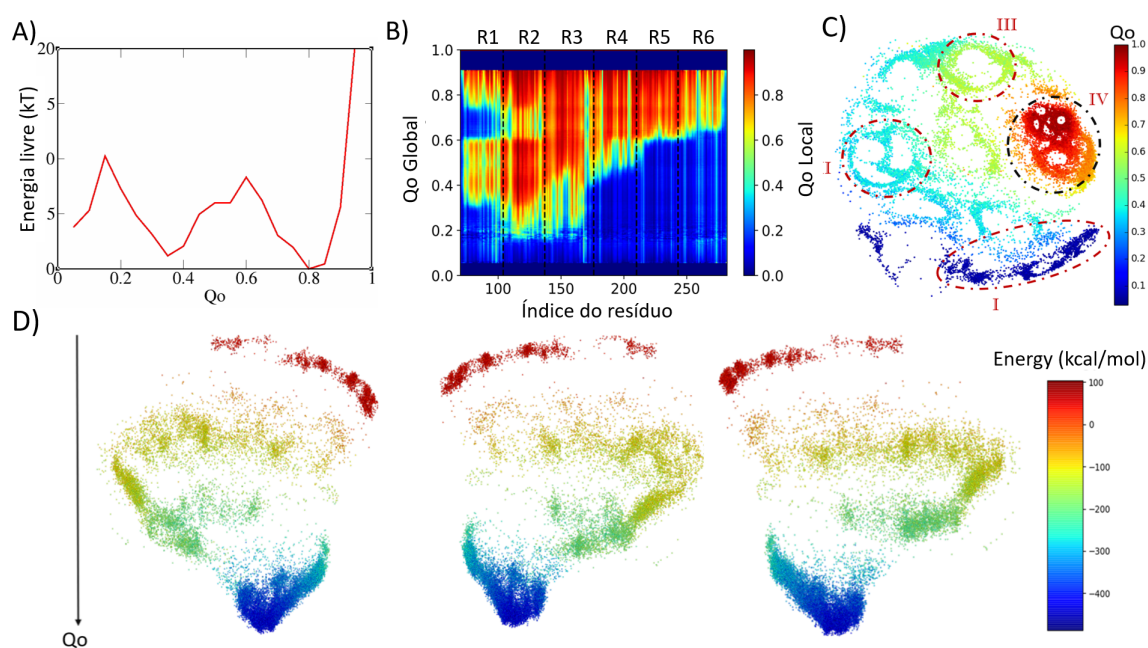


Figura 20: Análises das trajetórias da 6ANK. A) Perfil de energia livre em função da coordenada Q_o . B) Grau de enovelamento para cada resíduo em função de Q_o , com a cor indicando a média do Q_o local para um resíduo sobre o conjunto de estruturas de um dado Q_o global. C) Espaço de fase conformacional da 6ANK visualizado através do ELViM em função de Q_o . D) Projeção tridimensional das trajetórias em função da energia total da dinâmica, com o eixo Z correspondendo à coordenada Q_o .

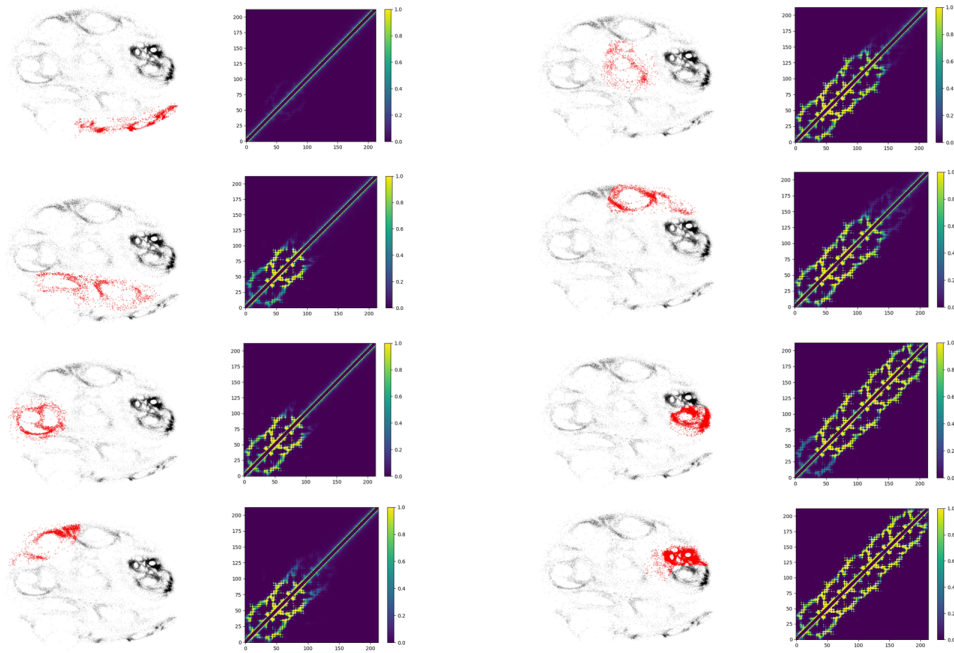


Figura 21: Mapa de contato médio das estruturas da 6ANK correspondendo às diferentes regiões da projeção do ELViM (destacadas em vermelho).

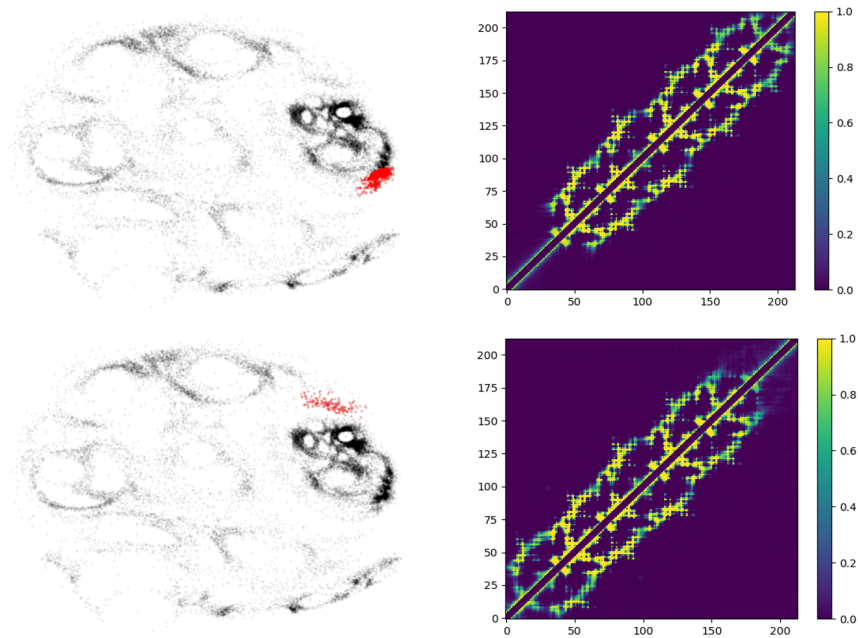


Figura 22: Mapa de contato médio dos estados de transição da 6ANK com mesmos valores de Q_o global. A diferença entre essas conformações combinada com o fato de que apresentam valores próximos de Q_o é o motivo por trás do aparente *backtracking*.

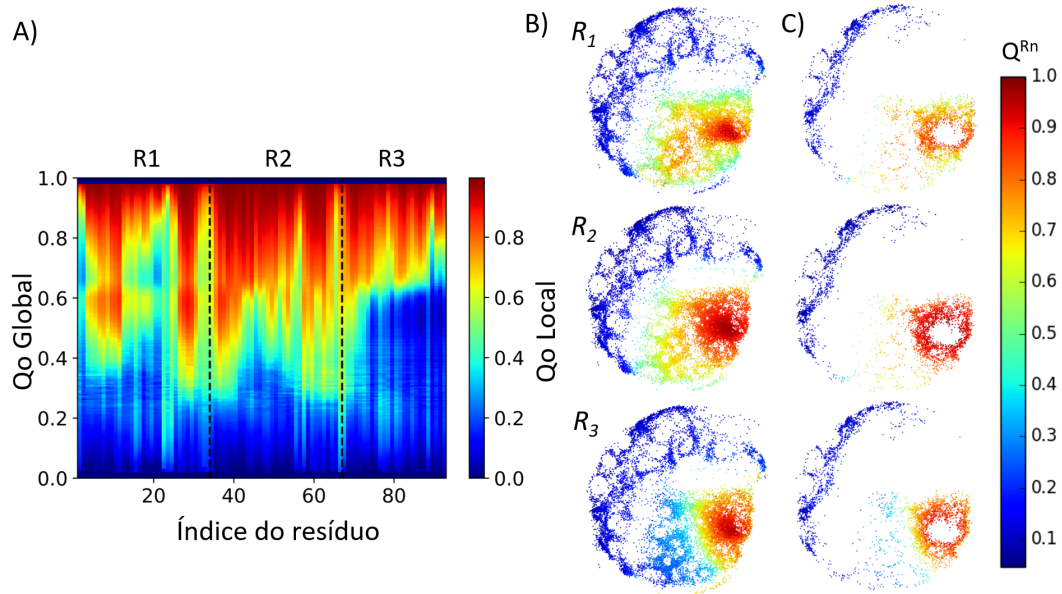


Figura 23: A) Grau de enovelamento de cada resíduo em função de Q . B) Projeções bidimensionais para as trajetórias enviesadas da 3ANK em função de Q^{Rn} para cada repetição. C) Projeções bidimensionais para as trajetórias não enviesadas da 3ANK em função de Q^{Rn} para cada repetição.

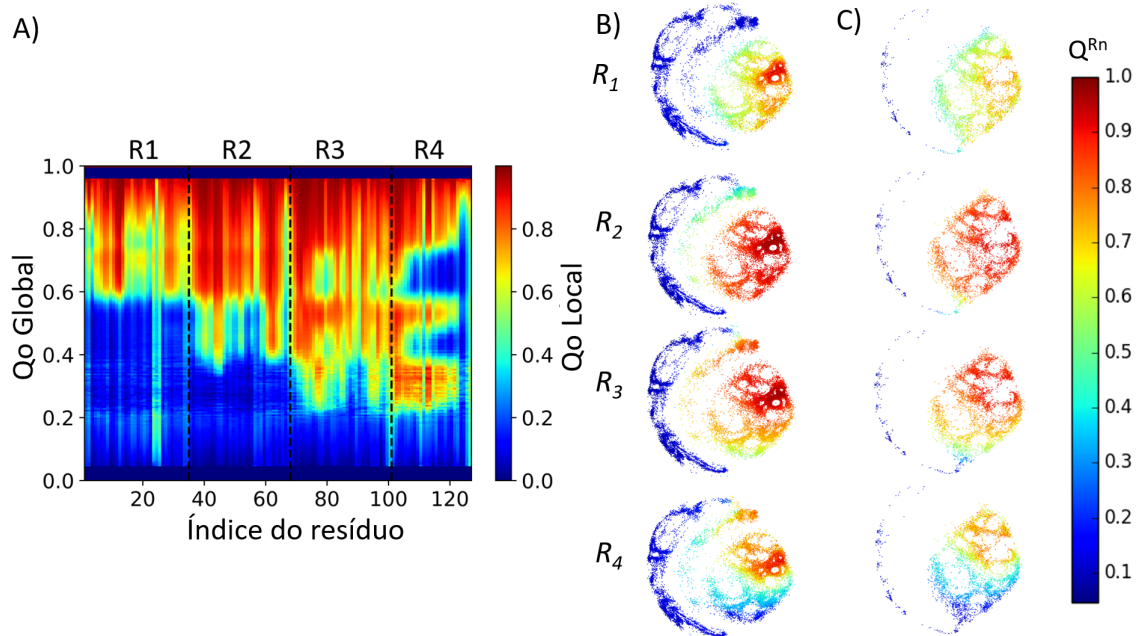


Figura 24: A) Grau de enovelamento de cada resíduo em função de Qo . B) Projeções bidimensionais para as trajetórias enviesadas da 4ANK em função de Q^{Rn} para cada repetição. C) Projeções bidimensionais para as trajetórias não enviesadas da 4ANK em função de Q^{Rn} para cada repetição.

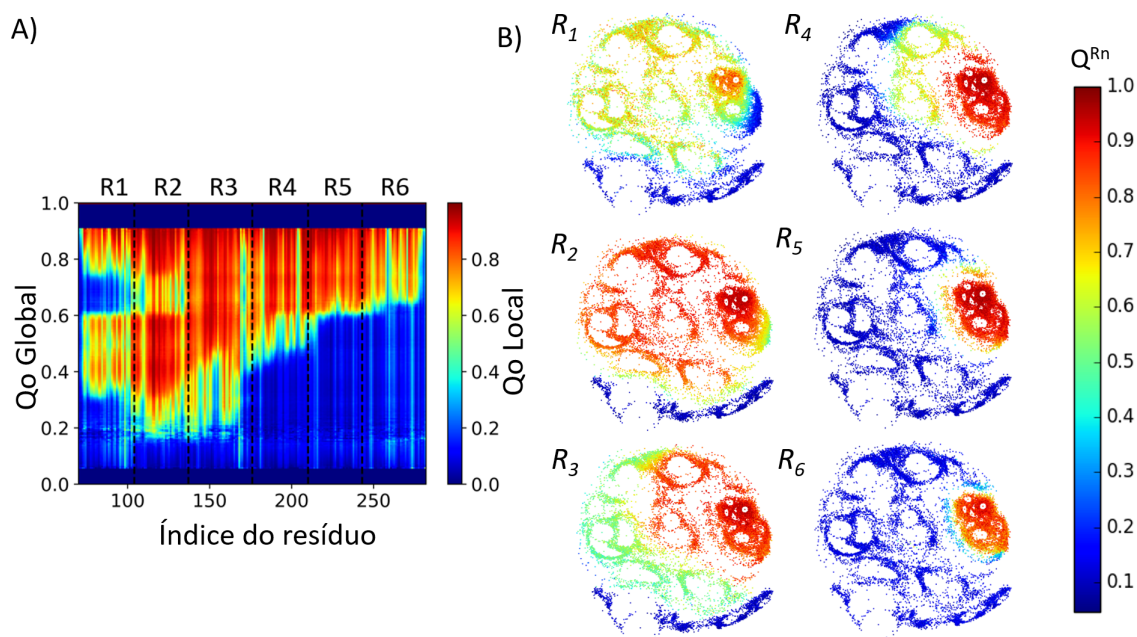


Figura 25: A) Grau de enovelamento de cada resíduo em função de Qo . B) Projeções bidimensionais para as trajetórias enviesadas da 6ANK em função de Q^{Rn} para cada repetição.

5 Conclusões

As superfícies de energia foram analisadas através do método ELViM para dois sistemas diferentes: A IDP β -amilóide e as repetições de anquirina, especificamente de três, quatro e seis repetições.

A natureza desordenada e a alta flutuação das IDPs resulta em uma superfície complexa de alta dimensionalidade, além do fato de que a falta de estrutura nativa também implica na falta de uma coordenada de reação adequada. Por meio do ELViM foi possível explorar as mudanças conformacionais nas formas monoméricas das $A\beta$, o que revelou que uma pré disposição para agregação possivelmente possui uma relação direta com o conteúdo de folhas- β dos monômeros, uma vez que a $A\beta_{42}$ e os mutantes, que possuem altas taxas de agregação, apresentam em suas rotas uma barreira de energia menor que a presente na $A\beta_{40}$. Além disso, a localização dos estados fundamentais nas projeções revelam que a maioria dos contatos necessários para iniciar a formação de fibras já estão formados na $A\beta_{42}$ e nas mutações, enquanto que a $A\beta_{40}$ precisa desfazer suas hélices primeiro.

Já os resultados para as proteínas de repetição confirmaram as expectativas de que suas superfícies são de fato afuniladas em direção ao estado nativo. Contudo, a

ocorrência de estados intermediários e a localização dos estados de transição são governadas pela topologia mais detalhada de cada proteína. Vemos também que, mesmo quando a topologia parece ser altamente simétrica, a população de rotas de enovelamento acaba sendo fortemente assimétrica. Todos os três sistemas parecem retroceder a formação de alguns elementos quando monitorados utilizando somente a média do Qo local, efeito que pode ser entendido como decorrente de serem estruturas diferentes de rotas de enovelamento paralelas que possuem o mesmo Qo global. É provável que à medida que essas matrizes de repetição crescem, a cooperatividade entre as repetições decai, conforme a energia das interfaces de tamanho finito deve ficar cada vez mais forte para superar a entropia de incorporar uma nova repetição [35]. Isso é observado para o maior sistema que foi analisado nesse trabalho, a $I\kappa B\alpha$, que tem uma região bastante plana da superfície que consiste em um subdomínio enovelado de cerca de 3 unidades de repetição consecutivas, o que é consistente com o conjunto estrutural proposto para esta proteína no estado livre não ligado e que está fortemente relacionado ao seu mecanismo funcional [36].

Através do *Energy Landscape Visualization Method* foi possível analisar as superfícies de energia de sistemas que até então apresentavam empecilhos em seu entendimento devido às suas naturezas. Os resultados obtidos confirmaram algumas hipóteses já existentes e revelaram novos mecanismos pelos quais esses sistemas atuam.

Referências

- [1] Vinícius de Godoi Contessoto, Antonio Bento de Oliveira Junior, Jorge Chahine, Ronaldo Junio de Oliveira, and Vitor Barbanti Pereira Leite. Introdução ao problema de enovelamento de proteínas: uma abordagem utilizando modelos computacionais simplificados. *Revista Brasileira de Ensino de Física*, 40(4):e4307, 2018.
- [2] John Hardy and Dennis J Selkoe. The amyloid hypothesis of alzheimer’s disease: progress and problems on the road to therapeutics. *science*, 297(5580):353–356, 2002.
- [3] P. E. Leopold, M. Montal, and J. N. Onuchic. Protein folding funnels: a kinetic approach to the sequence-structure relationship. 89(18):8721–8725.
- [4] José Nelson Onuchic, Nicholas D Socci, Zaida Luthey-Schulten, and Peter G Wolynes. Protein folding funnels: the nature of the transition state ensemble. *Folding and Design*, 1(6):441–450, 1996.
- [5] José Nelson Onuchic, Zaida Luthey-Schulten, and Peter G Wolynes. Theory of protein folding: the energy landscape perspective. *Annual review of physical chemistry*, 48(1):545–600, 1997.
- [6] Cheick Ahmed Tidiane Sissoko. *Profiling tau hyperphosphorylation in formalin-fixed paraffin embedded human brain with liquid chromatography tandem mass spectrometry*. PhD thesis, Icahn School of Medicine at Mount Sinai, 2020.
- [7] Weihua Zheng, Min-Yeh Tsai, Mingchen Chen, and Peter G Wolynes. Exploring the aggregation free energy landscape of the amyloid- β protein (1–40). *Proceedings of the National Academy of Sciences*, 113(42):11835–11840, 2016.
- [8] Debayan Chakraborty, John E Straub, and Dave Thirumalai. Differences in the free energies between the excited states of $\alpha\beta 40$ and $\alpha\beta 42$ monomers encode their aggregation propensities. *Proceedings of the National Academy of Sciences*, 117(33):19926–19937, 2020.
- [9] Weihua Zheng, Min-Yeh Tsai, and Peter G Wolynes. Comparing the aggregation free energy landscapes of amyloid beta (1–42) and amyloid beta (1–40). *Journal of the American Chemical Society*, 139(46):16666–16676, 2017.
- [10] Bogdan Barz, Qinghua Liao, and Birgit Strodel. Pathways of amyloid- β aggregation depend on oligomer shape. *Journal of the American Chemical Society*, 140(1):319–327, 2018.
- [11] Steve Plimpton. Fast parallel algorithms for short-range molecular dynamics. *Journal of computational physics*, 117(1):1–19, 1995.
- [12] Simona Tomaselli, Veronica Esposito, Paolo Vangone, Nico AJ van Nuland, Alexandre MJJ Bonvin, Remo Guerrini, Teodorico Tancredi, Piero A Temussi, and Delia Picone. The α -to- β conformational transition of alzheimer’s $\alpha\beta$ -(1–42) peptide in aqueous media is reversible: a step by step conformational analysis suggests the location of β conformation seeding. *ChemBioChem*, 7(2):257–267, 2006.

-
- [13] Subramanian Vivekanandan, Jeffrey R Brender, Shirley Y Lee, and Ayyalusamy Ramamoorthy. A partially folded structure of amyloid-beta (1–40) in an aqueous environment. *Biochemical and biophysical research communications*, 411(2):312–316, 2011.
- [14] Aida Attar, Derya Meral, Brigita Urbanc, and Gal Bitan. Assembly of amyloid β -protein variants containing familial alzheimer’s disease-linked amino acid substitutions. In *Bio-nanoimaging*, pages 429–442. Elsevier, 2014.
- [15] Asa Hatami, Sanaz Monjazez, Saskia Milton, and Charles G Glabe. Familial alzheimer’s disease mutations within the amyloid precursor protein alter the aggregation and conformation of the amyloid- β peptide. *Journal of Biological Chemistry*, 292(8):3172–3185, 2017.
- [16] Antonio B Oliveira Junior, Xingcheng Lin, Prakash Kulkarni, José N Onuchic, Susmita Roy, and Vitor BP Leite. Exploring energy landscapes of intrinsically disordered proteins: Insights into functional mechanisms. *Journal of Chemical Theory and Computation*, 2021.
- [17] Lisanna Paladin, Martina Bevilacqua, Sara Errigo, Damiano Piovesan, Ivan Mičetić, Marco Necci, Alexander Miguel Monzon, Maria Laura Fabre, Jose Luis Lopez, Juliet F Nilsson, Javier Rios, Pablo Lorenzano Menna, Maia Cabrera, Martin Gonzalez Buitron, Mariane Gonçalves Kulik, Sebastian Fernandez-Alberti, Maria Silvina Fornasari, Gustavo Parisi, Antonio Lagares, Layla Hirsh, Miguel A Andrade-Navarro, Andrey V Kajava, and Silvio C E Tosatto. RepeatsDB in 2021: improved data and extended classification for protein tandem repeat structures. *Nucleic Acids Research*, 49(D1):D452–D457, 11 2020.
- [18] R Gonzalo Parra, Rocío Espada, Nina Verstraete, and Diego U Ferreiro. Structural and energetic characterization of the ankyrin repeat protein family. *PLoS computational biology*, 11(12):e1004659, 2015.
- [19] R Gonzalo Parra, Rocío Espada, Ignacio E Sánchez, Manfred J Sippl, and Diego U Ferreiro. Detecting repetitions and periodicities in proteins by tiling the structural space. *The Journal of Physical Chemistry B*, 117(42):12887–12897, 2013.
- [20] Rocío Espada, R. Gonzalo Parra, Manfred J. Sippl, Thierry Mora, Aleksandra M. Walczak, and Diego U. Ferreiro. Repeat proteins challenge the concept of structural domains. *Biochemical Society Transactions*, 43:844–849, 2015.
- [21] Diego U. Ferreiro, Aleksandra M. Walczak, Elizabeth A. Komives, and Peter G. Wolynes. The energy landscapes of repeat-containing proteins: Topology, cooperativity, and the folding funnels of one-dimensional architectures. *PLOS Computational Biology*, 4(5):1–13, 05 2008.
- [22] Antonio B. Oliveira, Huan Yang, Paul C. Whitford, and Vitor B. P. Leite. Distinguishing biomolecular pathways and metastable states. *Journal of Chemical Theory and Computation*, 15(11):6482–6490, 2019. PMID: 31618581.

- [23] Antonio B Oliveira Jr, Huan Yang, Paul C Whitford, and Vitor BP Leite. Distinguishing biomolecular pathways and metastable states. *Journal of chemical theory and computation*, 15(11):6482–6490, 2019.
- [24] Corey Hardin, Michael P Eastwood, Zaida Luthey-Schulten, and Peter G Wolynes. Associative memory hamiltonians for structure prediction without homology: alpha-helical proteins. *Proceedings of the National Academy of Sciences*, 97(26):14235–14240, 2000.
- [25] 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, 2003.
- [26] Alan M. Ferrenberg and Robert H. Swendsen. New monte carlo technique for studying phase transitions. 61(23):2635.
- [27] Shankar Kumar, John M Rosenberg, Djamal Bouzida, Robert H Swendsen, and Peter A Kollman. THE weighted histogram analysis method for free energy calculations on biomolecules. i. the method. 13(8):1011–1021.
- [28] Aram Davtyan, Nicholas P Schafer, Weihua Zheng, Cecilia Clementi, Peter G Wolynes, and Garegin A Papoian. Awsem-md: protein structure prediction using coarse-grained physical potentials and bioinformatically based local structure biasing. *The Journal of Physical Chemistry B*, 116(29):8494–8503, 2012.
- [29] Michael P Eastwood and Peter G Wolynes. Role of explicitly cooperative interactions in protein folding funnels: a simulation study. *The Journal of Chemical Physics*, 114(10):4702–4716, 2001.
- [30] Robert B. Best, Gerhard Hummer, and William A. Eaton. Native contacts determine protein folding mechanisms in atomistic simulations. *Proceedings of the National Academy of Sciences*, 110(44):17874–17879, 2013.
- [31] Mona Habibi, Jörg Rottler, and Steven S Plotkin. As simple as possible, but not simpler: exploring the fidelity of coarse-grained protein models for simulated force spectroscopy. *PLoS computational biology*, 12(11):e1005211, 2016.
- [32] Sébastien Côté, Philippe Derreumaux, and Normand Mousseau. Distinct morphologies for amyloid beta protein monomer: A β 1–40, a β 1–42, and a β 1–40 (d23n). *Journal of chemical theory and computation*, 7(8):2584–2592, 2011.
- [33] 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*, 103(3):586–591, 2006.
- [34] Diego U Ferreiro and Elizabeth A Komives. Molecular mechanisms of system control of nf- κ b signaling by κ b α . *Biochemistry*, 49(8):1560–1567, 2010.
- [35] Ezequiel A Galpern, María I Freiburger, and Diego U Ferreiro. Large ankyrin repeat proteins are formed with similar and energetically favorable units. *Plos one*, 15(6):e0233865, 2020.

- [36] Stephanie ME Truhlar, Erika Mathes, Carla F Cervantes, Gourisankar Ghosh, and Elizabeth A Komives. Pre-folding $\text{i}\kappa\text{b}\alpha$ alters control of $\text{nf-}\kappa\text{b}$ signaling. *Journal of molecular biology*, 380(1):67–82, 2008.

A Descrição da representação *coarse-grained*

Na representação estrutural do AWSEM, a posição e orientação de cada aminoácido é descrita pela posição de seus átomos de C_α , C_β e O (com exceção da glicina, que não possui o átomo de C_β). A posição dos demais átomos da cadeia é calculada assumindo uma geometria ideal:

$$r_{Ni} = 0.48318r_{C_{\alpha i-1}} + 0.70328r_{C_{\alpha i}} - 0.18643r_{O_{i-1}}$$

$$r_{C'_i} = 0.44365r_{C_{\alpha i}} + 0.23520r_{C_{\alpha i+1}} + 0.32115r_{O_i} \quad (14)$$

$$r_{H_i} = 0.84100r_{C_{\alpha i-1}} + 0.89296r_{C_{\alpha i}} - 0.73389r_{O_{i-1}}$$

onde r_{H_i} é a posição do hidrogênio que está ligado ao nitrogênio da cadeia.

B Campo de força do AWSEM

A função de energia total é descrita da seguinte forma:

$$V_{total} = V_{backbone} + V_{contact} + V_{burial} + V_{HB} \quad (15)$$

$V_{backbone}$ é o termo responsável por manter a geometria da cadeia principal. A forma completa desse potencial de cadeia é expressa por:

$$V_{backbone} = V_{con} + V_{chain} + V_\chi + V_{rama} + V_{excl} \quad (16)$$

no qual V_{con} garante a conectividade da cadeia por meio de uma série de ligações harmônicas:

$$\begin{aligned} V_{con} = \lambda_{con} \sum_{i=1}^N & \left[(r_{C_{\alpha i} O_i} - r_{C_{\alpha i} O_i}^0)^2 + (r_{C_{\alpha i} C_{\beta i i}} - r_{C_{\alpha i} C_{\beta i}}^0)^2 \right] \\ & + \lambda_{con} \sum_{i=1}^{N-1} \left[(r_{C_{\alpha i} C_{\alpha i}} - r_{C_{\alpha i} C_{\alpha i}}^0)^2 + (r_{O_i C_{\alpha i}} - r_{O_i C_{\alpha i}}^0)^2 \right] \end{aligned} \quad (17)$$

onde os valores de λ_{con} e as distâncias de equilíbrio r^0 são tabelados a partir de dados experimentais, disponíveis na documentação original [28].

Os ângulos de ligação corretos em torno dos átomos de C_α são obtidos através do potencial V_{chain} , que também utiliza de potenciais harmônicos, como mostrado a seguir:

$$V_{chain} = \lambda_{chain} \left[\sum_{i=2}^N (r_{N_i C_{\beta i}} - r_{N_i C_{\beta i}}^0)^2 + \sum_{i=1}^{N-1} (r_{C'_i C_{\beta i}} - r_{C'_i C_{\beta i}}^0)^2 + \sum_{i=2}^{N-1} (r_{N_i C'_i} - r_{N_i C'_i}^0)^2 \right] \quad (18)$$

V_χ é o termo responsável pela quiralidade do C_α , expresso como:

$$V_\chi = \lambda_\chi \sum_{i=2}^{N-1} (\chi_i - \chi_0)^2 \quad (19)$$

onde

$$\chi_i = (\mathbf{r}_{C'_i C_{\alpha i}} \times \mathbf{r}_{C_{\alpha i} N_i}) \mathbf{r}_{C_{\alpha i} C_{\beta i}} \quad (20)$$

Para reproduzir a distribuição experimental dos ângulos diedrais da cadeia é utilizado o potencial de Ramachandran, V_{rama} .

$$V_{rama} = -\lambda_{rama} \sum_{i=2}^{N-1} \sum_j W_j \exp -\sigma_j (\omega_{\phi_j} \cos(\phi_i - \phi_{0_j}) - 1)^2 + \omega_{\psi_j} (\cos(\psi_i - \psi_{0_j}) - 1)^2 \quad (21)$$

O primeiro e o último resíduo, assim como o resíduo de glicina, não estão incluídos nesse potencial. ϕ_i é o ângulo diedral entre os átomos de C'_{i-1} , N_i , C_α e C' , e ψ_i é o ângulo diedral entre os átomos de N_i , C_α , C' e N_{i+1} .

V_{excl} é a interação de volume excluído que fornece uma repulsão entre os átomos próximos, impedindo-os de se sobreporem. Seu potencial é definido como:

$$V_{excl} = \lambda_{excl} \sum_{ij} \Theta(r_{Ci,Cj} - r_{ex}^C) (r_{Ci,Cj} - r_{ex}^C)^2 + \lambda_{excl} \sum_{ij} \Theta(r_{Oi,Oj} - r_{ex}^O) (r_{Oi,Oj} - r_{ex}^O)^2 \quad (22)$$

onde $r_{ex}^C = 3.5\text{\AA}$ para separação de sequência menor que 5 e $4.5\text{\AA}$ caso contrário, enquanto que $r_{ex}^O = 3.5\text{\AA}$ para qualquer separação de sequência. O subscrito C refere-se a ambos C_α e C_β . Os índices i e j são os índices dos átomos, que variam sobre todos os pares de C e O que não estão diretamente conectados por V_{con} . Por fim, Θ é a função de Heaviside.

Esses são os potenciais que compõem o potencial de cadeia $V_{backbone}$.

O próximo termo é o potencial de contato, $V_{contact}$ e para descrevê-lo é útil definir dois intervalos de distâncias entre C_β - C_β , que daqui em diante serão identificados pelos sobrescritos I e II . O primeiro intervalo de distância vai de $r_{min}^I = 4.5\text{\AA}$ à $r_{max}^I = 6.5\text{\AA}$. O segundo intervalo vai de $r_{min}^I = 6.5\text{\AA}$ à $r_{max}^I = 9.5\text{\AA}$. Se os átomos de C_β de dois resíduos i e j estão separados por uma distância entre r_{min}^μ e r_{max}^μ , então a função Θ_{ij}^μ , dada na Equação 23, será igual a 1; caso contrário será 0. Ela muda suavemente de 1 para 0 próximo aos extremos dos intervalos de distância.

$$\Theta_{ij}^\mu = \frac{1}{4}(1 + \tanh[\eta(r_{ij} - r_{min}^\mu)])(1 + \tanh[\eta(r_{ij} - r_{max}^\mu)]) \quad (23)$$

Ao somar Θ_{ij}^μ sobre todo o j é possível obter o número de resíduos no poço- μ dos resíduos i . A densidade local, ρ_i , dos resíduos i é definida como $\rho_i = \sum_{j=1}^N \Theta_{ij}^I$.

Definindo Θ_{ij}^μ é possível então descrever $V_{contact}$, que é o termo de interação de contato entre resíduos distantes na sequência. Esse termo é composto por dois outros potenciais: V_{direct} e V_{water} . V_{direct} é um potencial aditivo par a par:

$$V_{direct} = -\lambda_{direct} \sum_{j-i>9}^N \gamma_{ij}(a_i, a_j) \Theta_{ij}^I \quad (24)$$

onde r_{ij} é a distância C_β - C_β entre os resíduos i e j , e $\gamma_{ij}(a_i, a_j)$ é uma constante específica do tipo de resíduo.

V_{water} é um termo de interação de vários corpos que alterna entre pesos de interação mediados por água e mediados por proteínas dependendo da densidade local em torno dos resíduos de interação.

$$V_{water} = -\lambda_{water} \sum_{j-i>9}^N \Theta_{ij}^{II} (\sigma_{ij}^{wat} \gamma_{ij}^{wat}(a_i, a_j) + \sigma_{ij}^{prot} \gamma_{ij}^{prot}(a_i, a_j)) \quad (25)$$

onde σ_{ij}^{wat} e σ_{ij}^{prot} são as funções de comutação:

$$\sigma_{ij}^{wat} = \frac{1}{4}(1 + \tanh[\eta_\sigma(\rho_i - \rho_0)])(1 + \tanh[\eta_\sigma(\rho_i - \rho_0)])$$

$$\sigma_{ij}^{prot} = 1 - \sigma_{ij}^{wat} \quad (26)$$

no qual σ_{ij}^{wat} e σ_{ij}^{prot} variam suavemente de 0 a 1 e de 1 a 0, respectivamente, como nenhuma das densidades locais, ρ_i e ρ_j excedem o limite $\rho_0=2.6$.

O termo de enterramento V_{burial} é outra interação de vários corpos que é baseada em uma propensão particular do tipo de resíduo para estar em um ambiente de baixa, média ou alta densidade.

$$V_{burial} = -\frac{1}{2} \lambda_{burial} \sum_{i=1}^N \sum_{\mu=1}^3 \gamma_{burial}(a_i, \rho_i) (\tanh[\eta(\rho_i - \rho_{min}^\mu)] + \tanh[\eta(\rho_{max}^\mu - \rho_i)]) \quad (27)$$

O potencial de ligação de hidrogênio V_{HB} é composto pela soma de outros três termos:

$$V_{HB} = V_\beta + V_{P-AP} + V_{helical} \quad (28)$$

Os primeiros dois termos são termos de ligação de hidrogênio β . O potencial V_β é composto de termos que corrigem e determinam a geometria da ligação de hidrogênio β , enquanto que o potencial V_{P-AP} permite que uma cadeia de proteína adote conformações próximas as de folhas- β paralelas ou antiparalela antes que as ligações de hidrogênio sejam totalmente formadas. A força deste potencial é escolhida para que as estruturas possam facilmente se desmontar e se recompor. Ambos os potenciais são construído a partir do somatório valores experimentais tabelados, e se encontram no material suplementar da documentação do AWSEM [28].

Por fim, o termo $V_{helical}$ é o potencial responsável pela formação de hélices- α .

$$V_{helical} = -\lambda_{helical} \sum_{i=1}^{N-4} (f(a_i) + f(a_{i+4})) (\gamma_{prot} \sigma_{i,i+4}^{prot} + \gamma_{wat} \sigma_{i,i+4}^{wat}) \times \exp \left[-\frac{(r_{i,i+4}^{ON} - \langle r^{ON} \rangle)^2}{2\sigma_{ON}^2} - \frac{(r_{i,i+4}^{OH} - \langle r^{OH} \rangle)^2}{2\sigma_{OH}^2} \right] \quad (29)$$

onde $f(a_i)$ é a probabilidade de encontrar o resíduo i em uma hélice. Todos os tipos de aminoácidos tem valores positivos entre 0 e 1 com exceção da prolina, uma vez que não possui um hidrogênio de cadeia de amida e portanto só pode ser um acceptor de ligação de hidrogênio, nunca um doador. Para refletir esse comportamento é utilizado $f(a_{i+4})=3.0$ se o resíduo $i+4$ for uma prolina. $\sigma_{i,j}^{prot}$ e $\sigma_{i,j}^{wat}$ são os mesmo que os da Equação B. γ_{prot} é a constante da força de interação quando ambos os resíduos se encontram na parte interna da proteína (enterrados). Quando os resíduos estão expostos à água, é permitido que eles formem ligações de hidrogênio com as moléculas de água vizinhas, de forma que formas essas ligações entre eles mesmo não é tão favorável.

Todos os valores das constantes presentes nos potenciais estão tabeladas e disponíveis no material suplementar da documentação original [28].