Universidade de São Paulo
FFCLRP - Departamento de Física
Pós-graduação em Física aplicada à Medicina e Biologia

MATHEUS RODRIGUES DE MENDONÇA

Análise de modos normais dos movimentos conformacionais em proteínas

Tese apresentada à Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto da Universidade de São Paulo, como parte das exigências para a obtenção do título de Doutor em Ciências, Área: Física aplicada à Medicina e Biologia

Ribeirão Preto - SP
2015

MATHEUS RODRIGUES DE MENDONÇA

Análise de modos normais dos movimentos conformacionais em proteínas

Tese apresentada à Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto da Universidade de São Paulo, como parte das exigências para a obtenção do título de Doutor em Ciências.

Área de Concentração:
Física aplicada à Medicina e Biologia

Orientador:
Nelson Augusto Alves

Versão original
Disponível na FFCLRP - USP

Ribeirão Preto - SP
2015

FICHA CATALOGRÁFICA

Rodrigues de Mendonça, Matheus

Análise de modos normais dos movimentos conformacionais em proteínas / Matheus Rodrigues de Mendonça; orientador Nelson Augusto Alves. Ribeirão Preto - SP, 2015.

86 p.:il.

Tese (Doutorado - Programa de Pós-graduação em Física aplicada à Medicina e Biologia) - Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto da Universidade de São Paulo, 2015.

1. modelo de rede elástica. 2. fator-B. 3. dinâmica vibracional. 4. análise de modos normais.

Nome: Rodrigues de Mendonça, Matheus

Título: Análise de modos normais dos movimentos conformacionais em proteínas

Tese apresentada à Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto da Universidade de São Paulo, como parte das exigências para a obtenção do título de Doutor em Ciências.

Aprovado em:____/____/____.

Banca Examinadora

Prof(a). Dr(a). :_____ Instituição:_____

Julgamento: _____ Assinatura:_____

Prof(a). Dr(a). :_____ Instituição:_____

Julgamento: _____ Assinatura:_____

Prof(a). Dr(a). :_____ Instituição:_____

Julgamento: _____ Assinatura:_____

Prof(a). Dr(a). :_____ Instituição:_____

Julgamento: _____ Assinatura:_____

Prof(a). Dr(a). :_____ Instituição:_____

Julgamento: _____ Assinatura:_____

Dedico esta Tese aos meus pais Sônia Maria Rodrigues de Mendonça e Mailton Ribeiro de Mendonça.

Agradecimentos

Primeiramente, agradeço a Deus pela oportunidade de cursar o doutorado no programa de pós-graduação FAMB, um projeto de vida que nasceu em um dado período do curso de graduação em Licenciatura em Física pela Faculdade de Engenharia (FEIS) da Universidade Estadual Paulista (Unesp) do Campus de Ilha Solteira.

Aos meus amados pais, Sônia Maria Rodrigues de Mendonça e Mailton Ribeiro de Mendonça, pela minha educação, pelo exemplo de humildade, simplicidade e sabedoria, e pelo ambiente de paz e amor que eles me proporcionaram ao longo desses anos. Ao meu irmão Adriano Rodrigues de Mendonça pelo incentivo e sobrinho Lucas de Souza Rodrigues (hoje com 6 anos) pelo companherismo e momentos de alegria.

Ao Professor Dr. Nelson Augusto Alves, pela sua excelente orientação, seu exemplo de organização e rigor científico em cada etapa de desenvolvimento dos nossos trabalhos. Em particular, pelo seu empenho devotado na coordenação do nosso trabalho publicado na revista *Proteins*, o qual exigiu bastante de todos os autores; também pelo amplo aprendizado adquirido por meio dele ao longo desses anos e por sua amizade.

Ao Professor Dr. Vitor Barbanti Leite e ao Msc. Vinícius Contessoto, ambos do Instituto de Biociências, Letras e Ciências Exatas (IBILCE) da Unesp de São José do Rio Preto, pela colaboração nesse trabalho publicado na *Proteins*.

Desde já agradeço à todos os membros da banca por aceitarem gentilmente o convite.

Aos Professores que participaram da minha banca de qualificação: Dr. Renato Tinós, Dr. Osame Kinouchi Filho e Dr. Ubiraci Pereira da Costa Neves, pelas riquíssimas contribuições que foram importantíssimas para o desenvolvimento desta Tese.

Ao amigos e colaboradores do nosso grupo de pesquisa: Dr. Leandro Gutierrez Rizzi, pelas discussões técnicas e científicas. Ao Dr. Luiz Mostaço Guidolin pelo pacote computacional ctvpdb para manipular arquivos PDB e pela revisão do *bash script* inicial

do pacote que implementamos correspondente ao nosso modelo wGNM. À Msc. Jacyana Saraiva pela demonstração de preocupação com o andamento do trabalho de doutorado. Ao Rafael Frigori pelas discussões e esclarecimentos sobre o modelo AB. Ao Lucas Dadalt, aluno de iniciação científica, pelas discussões nos seminário de grupo e pela colaboração, efetuando simulações de Dinâmica Molecular, em alguns estudos preliminares.

Aos funcionários do Departamento de Física (DF), Julio Cezar, José L. Aziani, Ricardo G. F. dos Santos e Marcilio Mano Jr.; e da antiga Seção de Informática, Dr. Adriano Holanda, Fábio Moretti, Everton Bertolai, Tiago Carrer e Matheus Machado pelo auxílio prestado.

À Dra. Cynthia M. C. P. Manso pelas revisões de inglês dos textos publicados.

Aos amigos e colegas da pós-graduação FAMB: Doutores Olavo H. Menin, Tiago Arruda "Turco", Marcelo A. Pereira, Rodrigo S. González, Lindomar S. dos Santos, Mairon Santos e Natália Destefano; e aos Mestres, Sandro Reia e Cristiano Roberto Fabri Granzotti.

Às secretárias do Departamento de Física e Matemática, Nilza e Sônia, pelos serviços administrativos prestados com prontidão.

Aos Professores Dr. Marcelo Mulato e Dr. Antônio José da Costa Filho, coordenadores do programa de pós-graduação FAMB durante o meu doutorado, pelo apoio e por viabilizar auxílio financeiro para participação de congressos.

Ao Professor Dr. Alexandre Souto Martinez pelo supervisionamento no estágio PAE na disciplina de Estatística I Aplicada à Psicologia.

Ao técnico de informática André Luiz Girol, pelo suporte técnico com eficácia na manutenção dos computadores do nosso laboratório, na instalação de bibliotecas e por sua amizade.

Aos Professores das disciplinas que cursei na FCFRP durante o Doutorado, Prof. Dr. Antonio Caliri, Prof. Dr. Marco Antônio A. da Silva e Prof. Dra. Maria Cristina Nonato Costa.

À agência de fomento CAPES pelo suporte financeiro.

"The atoms come into my brain, dance a dance, and then go out - there are always new atoms, but always doing the same dance, remembering what the dance was yesterday."

(Feynman, 1988)

Resumo

DE MENDONÇA, M.R. **Análise de modos normais dos movimentos conformacionais em proteínas**. 2015. 86 f. Tese (Doutorado - Programa de Pós-graduação em Física aplicada à Medicina e Biologia) - Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto - SP, 2015.

A caracterização das flutuações dos resíduos da proteína em torno do seu estado nativo é essencial para estudar mudanças conformacionais, interação proteína-proteína e interação proteína-ligante. Tal caracterização pode ser capturada pelo modelo de rede gaussiana (GNM). Este modelo tem sido modificado e novas propostas têm surgido nos últimos anos. Nesta Tese, apresentamos um estudo sobre como melhorar o GNM e exploramos o seu desempenho em predizer os fatores-B experimentais. Modelos de redes elásticas são construídos a partir das coordenadas experimentais dos C_α levando em consideração pares de átomos de C_α distantes entre si até um dado raio de corte R_c. Estes modelos descrevem as interações entre os atómos por molas com a mesma constante de força. Desenvolvemos um método baseado em simulações numéricas com um campo de forças simplificado para atribuir pesos a estas constantes de mola. Este método considera o tempo em que dois átomos de C_α permanecem conectados na rede durante o desenovelamento parcial, estabelecendo assim uma forma de medir a intensidade de cada ligação. Examinamos dois diferentes campos de forças simplificados e exploramos o cálculo desses pesos a partir do desenovelamento das estruturas nativas. Nós comparamos o seu desempenho na predição dos fatores-B com outros modelos de rede elástica. Avaliamos tal desempenho utilizando o coeficiente de correlação entre os fatores-B preditos e experimentais. Mostramos como o nosso modelo pode descrever melhor os fatores-B.

Palavras-chave: 1. modelo de rede elástica. 2. fator-B. 3. dinâmica vibracional. 4. análise de modos normais.

Abstract

DE MENDONÇA, M.R. **Normal Mode Analysis of the conformational motions in proteins**. 2015. 86 f. Thesis (Ph.D. - Postgraduate program in Physics applied to Medicine and Biology) - Faculty of Philosophy, Sciences and Literature, University of São Paulo, Ribeirão Preto - SP, 2015.

The characterization of the fluctuations in protein residues around its native state is essential to study conformational changes, protein binding interaction and protein-protein interaction. Such characterization can be captured by simple elastic network models as the Gaussian Network Model (GNM). This model has been modified and new proposals have emerged in recent years. In this Thesis we propose an extended version of GNM, namely wGNM. Elastic network models are built on the experimental C_α coordinates, and they only take the pairs of C_α atoms within a given cutoff distance R_c into account. These models describe the interactions by elastic springs with the same force constant to predicted the experimental B-factors, providing insights into the structure-function properties of proteins. We have developed a method based on numerical simulations with a simple coarse-grained force field, to attribute weights to these spring constants. This method considers the time that two C_α atoms remain connected in the network during partial unfolding, establishing a means of measuring the strength of each link. We examined two different coarse-grained force fields and explored the computation of these weights by unfolding native structures. We compare the B-factors predicted by different elastic network models with the experimental ones employing the correlation coefficient between these two quantities. We show that wGNM performs better and consequently provides better evaluation of the B-factors.

Key-words: 1. elastic network model. 2. B-factor. 3. vibrational dynamics. 4. normal mode analysis.

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

GNM	Modelo de rede gaussiana (*Gausssian Elastic Network Model*).
pfGNM	Modelo de rede gaussiana livre de parâmetros (*parameter-free Gaussian Elastic Network Mode*).
WCN	Número de contatos ponderados (*Weighted-contact Number Model*).
NMA	Análise de modos normais (*Normal Mode Analysis*).
ST	Temperatura estatística (*Statistical Temperature*).
WHAM	Método de análise reponderando histogramas (*Weighted Histogram Analysis Method*).
wGNM	Modelo de rede gaussiana ponderada (*Weighted Gaussian Elastic Network Model*).
MD	Dinâmica molecular (*Molecular Dynamics*).
NMR	Ressonância magnética nuclear (*Nuclear Magnetic Resonance*).

Sumário

Capítulo 1

Introdução

É bem conhecido que as proteínas globulares evoluem de forma espontânea para conformações compactas únicas, cujo mecanismo chamado de enovelamento de proteínas (*protein folding*) é ainda pouco compreendido. Inúmeros estudos, teóricos e experimentais [1, 2, 3], continuam a ser realizados com a finalidade de identificar os possíveis mecanismos envolvidos. As estruturas compactas são chamadas de terciárias e correspondem ao estado biologicamente ativo. Tudo indica que este enovelamento espontâneo, conduzindo à formação da estrutura terciária, é determinado unicamente pela sua sequência primária [4]. As interações entre as cadeias laterais dos aminoácidos são importantes para a estabilidade das proteínas e determinação da função que desempenham em processos biológicos, além de estabelecer a dinâmica do processo de enovelamento. Essas interações, não covalentes entre cadeias laterais, fazem com que haja a estabilização da orientação mútua das estruturas secundárias no estado completamente enovelado, correspondendo ao estado nativo da proteína.

A flexibilidade da proteína está intimamente relacionada com a sua função biológica. Para interagir com uma molécula qualquer, a proteína tem que ser capaz de mudar a sua conformação. Tal mudança conformacional pode ser mínima, envolvendo o movimento de átomos de um resíduo, ou brusca, envolvendo o movimento de domínios inteiros de uma estrutura multimérica. Em organismos termofílicos, as proteínas evoluiram para resistir a altas temperaturas. As proteínas termofílicas são geralmente mais rígidas que suas homologas mesofílicas em temperaturas semelhantes. Entretanto, em temperaturas fisiológicas, as proteínas meso e termofílicas têm flexibilidades equivalentes. Como as enzimas termofílicas tendem a perder a sua atividade em temperaturas abaixo das tem-

peraturas fisiológicas, tem-se sugerido que a função da proteína está correlacionada com a sua estabilidade e que tal correlação deve ser mediada por mudanças na flexibilidade da proteína. Experimentalmente, as flutuações das posições dos resíduos em torno da estrutura nativa podem ser capturadas pelo fator-B por meio da técnica de cristalografia de raios-X. O fator-B de um dado átomo indica a sua mobilidade. Simulações clássicas de dinâmica molecular com todos os átomos fornecem informações detalhadas sobre essas flutuações. Entretanto, essas simulações são lentas computacionalmente e mesmo após semanas de processamento numérico somente amostram flutuações que ocorrem em uma escala temporal de nanosegundos. Cálculos utilizando os resultados da análise de modos normais (NMA) têm mostrado que a NMA é uma técnica robusta capaz de predizer flutuações associadas a largas escalas de tempo físico em um tempo computacional relativamente curto.

A análise de modos normais de moléculas poliatômicas é uma técnica tradicional em espectroscopia molecular. Neste contexto, um método bastante utilizado pelos espectroscopistas é o método da matriz GF [5]. Neste método, as interações entre os átomos são consideradas harmônicas e as vibrações são descritas por coordenadas internas. Além disso, simplificações são feitas por meio de argumentos de simetria da teoria de grupo. Detalhes do sistema, como a geometria da molécula, os ângulos entre as ligações e a natureza das ligações são relevantes neste método. Desde a década de 50, a NMA tem sido efetuada pelo método da matriz GF em moléculas pequenas e tem demonstrado ser uma ferramenta poderosa na predição de espectros de frequências vibracionais.

Os primeiros estudos da NMA em proteínas foram reportados na década de 80. Inicialmente, estes estudos foram realizados por meio de simulações de dinâmica molecular utilizando campos de forças altamente complexos e tendo como variáveis dinâmicas as coordenadas internas. Os modos normais e as frequências normais de vibração eram calculados via diagonalização da matriz Hessiana (matriz das constantes de força). Pelo fato de que os elementos da matriz Hessiana são calculados na conformação de equilíbrio, era necessário efetuar a minimização da energia.

Gō e colaboradores [6] estudaram a dinâmica da proteína bovina pancreática tripsina inibidora (BPTI) utilizando como variáveis dinâmicas apenas os ângulos torcionais entre os planos de ligação e potenciais relativamente mais simples que os usados atualmente. Eles consideraram os ângulos torcionais como sendo as únicas variáveis dinâmicas.

Isto decorre do fato de que as variações dos comprimentos e dos ângulos das ligações serem muito menores que as variações associadas aos ângulos torcionais. Assim, as quantidades comprimento e ângulos de ligação foram consideradas fixas. A vantagem de diminuir o número de variáveis dinâmicas está na redução do tamanho da matriz Hessiana e, consequentemente, na obtenção de maior ganho em eficiência computacional.

Computacionalmente, um grande empecilho nos estudos relatados anteriormente encontrava-se na etapa da minimização da energia. Assumindo a conformação nativa como sendo a conformação de equilíbrio, a autora Tirion [7] desenvolveu um modelo no qual esta etapa é eliminada. Além disso, nesse modelo as interações entre os átomos são descritas por um potencial harmônico com uma única constante de força, para todos os pares de átomos, escolhida por ajuste com o potencial experimental.

Inspirados no modelo desenvolvido por Tirion e na teoria de rede elástica de polímeros desenvolvida por Flory [8], Haliloglu e colaboradores [9] propuseram o modelo de rede gaussiana (GNM). No GNM, a proteína é representada por uma rede elástica tridimensional, onde os vértices da rede são identificados pelas coordenadas dos C_α e as molas representam as interações ligantes e não ligantes entre resíduos que estão distantes entre si até um dado limiar. Assume-se ainda que as flutuações são gaussianas e isotrópicas. No GNM, não há informações sobre as direções dos movimentos. O modelo de rede anisotrópica (ANM) é uma extensão do GNM em que as flutuações são consideradas como sendo anisotrópicas.

Os resultados da análise de modos normais com os modelos GNM e ANM demonstram que estes modelos descrevem muito bem o espectro vibracional de proteínas. Estes modelos têm sido aplicados no estudo do mecanismo e da dinâmica conformacional de diversos sistemas proteícos: hemoglobina [10], HIV *transcriptase reverse* [11, 12], *aspartate transcarbamylase* [13], GroEL-GroES [14], um segmento de *actin* [15], *ribosome* [16, 17], RNA polimerase [18], F1-ATpase [19] e *viral capsids* [20]. Recentemente, esses modelos vêm sendo revisados e modificados. Yang e colaboradores [21] propuseram o modelo pfGNM, no qual não há distância de corte entre as interações, ou seja, considera-se que cada resíduo interage com todos os outros resíduos. Entretanto, neste modelo, a constante de mola da interação entre os resíduos i e j é ponderada pelo inverso do quadrado da distância de separação entre eles na estrutura nativa.

Nesta Tese, propomos uma nova versão do GNM, a qual chamamos de wGNM.

Nesta abordagem, consideramos que o peso associado a interação entre dois C_α é proporcional ao tempo em que eles permanecem conectados quando a estrutura nativa é parcialmente desenovelada. Nosso objetivo é obter uma forma eficiente de atribuir estes pesos. Para isso, analisamos as trajetórias de desenovelamento parcial em diferentes temperaturas, incluindo simulações na temperatura de desenovelamento.

1.1 Organização da Tese

Organizamos a apresentação desta Tese como segue. Inicialmente, no capítulo 2, apresentamos uma visão geral sobre a teoria dos modelos de rede elástica. Historicamente, mostramos como os potenciais empregados para descrever as interações entre os átomos das proteínas evoluiram ao longo dos anos nos trabalhos envolvendo a análise de modos normais. Nos trabalhos pioneiros, adotavam-se potenciais relativamente complexos contendo vários termos para descrever cada tipo de interação entre os átomos. Nos trabalhos decorrentes, esses potenciais evoluiram no sentido da simplificação, tanto em termos da redução do número de parâmetros quanto em termos da redução do número de partículas consideradas. Exemplificamos, na seção 2.1, tal evolução na abordagem proposta por Gō e na abordagem proposta por Tirion. Nas subseções desta seção, introduzimos o modelo de rede gaussiana (GNM) e demonstramos a equação que prediz os fatores-B. Além do GNM, descrevemos também os modelos ponderados pfGNM e WCN. Em seguida, apresentamos na seção 2.2 uma breve revisão da literatura sobre a aplicação do GNM na identificação de resíduos importantes do ponto de vista funcional e estrutural. Ainda nesta seção de Metodologia, apresentamos os campos de forças descritos pelo modelo AB (seção 2.3) e SBM (seção 2.4). Na seção 2.5, descrevemos o algoritmo de análise ST-WHAM. Na seção 2.6, introduzimos a nossa proposta de modelo de rede elástica ponderada chamado de wGNM. No capítulo 3, apresentamos os resultados e discussão. Inicialmente, apresentamos na seção 3.1 os resultados do estudo de caráter exploratório com a finalidade de estabelecer os parâmetros adequados das simulações para o nosso modelo, por exemplo, a temperatura adequada para efetuar as simulações de desenovelamento parcial das proteínas e a escolha da escala de hidropatia para classificar cada resíduo da proteína como hidrofóbico ou hidrofílico no modelo AB. Tal estudo foi efetuado para um conjunto de 818 estruturas não redundantes extraídas do *Protein Data Bank* (PDB). Em seguida, apre-

sentamos, na seção 3.2, os resultados do trabalho que publicamos recentemente na revista *Proteins* no qual introduzimos o nosso modelo wGNM e o validamos para um conjunto de proteínas do PDB resolvidas pela técnica de cristalografia de raios-X. Finalizamos este capítulo, apresentando os resultados do estudo em que aplicamos o wGNM para um conjunto de estruturas resolvidas pela técnica de NMR. No capítulo 4, apresentamos as conclusões finais e perspectivas. Finalmente, deixamos para o Apêndice A, a apresentação da demonstração de equações e para o Apêndice B, a apresentação da determinação experimental do fator-B.

Capítulo 2

Metodologia

Descreveremos neste capítulo a teoria dos modelos de rede elástica. Inicialmente, apresentaremos a abordagem proposta por Gō e colaboradores [6] e, posteriormente, a proposta de Tirion [7]. Na sequência descreveremos o modelo de rede gaussiana e os modelos ponderados pfGNM e WCN. Por fim, apresentamos a nossa proposta de modelo rede de elástica intitulado "Modelo de rede gaussiana ponderada (wGNM)".

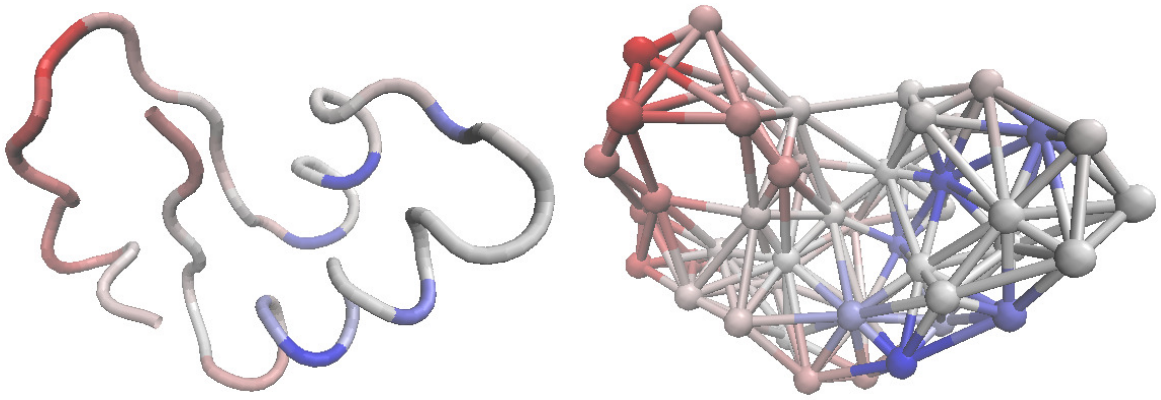
2.1 Modelos de rede elástica (ENM)

Nos ENM, a estrutura terciária da proteína (figura 2.1(a)) é representada por uma rede elástica de C_α conectados por molas virtuais. Nesta rede, os vértices são identificados pelas posições dos C_α e as hastes (molas) representam as interações ligantes e não ligantes entre os C_α conforme ilustrado na figura 2.1(b).

Os primeiros estudos da dinâmica de macromoléculas por meio da análise de modos normais iniciaram-se na década de 80. Em geral, utilizavam-se simulações de dinâmica molecular (MD) envolvendo todos os átomos. Nestas simulações, adotavam-se campos de forças relativamente complexos, por exemplo,

$$\begin{aligned} E_p &= \frac{1}{2}\sum_{\text{ligações}} k_b(b-b_0)^2 + \frac{1}{2}\sum_{\text{ângulos}} k_\theta(\theta-\theta_0)^2 \\ &+ \frac{1}{2}\sum_{\text{torsões}} k_\phi[(1+\cos(n\phi-\delta)] \\ &+ \sum_{\text{não ligantes}} \left[\frac{A}{r^{12}} - \frac{B}{r^6} + \frac{q_1 q_2}{Dr}\right]. \end{aligned} \tag{2.1}$$

Nesta expressão, o primeiro termo leva em conta a contribuição da energia associada às variações dos comprimentos das ligações, o segundo termo está associado às variações dos



(a) estrutura terciária (b) rede elástica

Figura 2.1: *(a) Diagrama da proteína cujo código PDB é 1CNR, (b) representação da estrutura protéica por uma rede elástica.*

ângulos entre as ligações, o terceiro termo inclui as variações dos ângulos torcionais e o último termo leva em conta as interações não ligantes.

Em MD, o custo computacional da simulação é diretamente proporcional ao número de átomos do sistema e da complexidade do campo de forças escolhido.

Com o intuito de reduzir o custo computacional nas simulações de dinâmica molecular, Gō e colaboradores [6] propuseram um campo de forças relativamente mais simples do que o apresentado na equação (2.1). Para isso, eles conjecturaram que a dinâmica dos átomos nas proteínas é governada pelas interações de curto alcance,

$$V \approx \sum_i \left[\frac{k_1}{2}(r_{i,i+1} - r^0_{i,i+1})^2 + \frac{k_2}{4}(r_{i,i+1} - r^0_{i,i+1})^4 \right] + \sum_{i,j} 4\epsilon \left(\frac{1}{r^6_{ij}} - \frac{1}{r^{12}_{ij}} \right), \qquad (2.2)$$

no qual o primeiro termo abrange as oscilações harmônicas e anarmônicas e o último as interações não ligantes descritas pelo potencial clássico de Lenard-Jonnes. Aqui r_{ij} é a distância entre os átomos i e j e r^0_{ij} é a correspondente distância na conformação de equilíbrio. Este campo de forças é usualmente conhecido como potencial de Gō, o qual foi amplamente empregado no estudo dos processos de desenovelamento de proteínas nas décadas passadas.

Inspirado no modelo proposto por Gō, Tirion [7] propõe um modelo no qual o processo de minimização da energia é eliminado. Em termos de eficiência computacional, o grande sucesso desse modelo que indubitavelmente acarretou em um avanço nas abordagens *coarse-grained*, foi a escolha da conformação da proteína no estado enovelado obtido

do PDB como sendo a conformação de equilíbrio já minimizada. As interações entre pares de resíduos são descritas pelo potencial simplificado com um único parâmetro,

$$E_p\left(\mathbf{r}_a, \mathbf{r}_b\right) = \frac{C}{2}(|\mathbf{r}_{a,b}| - \left|\mathbf{r}^0_{a,b}\right|)^2, \tag{2.3}$$

em que $\mathbf{r}_{a,b} \equiv \mathbf{r}_a - \mathbf{r}_b$ é o deslocamento do resíduo b em relação ao resíduo a, correspondendo ao elongamento da mola virtual conectando o resíduo a ao b. A energia potencial da proteína então é dada por

$$E = \sum_{(a,b)} E_p\left(\mathbf{r}_a, \mathbf{r}_b\right). \tag{2.4}$$

Nesta equação, o somatório é restrito aos pares de resíduos cuja distância entre eles é menor que $R_{\mathrm{vdW}}(a)+R_{\mathrm{vdW}}(b)+R_c$, onde $R_{\mathrm{vdW}}(a)$ é o raio de van der Walls do resíduo a, $R_{\mathrm{vdW}}(b)$ o raio de van der Walls do resíduo b e R_c o raio de corte delimitando o alcance das interações. Na equação (2.3), C representa uma constante de força fenomenológica, a qual assume o mesmo valor para todos os pares de resíduos conectados.

2.1.1 Modelo de rede gaussiana (GNM)

Inspirados no modelo de rede elástica proposto por Tirion [7] e na teoria de rede elástica de polímeros desenvolvida por Flory [8], Haliloglu e colaboradores [9] propuseram o modelo de rede gaussiana. A finalidade desse modelo é explorar a dinâmica da rede elástica, definida pela sua estrutura $3D$, para obter informações relevantes sobre as flutuações térmicas dos átomos nas proteínas.

No GNM, a proteína é descrita como uma rede elástica tridimensional (ver figura 2.2(a)), onde os vértices da rede são identificados pelas coordenadas dos resíduos e as molas representam as interações ligantes e não ligantes entre eles. Duas suposições são feitas em relação às flutuações:

(i) são isotrópicas;

(ii) seguem uma distribuição de probabilidade gaussiana.

É conveniente estudarmos a rede elástica em relação a um sistema de referência fixo no laboratório (ver figura 2.2(b)).

Uma vez escolhido o sistema de referência, a posição de equilíbrio do vértice i da rede é determinada pelo vetor posição $\mathbf{R}^0_i$ e a posição instantânea por $\mathbf{R}(i)$. As flutuações

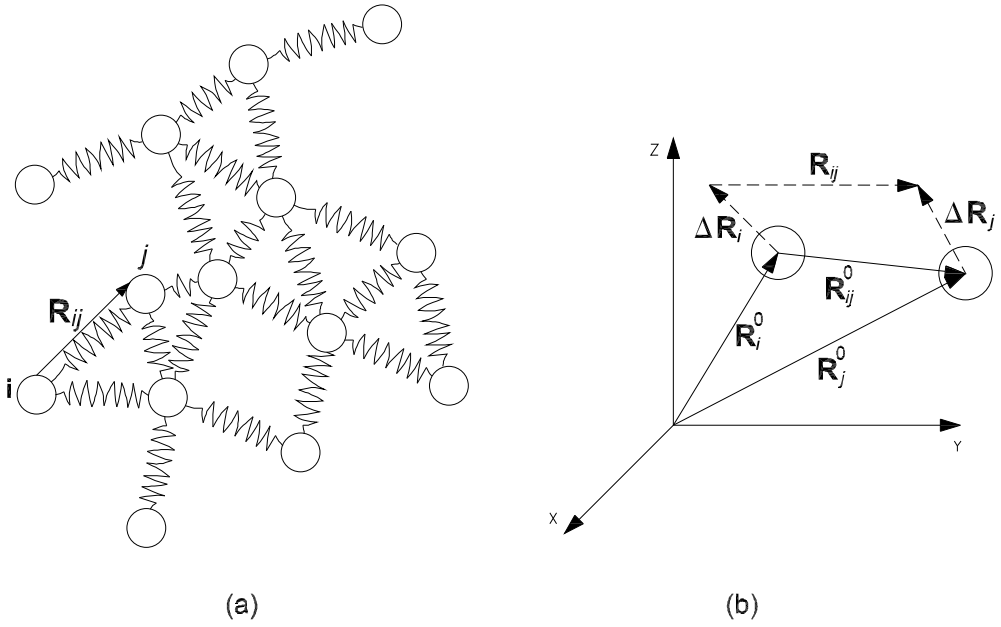


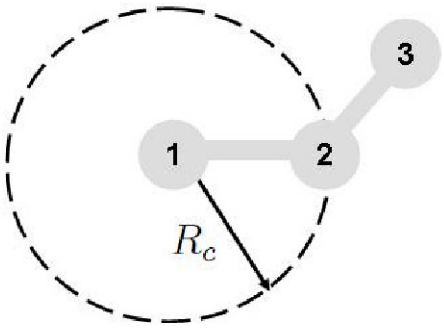
Figura 2.2: *(a) Representação da proteína por uma rede elástica tridimensional, onde os vértices da rede são definidos pelas coordenadas dos resíduos e as molas representam as interações não ligantes entre eles. Neste modelo, considera-se que todas as molas possuem constante de força γ, (b) representação do sistema de referência fixo no laboratório para descrever as posições dos resíduos i e j.*

ou deformações das posições médias podem ser definidas pelo vetor $\Delta\mathbf{R}_i = \mathbf{R}_i - \mathbf{R}_i^0$. Semelhantemente, as flutuações no vetor distância $\mathbf{R}_{ij}$, entre os resíduos i e j, são dadas por $\Delta\mathbf{R}_{ij} = \mathbf{R}_{ij} - \mathbf{R}_{ij}^0 = \Delta\mathbf{R}_j - \Delta\mathbf{R}_i$.

Tendo em vista as suposições (i) e (ii), podemos escrever a Hamiltoniana da rede de N resíduos, $\mathcal{H}_{\mathrm{GNM}}$, em função das deformações no vetor distância $\Delta\mathbf{R}_{ij}$ da seguinte forma

$$\begin{aligned}
\mathcal{H}_{\mathrm{GNM}} &= \frac{\gamma}{2}\sum_{i=1}^{N}\sum_{j>i}^{N}(\boldsymbol{\Delta}\mathbf{R}_{ij})^2 H(R_c - \left|\mathbf{R}_{ij}^0\right|) \\
&= \frac{\gamma}{2}\sum_{i=1}^{N}\sum_{j>i}^{N}(\mathbf{R}_{ij} - \mathbf{R}_{ij}^0).(\mathbf{R}_{ij} - \mathbf{R}_{ij}^0) H(R_c - \left|\mathbf{R_{ij}}^0\right|) \qquad (2.5) \\
&= \frac{\gamma}{2}\sum_{i=1}^{N}\sum_{j>i}^{N}(\Delta\mathbf{R}_j - \Delta\mathbf{R}_i).(\Delta\mathbf{R}_j - \Delta\mathbf{R}_i) H(R_c - \left|\mathbf{R_{ij}}^0\right|) \qquad (2.6)
\end{aligned}$$

onde $H(R_c - \left|\mathbf{R_{ij}}^0\right|)$ é a função de passo Heaviside (assumindo valor unitário se o argumento for maior do que zero e nulo, caso contrário) e γ é a constante de força da mola.



$$
\begin{aligned}
\mathcal{H}_{\mathrm{GNM}} &= \frac{\gamma}{2}\left[(\Delta\mathbf{R}_{12})^2 + (\Delta\mathbf{R}_{23})^2\right] \\
&= \frac{\gamma}{2}\left[(\Delta\mathbf{R}_2 - \Delta\mathbf{R}_1)^2 + (\Delta\mathbf{R}_3 - \Delta\mathbf{R}_2)^2\right] \\
&= \frac{\gamma}{2}\begin{bmatrix} \Delta\mathbf{R}_1 & \Delta\mathbf{R}_2 & \Delta\mathbf{R}_3 \end{bmatrix} \underbrace{\begin{bmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{bmatrix}}_{\text{matriz de kirchhoff } \mathbf{\Gamma}} \begin{bmatrix} \Delta\mathbf{R}_1 \\ \Delta\mathbf{R}_2 \\ \Delta\mathbf{R}_3 \end{bmatrix}
\end{aligned}
$$

Figura 2.3: *Energia potencial de um tripeptídeo. Como a distância entre o resíduo 1 e o 3 é maior que R_c, desconsidera-se esta interação. A equação final da Hamiltoniana $\mathcal{H}_{\mathrm{GNM}}$ na forma matricial é dada pela equação (2.7).*

A equação acima pode ser expressa na seguinte forma quadrática

$$\mathcal{H}_{\mathrm{GNM}} = \frac{\gamma}{2}\mathbf{\Delta R}^T\mathbf{\Gamma}\mathbf{\Delta R}, \tag{2.7}$$

onde $\mathbf{\Delta R} = [\mathbf{\Delta R}_1, \mathbf{\Delta R}_2, ..., \mathbf{\Delta R}_N]$ é um hipervetor de dimensão N cujas componentes são vetores e $\mathbf{\Gamma}$ é a matriz de Kirchhoff (ou matriz conectividade) das interações de contatos entre resíduos. Esta matriz tem elementos Γ_{ij}:

$$\Gamma_{ij} = \begin{cases} -1, & \text{se} \quad i \neq j \quad \text{e} \quad R_{ij} \leq R_c \\ 0, & \text{se} \quad i \neq j \quad \text{e} \quad R_{ij} > R_c \\ -\sum_{i=1,k\neq i}^{N} \Gamma_{ik}, & \text{se} \quad i = j. \end{cases}$$

Demonstramos a equação (2.7) ilustrativamente na figura 2.3.

Neste modelo, demonstra-se (ver apêndice A.1) que os deslocamenteos quadráticos

médios associados aos resíduos i e j podem ser escritos como

$$\left\langle \mathbf{\Delta R_i}^{\mathbf{T}} \mathbf{\Delta R_i} \right\rangle = \frac{3k_B T}{\gamma} \left(\mathbf{\Gamma}^{-1} \right)_{ii} \tag{2.8}$$

e

$$\left\langle \mathbf{\Delta R_i}^{\mathbf{T}} \mathbf{\Delta R_j} \right\rangle = \frac{3k_B T}{\gamma} \left(\mathbf{\Gamma}^{-1} \right)_{ij}. \tag{2.9}$$

O determinante da matriz de Kirchhoff $\mathbf{\Gamma}$ é nulo, portanto, a princípio ela não é inversível (matriz singular). Entretanto, ela pode ser decomposta em termos de uma matriz $\mathbf{U}$, cujas colunas são os autovetores $\mathbf{u}_i$, e da matriz diagonal $\mathbf{\Lambda}$, com autovalores $\lambda_\mathbf{i}$, da seguinte forma

$$\mathbf{\Gamma} = \mathbf{U} \mathbf{\Lambda} \mathbf{U}^T, \tag{2.10}$$

onde os autovalores da matriz de Kirchhoff representam as frequências dos modos normais individuais enquanto que os autovetores definem as direções dos modos. Por conveniência os autovalores são ordenados em ordem crescente. O primeiro autovalor é nulo, $\lambda_1 = 0$, o qual corresponde à translação da proteína como um todo. Ignorando o primeiro autovalor, a matriz pseudo-inversa de Kirchhoff pode ser escrita como

$$\mathbf{\Gamma}^{-1} = \sum_{k=2}^{N} \lambda_k^{-1} \mathbf{u}_k \mathbf{u}_k^T. \tag{2.11}$$

Consequentemente, as correlações cruzadas e as autocorrelações associadas às flutuações dos resíduos i e j podem ser expressas respectivamente como:

$$\left\langle \mathbf{\Delta R}_i^T \mathbf{\Delta R}_j \right\rangle = \frac{3k_B T}{\gamma} \sum_{k=2}^{N} \lambda_k^{-1} \mathbf{u}_{ik} \mathbf{u}_{jk}^T, \tag{2.12}$$

$$\left\langle \mathbf{\Delta R}_i^T \mathbf{\Delta R}_i \right\rangle = \frac{3k_B T}{\gamma} \sum_{k=2}^{N} \lambda_k^{-1} \mathbf{u}_{ik}^2. \tag{2.13}$$

O fator-B, o qual mede a mobilidade dos resíduos, é definido por

$$B_i = \frac{8\pi^2}{3} \left\langle (\mathbf{\Delta R_i})^2 \right\rangle. \tag{2.14}$$

Utilizando a equação (2.13), podemos escrevê-lo como

$$B_i = \frac{8\pi^2 k_B T}{\gamma} \sum_{k=2}^{N} \lambda_k^{-1} \mathbf{u}_{ik}^2. \tag{2.15}$$

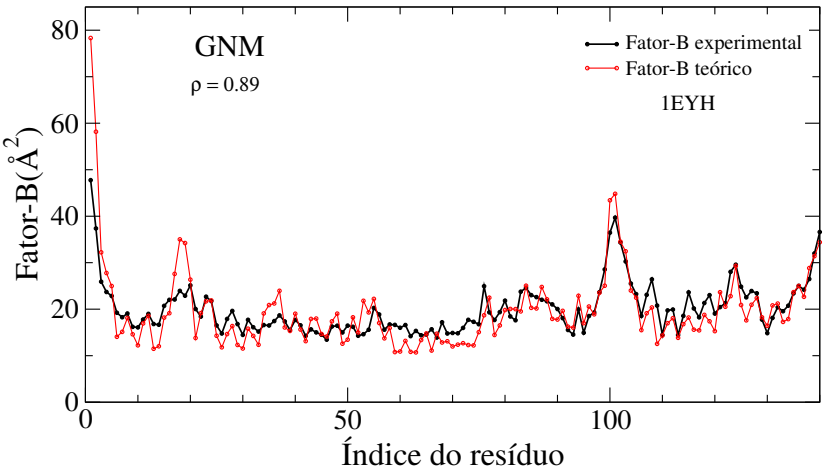


Figura 2.4: *Comparação entre os fatores-B teóricos e experimentais para proteína 1EYH.*

Para validar o GNM, é utilizado o coeficiente de correlação de Pearson,

$$\rho = \frac{\sum_{i=1}^{N}(B_i^{\text{teórico}} - \bar{B}^{\text{teórico}})(B_i^{\text{experimental}} - \bar{B}^{\text{experimental}})}{\sqrt{\sum_{i=1}^{N}(B_i^{\text{teórico}} - \bar{B}^{\text{teórico}})^2}\sqrt{\sum_{i=1}^{N}(B_i^{\text{experimental}} - \bar{B}^{\text{experimental}})^2}},$$

como medida de comparação entres os fatores-B teóricos e experimentais. A figura 2.4 mostra esta comparação para a proteína 1EYH. Desenvolvemos um programa de computador para calcular os valores B_i e a correlação com os valores experimentais apresentados nos arquivos PDB.

Os pontos nas curvas ilustradas na figura 2.4 para proteína 1EYH estão associados ao fator-B do C_α no aminoácido correspondente. Observa-se que as curvas teórica e experimental estão bem correlacionadas, apresentando comportamentos similares. O coeficiente de correlação de Pearson entre elas é $\rho = 0.89$.

Na equação (2.15), γ é uma constante a ser determinada. Usualmente, ela é obtida por meio da normalização dos fatores-B teóricos com os experimentais. Razão pela qual nos referimos a ela como constante de normalização. Fazendo $B_i = B_i^{\text{GNM}}$, podemos obter o seu valor (em unidades de k_BT) através da expressão

$$\gamma = \frac{\sum_{i=1}^{N} B_i^{\text{GNM}}}{\sum_{i=1}^{N} B_i^{\text{experimental}}}.$$

2.1.2 Modelo de rede gaussiana livre de parâmetro (pfGNM)

Este modelo é uma versão alternativa ao GNM, proposta por Yang e colaboradores [21], na qual:

i) não existe uma distância de corte delimitando as interações entre os C_α, de modo que cada C_α interage com todos os demais C_α da rede;

ii) os elementos i e j da matriz de Kirchhoff do GNM são ponderados pelo inverso do quadrado da distância R_{ij}, avaliada na conformação nativa, de acordo com a equação:

$$\mathbf{\Gamma}_{ij}^{pf} = \begin{cases} R_{ij}^{-2} & \text{se } i \neq j \\ -\sum_{i,j\neq i} \mathbf{\Gamma}_{ij}, & \text{se } i = j. \end{cases}$$

Utilizando-se esta matriz ponderada no GNM para predizer os fatores-B téoricos, Yang e colaboradores [21] mostraram que tal predição melhorou para as proteínas analisadas, quando comparada com o modelo não ponderado.

2.1.3 Número de contatos ponderados (WCN)

Esta abordagem, proposta por Lin e colaboradores [22], utiliza o conceito número de contatos (NC) de um vértice qualquer da rede em termos do número de vizinhos com os quais ele está conectado. Considera-se que dois vizinhos estão conectados se a distância entre eles for menor que R_c.

No WCN, define-se o NC dependente da distância da seguinte forma

$$\nu_i = \sum_{j\neq i}^{N} H\left(R_c - R_{ij}\right)/R_{ij}^2, \tag{2.16}$$

onde $H(R_c - |R_{ij}|)$ é a função de passo Heaviside (assumindo valor unitário se o argumento for maior do que zero e nulo, caso contrário). Nesta equação, a soma é ponderada pelo recíproco do quadrado da distância entre pares de C_α. Como o termo $1/R_{ij}^2$ decai rapidamente para grandes separações, então podemos simplificá-la para

$$\nu_i = \sum_{j\neq i}^{N} \frac{1}{R_{ij}^2}. \tag{2.17}$$

O perfil WCN do fator-B de uma proteína com N resíduos é definido por

$$\mathbf{w} = (\omega_1, \omega_2, ..., \omega_N)\,, \tag{2.18}$$

onde $w_i = 1/\nu_i$.

O coeficiente de correlação entre o fator-B experimental e o teórico calculado pelo WCN pode ser obtido por meio da correlação entre os dois vetores $\mathbf{b}$ e $\mathbf{w}$, sendo $\mathbf{b}$ o perfil do fator-B experimental descrito por

$$\mathbf{b} = (b_1, b_2 ... b_N)\,, \tag{2.19}$$

onde b_i é o fator-B experimental do i-ésimo C_α da estrutura nativa obtida no PDB.

Assim, nesta abordagem não é necessário diagonalizar a matriz de Kirchhoff. As flutuações térmicas são calculadas diretamente da equação (2.18). A eliminação dessa etapa resulta em uma diminuição significativa das operações de ponto flutuante, reduzindo o custo computacional. Lin e colaboradores [22] mostraram que os resultados obtidos pelo WCN são competitivos com os determinados pelo GNM.

2.1.4 Aplicação dos ENM para estruturas de NMR

A estrutura de uma proteína resolvida pela técnica de ressonância magnética nuclear (NMR) é composta por um conjunto (ensemble) de vários modelos. Considerando cada modelo dessa proteína como sendo um *frame*, se visualizarmos a representação 3D de todos os modelos de forma sequencial, teremos o filme da flutuação (oscilação) da estrutura dessa proteína em torno da sua estrutura de equilíbrio. Quantitativamente, podemos avaliar estas flutuações por meio do *rmsd* (*root mean square deviation*) entre os diferentes modelos de NMR. Para um ensemble composto por m conformações, o $rmsd_i$ do resíduo i é calculado pela equação

$$rmsd_i = \sqrt{\frac{\sum_{k=1}^{m} |\mathbf{r}_{i,k} - \bar{\mathbf{r}}_i|^2}{m}}, \tag{2.20}$$

onde $\bar{\mathbf{r}}_i = \dfrac{\sum_{k=1}^{m} \mathbf{r}_{i,k}}{m}$, o qual representa a posição média ou de equilíbrio entre os diferentes modelos [23].

Tomando a conformação de um dado modelo de NMR como referência, podemos predizer os deslocamentos quadráticos médios de cada C_α por meio do GNM, por exemplo.

O poder de tal predição pode ser avaliado pelo coeficiente de correlação entre o perfil das flutuações quadráticas médias teóricas calculadas pelo GNM e o perfil do *rmsd* da proteína (flutuações quadráticas médias experimentais). Como no GNM a estrutura 3D da proteína define os movimentos coletivos acessíveis em torno do estado nativo, esperamos que o GNM seja capaz de amostrar melhor as flutuações para as estruturas de NMR do que para estruturas de raios-X, visto que a liberdade dos átomos se movimentar em solução é maior que na estrutura rígida do cristal.

2.2 Identificação de resíduos funcional e estruturalmente importantes

Nos modelos de rede elástica, os modos normais mais lentos (de baixas frequências) fornecem informações sobre os movimentos coletivos dos átomos, por exemplo, movimento de domínios ou de *loops*, enquanto que os modos mais rápidos (de altas frequências) dizem respeito aos movimentos individuais dos átomos. Mostramos, neste capítulo, que a análise do perfil dos modos mais rápidos permite identificar, em boa concordância com os dados experimentais, os resíduos *hot spots* e sítios ligantes. Além disso, por meio dessa análise é possível identificar os resíduos conservados e os resíduos que compõem o núcleo do enovelamento (*folding nuclei*) [24, 25, 26, 27, 28].

2.2.1 Resíduos conservados e o núcleo do enovelamento

É conhecido que proteínas com sequências primárias muito diferentes entre si podem apresentar conformações 3D semelhantes. Acredita-se que isso ocorre porque existem alguns resíduos em comum entre essas sequências que desempenham papéis importantes no *folding* e na estabilidade de suas estruturas. Esses resíduos em comum são chamados de resíduos conservados ou resíduos chave. De acordo com o seu papel na proteína, os resíduos conservados podem ser classificados em funcionais ou estruturais. Os resíduos funcionais localizam-se apenas nos sítios ativos das proteínas enquanto que os estruturais são distribuídos, geralmente, no núcleo da proteína [29].

É amplamente aceito que o enovelamento de proteínas pequenas com apenas um domínio é bem descrito pelo mecanismo de nucleação e condensação (ver figura 2.5). Galzitskaya e colaboradores [30] chamam de *folding nucleus* a estrutura formada pela

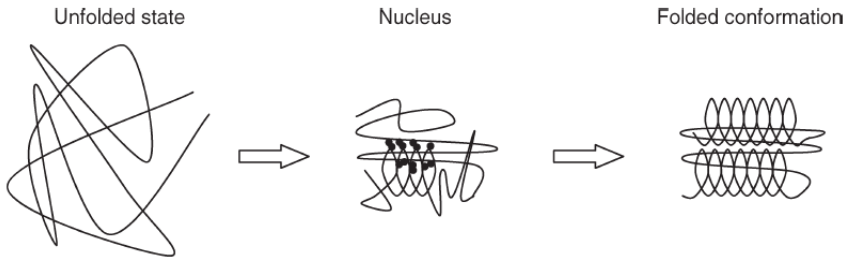


Figura 2.5: *Mecanismo de nucleação e condensação do folding protéico: o enovelamento é guiado pela formação inicial de um núcleo constituído de aminoácidos o qual possui algumas interações características de estruturas secundárias e terciárias corretas de modo que a estrutura adicional pode rapidamente condensar a este núcleo. Figura extraída do artigo [33].*

cadeia da proteína quando ela está no estado de transição. Por outro lado, Shmygelska e colaboradores [31], se referem ao núcleo do enovelamento como sendo um conjunto de contatos nativos que desempenham um papel importante no *folding*. Experimentalmente, identificar o núcleo do enovelamento é uma tarefa extremamente difícil, pois é necessário encontrar os resíduos cujas mutações afetam a taxa de *folding* de modo que a estabilidade dos estados de transição seja tão forte quanto a do estado nativo. Métodos experimentais para identificar o núcleo do enovelamento inclui a mutagênesis de sítio dirigida. Nesta técnica, um resíduo é classificado como pertencente ao núcleo do enovelamento de acordo com o seu valor ϕ. [1]

Shakhnovich e colaboradores [32] mostraram que, para a proteína inibidora de quimotripsina 2 (CI2), a maioria dos resíduos conservados A35, I39, L68, I70 e I76 participam do núcleo do enovelamento. Assim, encontrar o núcleo do enovelamento da proteína junto com o caminho do enovelamento (*folding pathway*) é um dos problemas mais relevantes em Biologia Estrutural. A predição do núcleo do enovelamento pode levar a descrição detalhada da hierarquia do processo de *folding* e possibilitar a identificação de contatos de alta importância estrutural. Tal predição contribui assim para a solução do problema do *folding*, fazendo com que o espaço conformacional que precisa ser pesquisado torne-se mais restrito. Conhecimentos obtidos sobre o estudo do núcleo do enovelamento podem ser úteis também para o desenho racional de drogas.

[1] $\phi = \frac{\Delta lnk_f}{\Delta lnk}$, onde k_f é a constante de *folding* e k é a constante de equilíbrio (k_f/k_u).

2.2.2 Definições de resíduos *hot spots*

Para realizar as suas funções biológicas, a grande maioria das proteínas interage com outras proteínas de modo a formar um complexo estável. A compreensão da formação desse complexo leva a muitas aplicações práticas como o desenho racional de novas drogas. Descobriu-se que a energia livre de ligação do complexo proteína-proteína não é uniforme e que alguns resíduos da interface contribuem significativamente para essa energia, os quais são chamados de *hot spots*. Por outro lado, trabalhos apresentados na literatura envolvendo a técnica *alanine scanning mutagenesis* definem os resíduos *hot spots* como sendo aqueles que quando mutados para alanina experimentam um aumento na energia livre de ligação de pelo menos $\Delta\Delta\mathbf{G} > 2$ kcal/mol [34, 35]. Como o aminoácido alanina não possui uma cadeia lateral além do C_β, esta técnica testa a importância de grupos de cadeias laterais individuais para a formação do complexo. Tem sido mostrado que os resíduos *hot spots* coincidem com os resíduos conservados e com os resíduos do núcleo do enovelamento (importante para estabilidade da proteína).

Bogan e colaboradores [34] estimaram a porcentagem de cada tipo de aminoácido em uma base de dados de 2325 mutações de alanina. Eles mostraram que apenas 3 aminoácidos aparecem com frequência acima que 10% com $\Delta\mathbf{G} > 2$ kcal/mol (*hot spots*) na seguinte proporção; 21 % de triptofano, 13% de arginina e 12.3 % de tirosina.

2.2.3 Análise do perfil dos modos mais rápidos

No GNM, a correlação cruzada associada com o k-ésimo modo é dada por

$$\langle \Delta\mathbf{R}_i \Delta\mathbf{R}_j \rangle = (3k_\mathrm{B}T/\gamma)\lambda_k^{-1}[\mathbf{u}_k]_i[\mathbf{u}_k]_j. \tag{2.21}$$

Logo, a contribuição para as flutuações associada a um dado subconjunto de modos $k_1 \leqslant k \leqslant k_2$ é:

$$\langle (\Delta R_i)^2 \rangle_{k_1-k_2} = (k_\mathrm{B}T/\gamma) \sum_{k_1}^{k_2} \lambda_k^{-1}[\mathbf{u}_k]_i^2 \Bigg/ \sum_{k_1}^{k_2} \lambda_k^{-1}. \tag{2.22}$$

A figura 2.6 mostra o perfil das flutuações quadráticas médias, considerando-se apenas os cinco modos mais rápidos para a proteína CI2 utilizando a equação (2.22) e

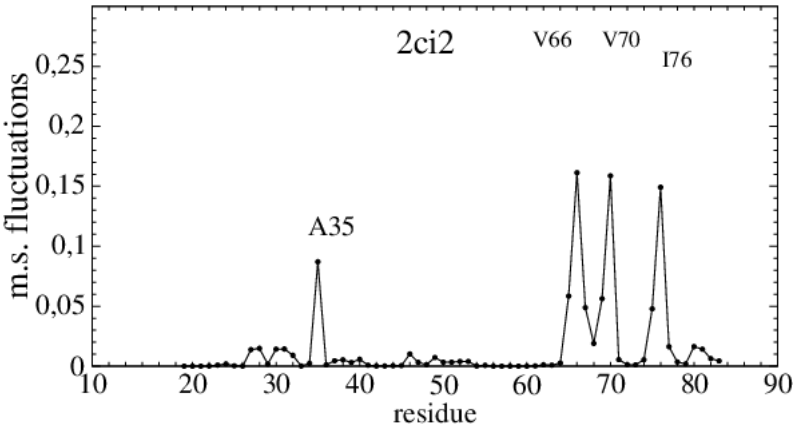


Figura 2.6: *Flutuações quadráticas médias dos resíduos considerando-se somente os cinco modos mais rápidos para proteína de código PDB 2CI2 constituída por 65 resíduos. Estes resultados foram obtidos com o programa* **identify_hot_residues_gnm.f** *desenvolvido pelo nosso grupo e reproduz a figura (6a) da referência [24] sob as mesmas condições.*

$R_c = 7$ Å. A maioria dos picos coincide com os resíduos indenticados por Shakhnovich e colaboradores [32], os quais são conservados e fazem parte do núcleo do enovelamento.

2.2.4 Identificação de sítios ligantes

Via GNM, obtêm-se as flutuações quadráticas médias

$$\langle \Delta\mathbf{R}_i \Delta\mathbf{R}_j^{\mathrm{T}} \rangle = k_B T \Gamma^{-1}. \tag{2.23}$$

Demonstra-se que a correlação entre as flutuações da energia (ΔU) e as flutuações das posições dos resíduos i e j é dada por

$$\langle \Delta U \Delta\mathbf{R}_i \Delta\mathbf{R}_j^{\mathrm{T}} \rangle = k_B T \langle \Delta\mathbf{R}_i \Delta\mathbf{R}_j^{\mathrm{T}} \rangle. \tag{2.24}$$

Assim, as flutuações da energia são distribuídas aos resíduos em proporção a correlação das flutuações.

Uma situação mais relevante fisicamente é encontrar como que as flutuações da energia afetam a distância entre dois resíduos. Da equação (2.24), pode-se demonstrar que

$$\langle \Delta U (\boldsymbol{\Delta\mathbf{R}}_{\mathbf{ij}})^2 \rangle = (k_B T)^2 [(\boldsymbol{\Gamma})_{ii} - 2(\boldsymbol{\Gamma}_{ij}^{-1}) + (\boldsymbol{\Gamma}_{jj}^{-1})]. \tag{2.25}$$

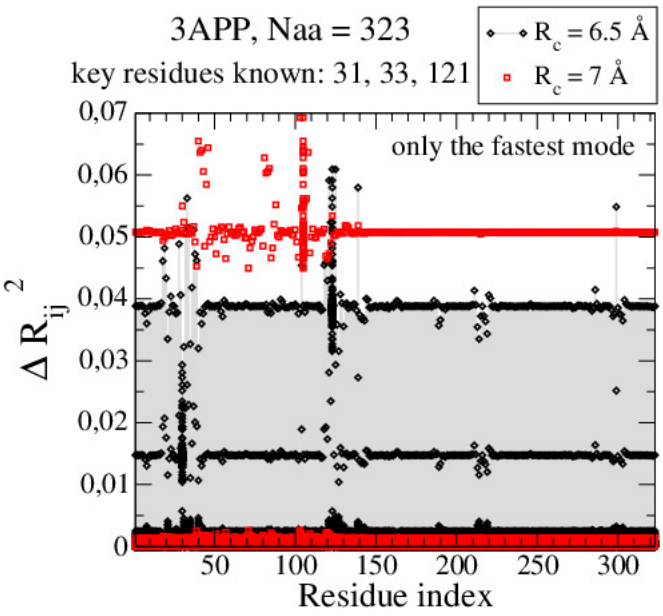


Figura 2.7: *Mapa de distâncias entre o resíduo i e todos os resíduos restantes, $j = 1, \cdots, N$, da proteína 3APP. Os picos indicam aqueles resíduos que exibem maior correlação da flutuação da distância com os demais resíduos da estrutura. Estes resultados foram obtidos como o pacote desenvolvido pelo nosso grupo o qual reproduz a figura do material suplementar da referência [25].*

A figura 2.7 exibe $\langle \Delta R_{ij}^2 \rangle$ correspondente à distância entre o resíduo i e todos os resíduos restantes, $j = 1, \cdots, N$, para proteína 3APP. Os picos indicam aqueles resíduos que exibem maior correlação da flutuação da distância com os demais resíduos da estrutura. Pares de resíduos são classificados de acordo com seus valores relativos das flutuações. Em seguida, os resíduos que aparecem no topo de 2-3 % entre todos os pares de resíduos são aqueles com maiores interações com outros resíduos na estrutura. Estes são interpretados como sendo os resíduos chave de ligação.

2.3 Modelo AB

No seção 2.6, apresentaremos o nosso modelo de rede elástica, onde atribuímos pesos à interação entre os vértices dessa rede. Ou seja, nesta representação atribuímos pesos às bordas da rede como sendo proporcionais ao tempo em que os resíduos permanecem conectados durante um intervalo de desenovelamento parcial. Para simular o desenovelamento proteíco, utilizamos o modelo AB via simulações de Monte Carlo (MC). O modelo AB é um modelo minimalista no qual os resíduos hidrofóbicos são representados por A e os hidrofílicos por B [36, 37]. As interações são descritas pelo potencial

$$E = \frac{1}{4}\sum_{k=2}^{N-2}(1-\cos\vartheta_k) + 4\sum_{k=1}^{N-2}\sum_{j=i+1}^{N}\left(\frac{1}{r_{ij}^{12}} - \frac{C(\sigma_i,\sigma_j)}{r_{ij}^{6}}\right), \tag{2.26}$$

onde $0 \leq \vartheta_k \leq \pi$ é o ângulo entre os sucessivos vetores de ligação e $\sigma_i = \{A, B\}$. O primeiro termo na equação (2.26) corresponde a uma interação ferromagnética, no sentido de que custa energia para dobrar a cadeia. O segundo termo depende da distância entre os monômeros e introduz competição. O fator $C(\sigma_i, \sigma_j)$ controla a dependência da valência do potencial de Lennard-Jones. $C(\sigma_i, \sigma_j)$ possui valor $+1$ para o par $\{A, A\}$ (forte atração), $+0.5$ para o par $\{B, B\}$ (fraca atração) e -0.5 para os pares $\{A, B\}$ e $\{B, A\}$ (fraca repulsão). O programa para simulação utilizando o potencial (2.26) foi desenvolvido por nosso grupo. Utilizamos o algoritmo *spherical-cap* [37] para atualizar as configurações nas simulações de Monte Carlo. Partindo-se da configuração da estrutura nativa, obtida no PDB, propõe-se uma nova configuração a uma temperatura adequada para a proteína se desenovelar. Tal configuração é aceita se, e o somente se, o critério de Metrópolis

$$p(E - E') = \min\left[1, e^{-\beta(E'-E)}\right], \tag{2.27}$$

for satisfeito. Sendo $p(E-E')$ a probabilidade de transição do estado com energia E para o estado com energia E' e β o inverso da temperatura.

Na equação (2.26), temos que a energia potencial do *backbone* depende da hidropatia (caráter hidrofóbico ou hidrofílico) dos resíduos que o compõe, a qual é definida por meio de escalas experimentais de hidrofobicidade. Definimos um resíduo como sendo A ou B, de acordo com o seu valor na escala de hidrofobicidade. Se o valor da hidrofobicidade for menor ou igual a zero, dizemos que o resíduo é B, hidrofílico, caso contrário ele é A, hidrofóbico. Neste trabalho, utilizamos as escalas de hidrofobicidade *Kyte_Doolittle*

[38] (denominada também como escala KD), *OONS* [39], *Roseman* [40] e *Eisenberg* [41] que denotamos por *Eisenberg_ em86*, as quais são amplamente empregadas em simulações computacionais de proteínas.

2.4 Modelo SBM

Para fins de comparação, também estimamos os pesos com outro modelo de simulação minimalista onde uma simples esfera de massa unitária localizada na posição do átomo C_α na cadeia principal dos resíduos representa os resíduos porém com um potencial mais detalhado [42, 43, 44, 45, 46]. Este é um modelo baseado em estrutura, ou seja, a Hamiltoniana que dá a energia de interação da proteína é baseada na geometria de seu estado nativo, de modo que a superfície de energia potencial alcança o mínimo global neste estado de referência. A forma funcional da energia potencial de uma dada estrutura S em relação ao seu estado nativo S_o é definida por

$$\begin{aligned} V(S,S_o) &= \sum_{\text{ligações}} \epsilon_r (r-r_o)^2 + \sum_{\text{ângulos}} \epsilon_\theta (\theta-\theta_o)^2 \\ &+ \sum_{\text{torsões}} \epsilon_\phi \left\{ [1-\cos(\phi-\phi_o)] + \frac{1}{2}[1-\cos(3(\phi-\phi_o))] \right\} \\ &+ \sum_{\text{ligantes}} \epsilon_C \left[5\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - 6\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{10} \right] + \sum_{\text{não ligantes}} \epsilon_{NC} \left(\frac{\sigma_{NC}}{r_{ij}}\right)^{12}, \quad (2.28) \end{aligned}$$

onde $\epsilon_r = 100\epsilon_0$, $\epsilon_\theta = 20\epsilon_0$, $\epsilon_\phi = \epsilon_0$, $\epsilon_C = \epsilon_0$, $\epsilon_{NC} = \epsilon_0$, $\sigma_{NC} = 4.0$, and ϵ_0 é a energia de interação por contato em unidades reduzidas. As constantes r_o, θ_o, ϕ_o, e σ_{ij} são extraídas das coordenadas da estrutura nativa. Na equação (2.28), as esferas conectadas e os ângulos de ligação interagem via termos harmônicos. O terceiro termo representa a rotação diedral na cadeia principal. A interação de esfera dura entre resíduos não ligantes em contato no estado nativo é dado por um potencial 10-12 de Lennard-Jonnes. Todos os pares de resíduos que não estão em contato na estrutura nativa interagem através de repulsão não específica. Mapas de contatos nativos das distâncias σ_{ij} foram determinadas pelo programa *Contact of Structural Units* (CSU) [47].

2.5 ST-WHAM

Duas importantes técnicas de análise de dados surgiram com os métodos de reponderação (*reweighting*) e de histogramas múltiplos. Em particular, o método conhecido pela sigla WHAM [48] combina dados de energia armazenados em histogramas $H_\alpha(E)$, $\alpha = 1, \cdots, M$, obtidos de M simulações independentes nas temperaturas T_α e fornece a correspondente variação de energia livre associada à mudança de fase termodinâmica. Mais recentemente, o método chamado ST-WHAM (*statistical temperature weighted histogram analysis method*) [49], foi desenvolvido de forma que as equações WHAM possam ser resolvidas evitando-se o processo numérico iterativo que caracteriza a resolução destas equações. O método ST-WHAM fornece uma estimativa numérica para a curva calórica $\beta(E)$, a qual estão relacionada com a entropia,

$$\beta(E) = \frac{1}{\sum_\gamma H_\gamma(E)} \sum_\alpha H_\alpha(E) \left(\frac{d \ln H_\alpha(E)}{dE} - \frac{d \ln \omega_\alpha(E)}{dE} \right) , \tag{2.29}$$

onde $\omega_\alpha(E) = e^{-\beta_\alpha E}$.

2.6 Proposta de novo modelo (wGNM)

Nesta abordagem, consideramos que o peso da interação entre os resíduos i e j é proporcional ao tempo que eles permanecem conectados quando a proteína sofre uma pequena perturbação de desenovelamento a partir do seu estado nativo. Matematicamente, representamos este peso como

$$w_{ij} = \frac{(t_{\text{conectado}})_{ij}}{t_{\text{total}}}, \tag{2.30}$$

onde $t_{\text{conectado}}$ é o tempo que os resíduos permanecem conectados e t_{total} é a duração do desenovelamento, o qual está diretamente relacionado com o número de *sweeps* ou de atualizações na simulação de Monte Carlo.

Mostramos na tabela 2.1 um exemplo de como os pesos w_{ij} são avaliados durante 3 passos de Monte Carlo (tempo de desenovelamento) à uma dada temperatura T. Consideramos que dois resíduos estão conectados se a distância entre eles for menor que o raio de corte $R_c = 1.98$ (unidades do modelo AB) que corresponde a aproximadamente 7.5 Å.

Tabela 2.1: *Exemplo da avaliação dos pesos w_{12}, w_{13} e w_{23} para 3 passos na simulação de Monte Carlo.*

	Tempo	w_{12}	w_{13}	w_{23}
	1	1	1	1
	2	1	0	1
	3	1	0	0
Total		3/3	1/3	2/3

Supondo uma configuração em que no instante $t = 1$ o resíduo 1 está conectado com o resíduo 2 e 3, que por sua vez, o 2 está conectado com o resíduo 3, temos $w_{12} = w_{13} = w_{23} = 1$. No tempo subsequente, por exemplo, devido ao desdobramento da cadeia, apenas o resíduo 3 deixa de estar conectado com o resíduo 1, logo $w_{13} = 0$ e $w_{12} = w_{23} = 1$. No passo seguinte somente prevalece o contato do resíduo 1 com o 2, resultando em $w_{12} = 1$, $w_{13} = w_{23} = 0$. Portanto, podemos dizer que a constante de força da mola que conecta o resíduo 1 ao 3, w_{13}, possui uma intensidade 3 vezes menor que w_{12} e duas vezes menor que w_{12}.

Tendo-se os valores dos pesos w_{ij}, a correlação entre os fatores-B teóricos e experimentais é calculada de modo análogo ao GNM pela equação:

$$B_i^{\mathrm{wGNM}} = \frac{8\pi^2 k_B T}{\gamma} \sum_{k=2}^{N} \lambda_k^{-1} \mathbf{u}_{ik}^2, \tag{2.31}$$

onde λ_k são os autovalores da matrix (w_{ij}) e $\mathbf{u}_{ik}$ os seus autovetores.

Para mensurar o quanto a proteína se desenovelou em relação à sua estrutura nativa, nós utilizamos a quantidade *fração de contatos nativos que permanecem conectados*, a qual nomeamos apenas como fração, definida pela relação

$$\text{fração} = \frac{\text{número de contatos nativos final}}{\text{número de contatos nativos inicial}}. \tag{2.32}$$

Capítulo 3

Resultados e discussão

Este capítulo abarca 3 seções. Na primeira seção, apresentamos os resultados do estudo de caráter exploratório com objetivo de estabelecer os parâmetros adequados do nosso modelo. Na segunda seção, apresentamos o resultado do estudo mais elaborado para validar nosso modelo. Para essa finalidade, realizamos estudos comparativos do wGNM com os modelos GNM, pfGNM e WCN da literatura para um conjunto de proteínas de raios-X com alta resolução. Na terceira seção, aplicamos o nosso modelo para estruturas de NMR como outra forma de avaliarmos o seu desempenho na predição dos deslocamentos quadráticos médio dos resíduos.

3.1 Estudo exploratório dos parâmetros para o nosso modelo

Inicialmente, para investigar os parâmetros adequados do modelo que propomos na seção 2.8, assim como a temperatura adequada na qual a proteína se desenovela, utilizamos como objeto de estudo as proteínas mostradas na tabela 3.1.

Bachmann e colaboradores [37] mostraram que, para heteropolímeros de 20 monômeros, a temperatura critica T_c do modelo AB encontra-se entre 0.1 e 1. Tendo em vista esta informação, investigamos o comportamento da fração e da correlação em função do tempo de desenovelamento para este intervalo de temperatura. Para cada temperatura, nós calculamos a fração, de acordo com a equação (2.32), e a correlação ρ_{wGNM}, entre os fatores-B teóricos e experimentais, obtida da matriz ponderada.

Na figura 3.1, os gráficos (a), (c) e (e), mostram o comportamento da fração em função do tempo de desenovelamento para a proteína 1CNR para 4 escalas de hidrofobi-

Tabela 3.1: *Proteínas utilizadas no nosso estudo inicial para a escolha dos parâmetros do modelo wGNM. Na coluna PDB temos o código de identificação da proteína, N representa o número de resíduos e ρ_{GNM} a correlação entre os fatores-B teóricos e experimentais calculada pelo GNM.*

PDB	N	ρ_{GNM}
1CNR	46	0.64
1EYH	144	0.89
1OAL	151	0.54
1LYX	246	0.39
1QAZ	351	0.70

cidade. Os gráficos (b), (d) e (f) exibem a dependência da correlação ρ_{wGNM} em função do tempo de desenovelamento. Comparando-se os gráficos (a), (c) e (e), observa-se que a fração assume valores menores à medida que a temperatura aumenta. Isto é plausível fisicamente porque a proteína recebe mais energia para tentar vencer as interações entre os seus contatos nativos e, consequentemente, adquire conformações distantes da nativa. Dentre esses valores de temperatura, supomos que $T = 1$ corresponda a um valor adequado para o nosso modelo. Como para $T = 1$ a curva da fração decresce de forma mais acentuada, conjecturamos que, para este valor de temperatura, a proteína 1CNR é levada mais rapidamente a conformações afastadas da nativa. Para verificar esta hipótese, efetuamos o cálculo da temperatura de transição do modelo AB para esta proteína. Em concordância com a nossa hipótese, verificou-se que T_c está em torno 0.67. Portanto, $T = 1$ pode ser considerado um valor de temperatura adequado para a proteína 1CNR se desenovelar. Além disso, no gráfico (f), para $Nsweeps$ entre 0 e 140, ocorre um colapso entre as curvas da correlação ρ_{wGNM}(nsweeps) para as escalas *Eisenberg_ em86*, *KD* e *OONS*. Assim, para este valor de temperatura, a influência da escala de hidrofobicidade no valor da correlação é fortemente minimizada.

O gráfico (f) mostra a evolução temporal da correlação ρ_{wGNM}. No tempo inicial, $Nsweeps = 0$, para as 4 escalas de hidrofobicidade, $\rho_{\mathrm{wGNM}} = \rho_{\mathrm{GNM}}$. Imediatamente após, o valor da correlação ρ_{wGNM} aumenta continuamente até atingir um máximo, a partir do qual ele começa a diminuir com o tempo. Nota-se que o valor da correlação ρ_{wGNM}

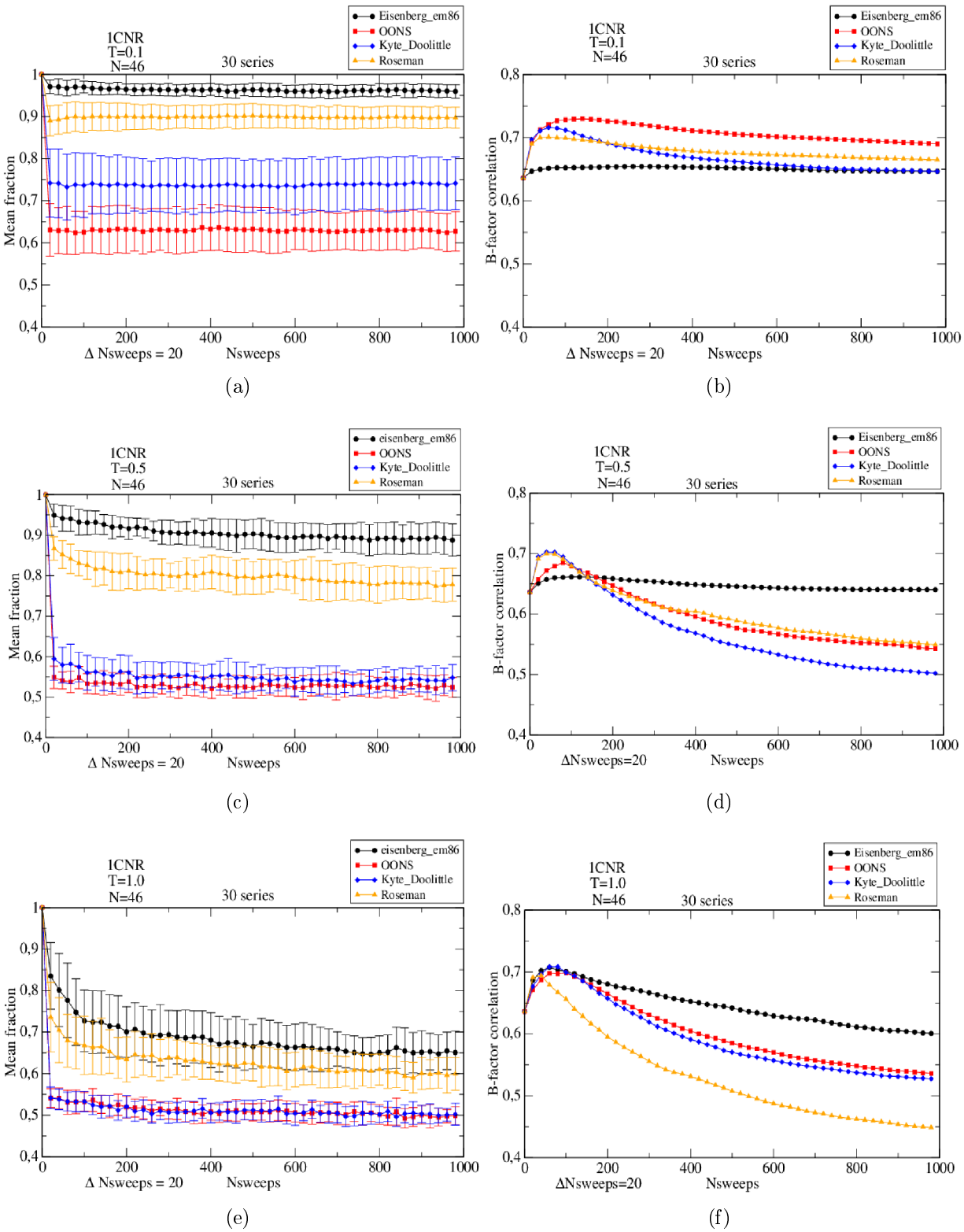


Figura 3.1: *Gráficos da fração média sobre 30 séries, painel à esquerda, e de ρ, painel à direita, em função do número de sweeps para a proteína 1CNR para valores de $T = 0.1$ em (a) e (b), $T = 0.5$ em (c) e (d) e $T = 1.0$ em (d) e (e). Tanto os valores da fração quanto os valores de ρ foram coletados em intervalos de 20 sweeps.*

diminui quando a fração tende a permanecer constante, ou seja, quando a proteína atinge um estado de equilíbrio longe do que seriam estados perturbados em torno do estado nativo. A partir deste estado, estaríamos avaliando os pesos w_{ij} de forma inadequada por meio da equação (3.3).

O comportamento da curva $\rho_{\mathrm{wGNM}}(\mathrm{nsweeps})$ para proteína 1CNR é mantido para as proteínas 1EYH, 1OAL, 1LYX e 1QAZ, conforme ilustrado nas figuras 3.2, 3.3, 3.4 e 3.5. Sugerindo que tal comportamento é fortemente dependente de T.

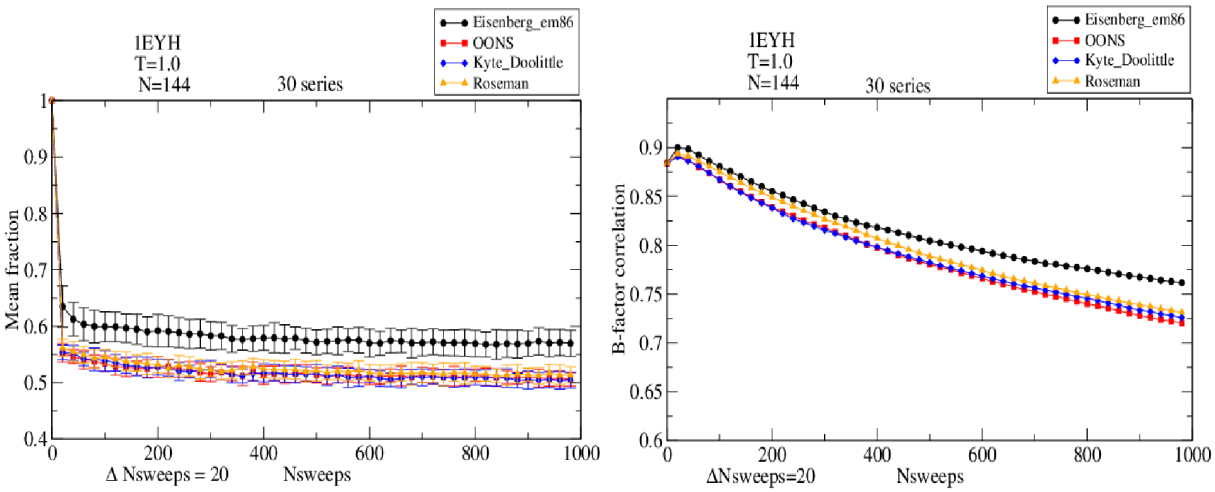
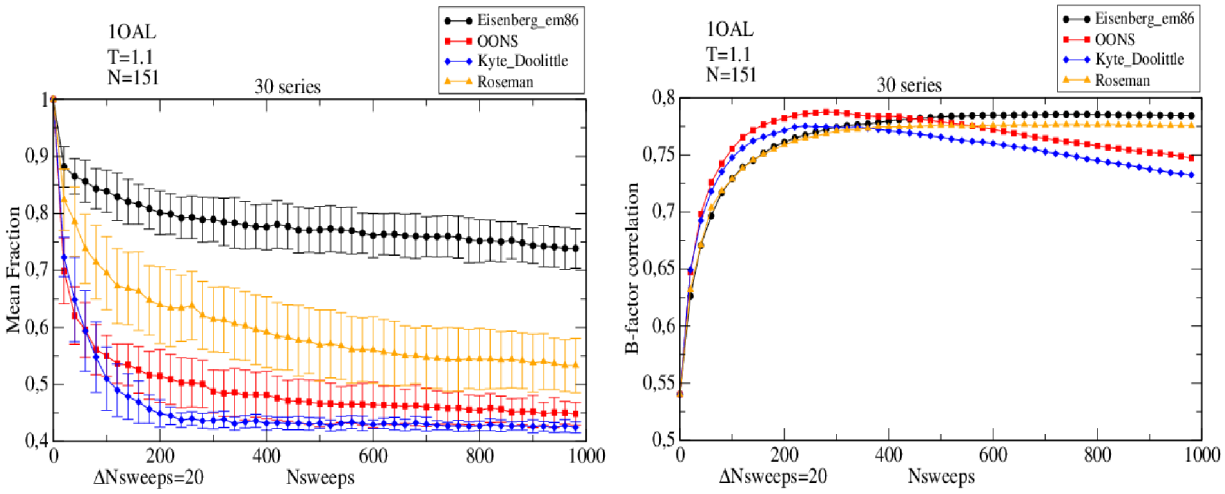
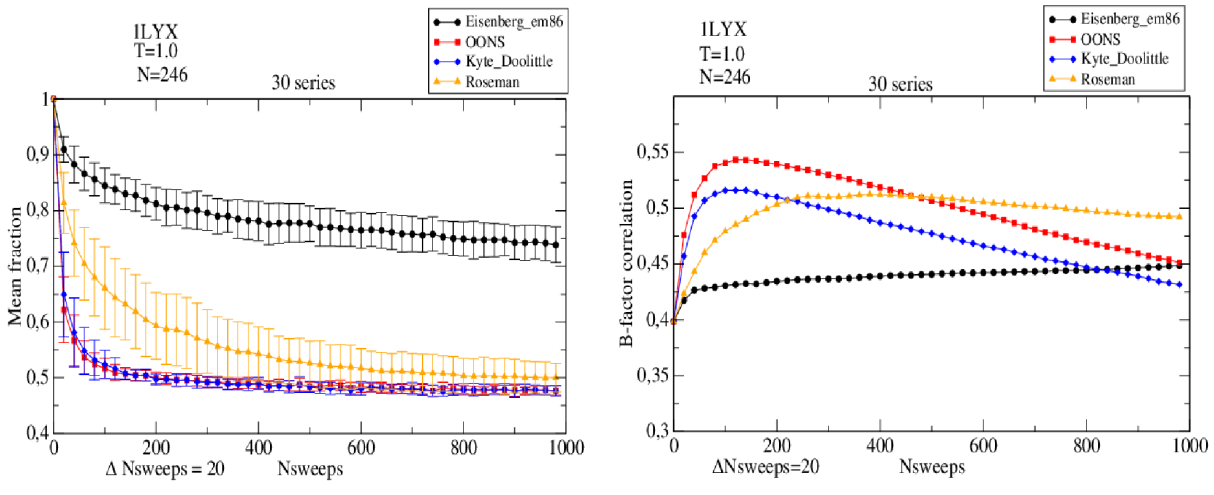


Figura 3.2: *(a) Comportamento da fração média sobre 30 séries durante 1000 sweeps para a proteína 1EYH de tamanho $N = 144$ à temperatura $T = 1$, para as escalas de hidrofobicidade Eisenberg_ em86 (●), OONS (■), Kyte_ Doolittle (◆) e Roseman (▲).*



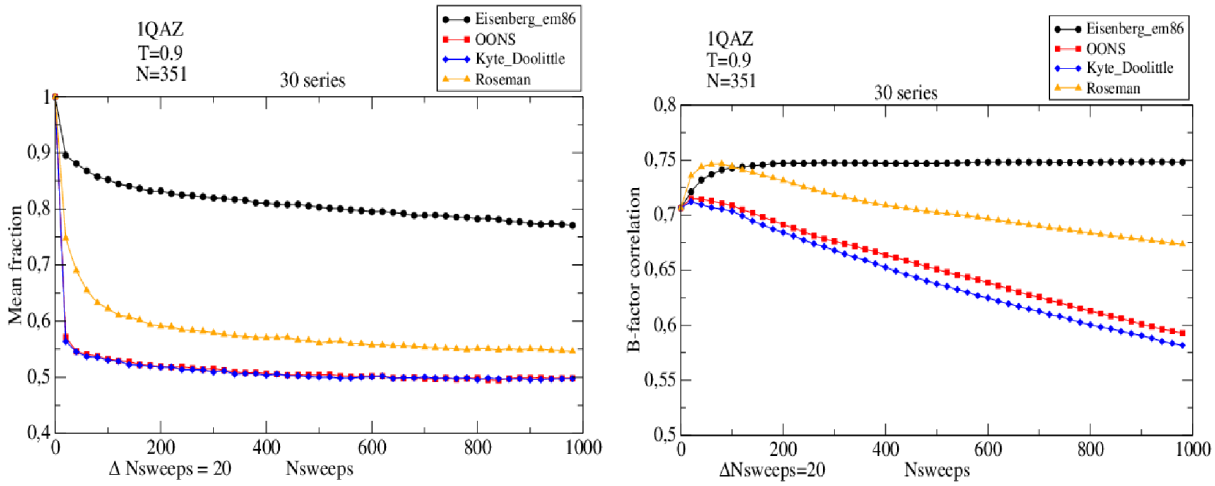
(a) (b)

Figura 3.3: *(a) Comportamento da fração média sobre 30 séries durante 1000 sweeps para a proteína 1OAL de tamanho $N = 151$ à temperatura $T = 1.1$, para as escalas de hidrofobicidade Eisenberg_ em86 (●), OONS (■), Kyte_ Doolittle (◆) e Roseman (▲).*



(a) (b)

Figura 3.4: *(a) Comportamento da fração média sobre 30 séries durante 1000 sweeps para a proteína 1LYX de tamanho $N = 246$ à temperatura $T = 1$, para as escalas de hidrofobicidade Eisenberg_ em86 (●), OONS (■), Kyte_ Doolittle (◆) e Roseman (▲).*



(a) (b)

Figura 3.5: *(a) Comportamento da fração média sobre 30 séries durante 1000 sweeps para a proteína 1QAZ de tamanho $N = 351$ à temperatura $T = 0.9$, para as escalas de hidrofobicidade Eisenberg_ em86 (●), OONS (■), Kyte_ Doolittle (◆) e Roseman (▲).*

Aplicamos o wGNM para 818 proteínas com tamanho entre 40 e 500 resíduos. Para cada proteína, estabelecemos T=1 no modelo AB e utilizamos a escala de hidrofobicidade *KD*. Para cada série na simulação de Monte Carlo, coletamos a matriz ponderada (w_{ij}) quando o valor da fração atinge 70%. Após obtidas as 30 séries, calculamos a matriz ponderada média e, a partir desta, os valores da correlação ρ_{wGNM}. A tabela 3.2 mostra os valores médios das correlações ρ_{GNM} e ρ_{wGNM}. Como estes valores médios e os seus respectivos desvios são aproximadamente iguais, temos que para as 818 proteínas ambos os modelos são equivalentes. A figura 3.6(a) apresenta os histogramas das correlações ρ_{GNM} e ρ_{wGNM}. A porcentagem de proteínas que obtiveram valores de correlação acima de 0.8 utilizando o wGNM é aproximadamente 4%, que é melhor que 2.5% encontrado utilizando o GNM. Entretanto, o desempenho do wGNM é pior que o do GNM, pois o número de casos fracamente correlacionados com os fatores-B experimentais é maior que no GNM. Presumimos que um dos motivos pelos quais o wGNM forneceu resultados inferiores ao GNM pode ser devido aos parâmetros utilizados. Para tentar otimizar estes parâmetros, efetuamos testes, por exemplo, variando o valor da fração, para uma amostra de 99 proteínas das 818. A tabela 3.3 mostra os valores da correlação ρ_{wGNM} coletados quando a fração atinge 70%, 80% e 90% para $T = 1$.

Tabela 3.2: *Correlação média sobre 818 proteínas dos fatores-B teóricos e experimentais calculados pelos modelos GNM e wGNM à $T = 1$. No wGNM, coletamos o valor da correlação no instante em que a fração atinge 70%.*

	GNM	wGNM (70%)
818 proteínas	0.55 ± 0.18	0.53 ± 0.2

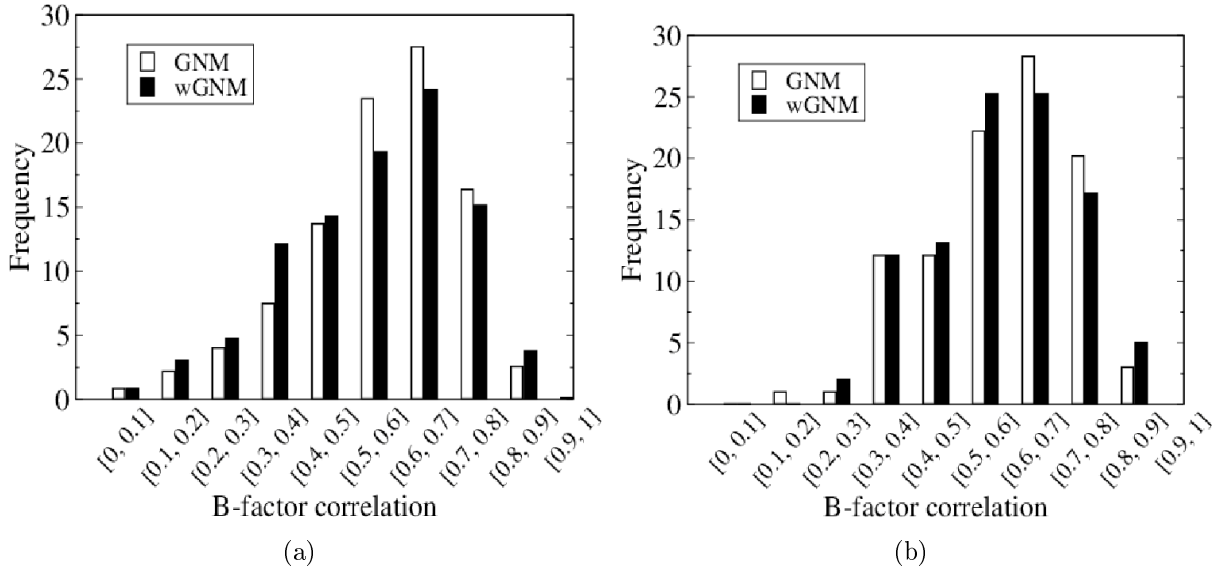


Figura 3.6: *Histogramas da correlação entre os fatores-B teóricos e experimentais calculados pelos os modelos GNM e wGNM à T=1 para 818 proteínas (a) e 99 proteínas (b) quando a fração atinge 70%.*

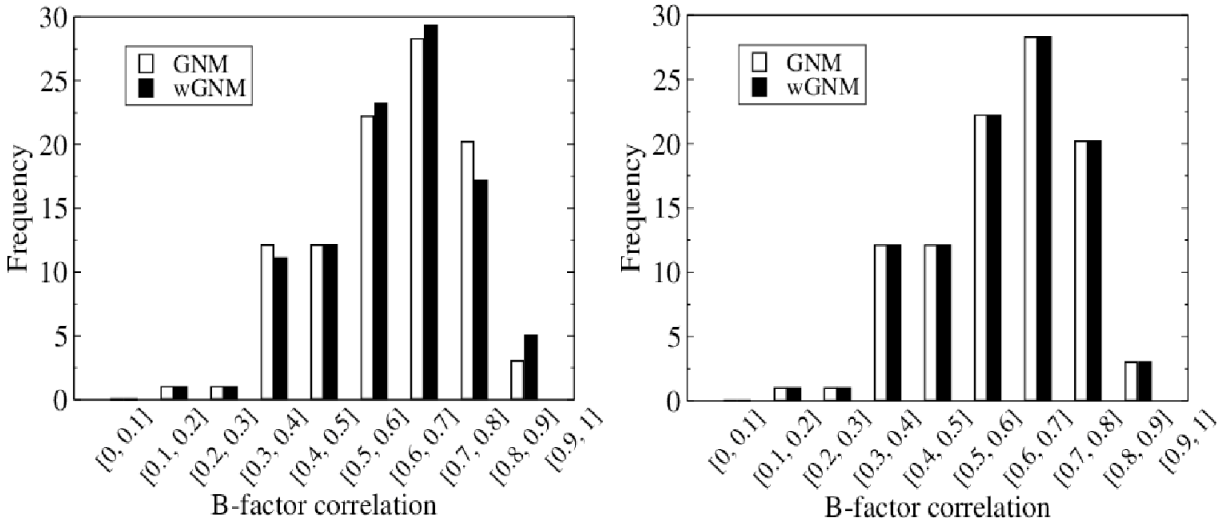
Dentre os valores da fraçao, 80% fornece um valor médio da correlação ρ_{wGNM} ligeiramente melhor que o valor médio da correlação ρ_{GNM}, entretanto, quando se compara as distribuições destas correlações, figura 3.7(a) e figura 3.6(b), não se observa nenhuma melhora significativa. Para o valor da fração igual a 90%, ambos os histogramas são idênticos. Como o tempo de desenovelamento é relativamente curto, isso faz com que o valor da correlação do GNM seja infimamente modificado.

Efetuamos testes semelhantes para $T = 1.2$ para as 99 proteínas. Obtivemos os valores médios $\bar{\rho}_{\mathrm{GNM}} = 0.59 \pm 0.14$ e $\bar{\rho}_{\mathrm{GNM}} = 0.58 \pm 0.15$ quando a fração atinge, respectivamente, 80% e 90%. Mesmo variando a fração e a temperatura, o wGNM não se mostrou significativamente superior ao GNM.

Analisando as séries temporais das energias produzidas pelo modelo AB, verifica-

Tabela 3.3: *Correlação média sobre 99 proteínas dos fatores-B teóricos e experimentais calculados pelos modelos GNM e wGNM à T=1.*

		wGNM		
	GNM	70%	80%	90%
99 proteínas	0.58 ± 0.15	0.58 ± 0.15	0.59 ± 0.14	0.58 ± 0.15



(a) (b)

Figura 3.7: *Histogramas da correlação entre os fatores-B teóricos e experimentais calculados pelos os modelos GNM e wGNM à T=1 para 99 proteínas quando a fração atinge 80% (a) e 90% (b).*

mos que até mesmo para valores de temperaturas em torno de T_c, a proteína pode ser levada à conformações mais compactadas que a nativa ao invés de se desenovelar. Isso porque não conhecemos, para cada proteína, a temperatura do potencial AB correspondente ao seu estado nativo. Para tentar contornar este problema e validar a nossa metodologia, utilizaremos o potencial proposto por Gō [42], para desenovelar a proteína. Escolhemos este potencial porque seus parâmetros são ajustados de modo que a sua energia mínima corresponda ao estado nativo experimental do PDB.

Apresentamos uma nova abordagem do modelo de rede elástica gaussiana, a qual considera um potencial mais sofisticado que o harmônico e o caráter hidrofóbico e hidrofílico dos resíduos. Mostramos que, para cinco proteínas, o nosso modelo fornece resultados significativamente melhores que o GNM. Entretanto, quando efetuamos testes para um

conjunto maior de proteínas, os modelos GNM e wGNM são equivalentes em média. Por outro lado, para uma amostra de 99 proteínas, obtivemos resultados ligeiramente melhores que o GNM para $T = 1$ quando a fração atinge 80%.

Verificamos que o modelo AB não descreve satisfatoriamente a evolução do desenovelamento da proteína a partir do estado nativo experimental. Para tentar contornar este problema, utilizaremos o modelo proposto por Gō, o qual parte da estrutura nativa do PDB como conformação de menor energia. Portanto, acreditamos que este modelo descreverá melhor o desenovelamento e isso refletirá em uma melhor concordância com os dados experimentais no wGNM.

3.2 Estudo comparativo entre os modelos wGNM, GNM, pfGNM e WCN

Tabela 3.4: *Conjunto de proteínas selecionadas para efetuar as simulações AB e SBM. As temperaturas de enovelamento (T_f) são dadas para cada proteína em função dos modelos AB e SBM.*

Proteína	código PDB	Resíduos	Resolução (Å)	modelo AB (T_f)	SBM (T_f)
Crambina	1CNR	46	1.05	0.65	0.84
CI-2	1YPA	64	2.0	0.70	1.04
Ubiquitina	1UBQ	76	1.8	0.50	1.10
RNase SA	1LNI	96	1.0	0.53	1.08
Citocromo c	1HRC	104	1.9	0.51	1.04
Azurina	1E65	128	1.85	0.53	1.092
Ciclofilina	1LOP	164	1.8	0.63	1.183

Apresentamos aqui os resultados do trabalho que publicamos recentemente [50] no qual comparamos o desempenho do wGNM com os modelos de rede elástica tradicionais GNM, pfGNM e WCN para um conjunto de sete proteínas do PDB de alta resolução (Tabela 3.4). Esta seleção permite assim explorar como os fatores-B preditos se comportam em função dos pesos w_{ij} entre pares de átomos de C_α. Temos como objetivo obter uma forma eficiente de atribuir estes pesos. Para isso, analisamos as trajetórias de desenovelamento parcial em diferentes temperaturas, incluindo simulações na temperatura de enovelamento.

Trajetórias de desenovelamento induzido pela temperatura são analisadas como função do tempo de simulação calculado em três temperaturas para as proteínas mostradas na tabela 3.4. Esta tabela também apresenta as temperaturas de enovelamento T_f obtidas com as simulações com modelos AB e SBM. Avaliamos a fração de contatos ligantes e não ligantes que permanecem conectados dentro de um raio $R_c = 7.5$ Å em função do tempo nas temperaturas $0.2T_f, 0.4T_f$ e T_f. As médias são calculadas com 50 trajetórias para

ambos modelos simulados iniciando da conformação nativa. Medidas são obtidas a cada passo de Monte Carlo no modelo AB e a cada 1 ps no modelo SBM.

3.2.1 Temperatura de transição

Estimamos a temperatura de transição de enovelamento T_f com o modelo AB usando as simulações de Monte Carlo (MC) envolvendo o método de troca entre réplicas (*replica exchange method*) [51]. Para cada proteína, consideramos $M = 8$ réplicas, escolhendo assim 8 temperaturas determinadas por meio de uma progressão aritmética em termos do inverso da temperatura, $\beta_{M-\alpha} = \beta_M + \alpha\delta$, com $\alpha = 0, \ldots, M-1$, e $\delta = 0.2857$, onde $\beta_M = 1/T_M$. Após um estudo exploratório inicial para tentar localizar os pontos de transição, ajustamos a maior temperatura T_M para 1 para as estruturas proteicas 1YPA e 1CNR, e $T_M = 0.7778$ para 1UBQ, 1HRC e 1LNI. Utilizamos a estatística de 3.1×10^7 passos de MC para cada temperatura, descartando 10^6 passos de MC para a termalização. As tentativas de trocas de réplicas adjacentes ocorreram a cada 2×10^3 passos de MC.

A partir dos histogramas, calculamos a densidade de estados $n(E)$ usando o método ST-WHAM-MUCA [52], o qual é baseado no ST-WHAM, método de análise ponderando histogramas obtidos em diferentes temperaturas [49]. A partir desta densidade de estados, obtemos a distribuição de probabilidade canônica $p(E) \propto n(E)\exp(-\beta E)$, para encontrar a temperatura T_f na qual aparace dois picos com a mesma altura.

Obtemos arquivos de entradas de modelos baseados em estruturas usando o *Structure-based MOdels* in Gromacs (SMOG) *webtool* [46]. Efetuamos as simulações de dinâmica molecular com o pacote GROMACS [53]. As proteínas são inicializadas em um configuração aleatória e simuladas durante 5×10^8 ps com passos de tempo de 0.4 fs e equilibração depois de 1×10^7 passos. Simulamos uma série de evoluções temporais com temperaturas constantes e usamos o WHAM [48, 54] para analisar os dados obtidos. Aqui, a temperatura de enovelamento T_f foi determinada pelo pico na curva do calor específico.

Tabela 3.5: *Coeficientes de correlação entre os fatores-B experimentais e os fatores-B preditos pelos modelos de rede elástica.*

código PDB	GNM	pfGNM	WCN	wGNM (Modelo AB)	wGNM (SBM)
1CNR	0.64	0.55	0.56	0.75	0.67
1YPA	0.83	0.69	0.70	0.88	0.85
1UBQ	0.82	0.84	0.83	0.85	0.84
1LNI	0.49	0.42	0.43	0.55	0.52
1HRC	0.12	0.29	0.29	0.52	0.34
1E65	0.57	0.57	0.57	0.60	0.57
1LOP	0.66	0.68	0.68	0.83	0.71

3.2.2 Citocromo c

Iniciamos a análise com a estrutura tridimensional do citocromo c (do coração de cavalo) com código PDB 1HRC e tendo resolução de 1.9 Å. O GNM fornece o menor coeficiente de correlação com os fatores-B experimentais para esta proteína ($\rho_{\mathrm{GNM}} = 0.12$, Tabela 3.5).

Esta proteína apresenta um domínio com 3 hélices maiores e 2 hélices menores, enoveladas em uma forma de bolsa que acomoda o grupo heme. O grupo heme do citocromo c de cavalo está altamente enterrado dentro da proteína: apenas 7.5% da superfície heme interage com as moléculas do solvente [55]. Apesar da sua estrutura de domínio relativamente simples, o citocromo c de cavalo precisa exibir conformações flexíveis para posicionar e manter o grupo heme. A figura 3.8 apresenta o comportamento da fração média de contatos nativos Q que permanecem conectados em função do tempo.

Um contato nativo permanece conectado se sua distância de separação está dentro do alcance do corte. As barras de erro foram obtidas sobre 50 séries temporais. Estas evoluções temporais foram usadas para obter um valor médio para os pesos w_{ij} em cada passo de tempo. Simulações a baixas temperaturas rapidamente alcançam as conformações de equilibrio caracterizadas por estados parcialmente enovelados, então elas não revelam a

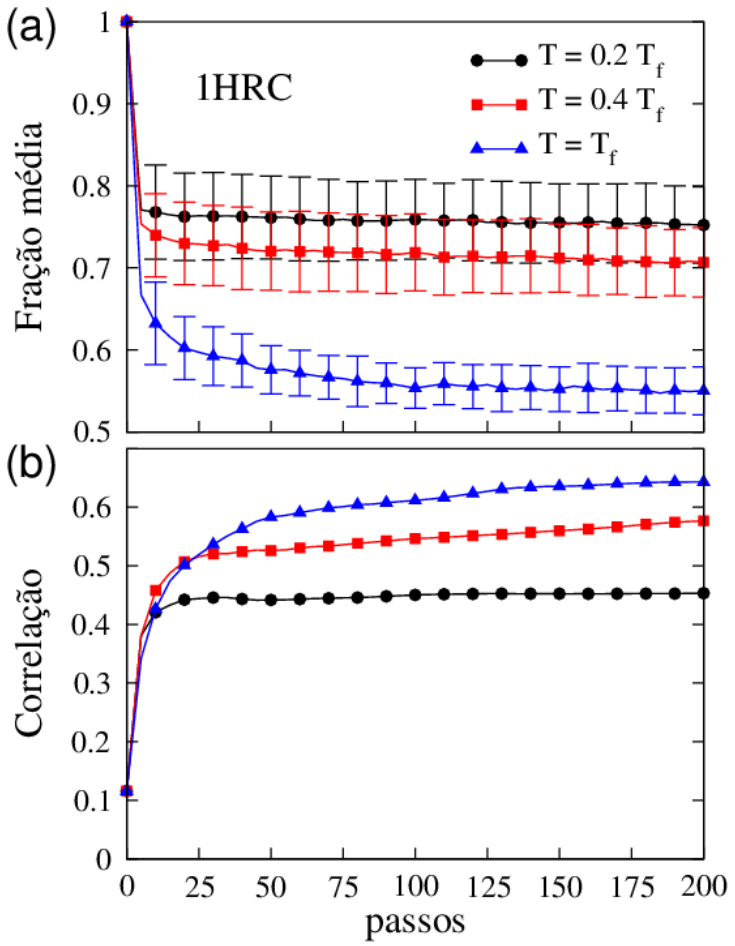


Figura 3.8: *Fração média de contatos nativos Q e coeficiente de correlação ρ em função dos passos de MC para proteína 1HRC. O modelo AB governa a dinâmica do desenovelamento.*

importância dos *links* na rede.

É preferível realizar esta avaliação na temperatura de enovelamento T_f (*folding temperature*), onde podemos observar as transições caracterizadas por um decaimento na fração média quando o sistema se aproxima do seu estado de equilíbrio nesta temperatura. Temperaturas mais baixas não dão um número apreciável de *links* quebrados e somente acarreta pequenos desvios da conformação nativa, a qual não altera significativamente o coeficiente de correlação ρ quando comparado com aquelas observados em T_f, ver figura 3.8(b).

A figura 3.8(b) mostra como o coeficiente de correlação se comporta em função do tempo para diferentes temperaturas de desenovelamento. A introdução dos pesos w_{ij} per-

mite uma melhor avalição dos fatores-B. Como regra geral, a avaliação dos pesos relativos deve continuar até que a fração média de contatos entre átomos de C_α tenha atingido um valor estável. Assim, a simulação deve prosseguir até que as ligações facilmente quebráveis tenham sido eliminadas da rede.

A figura 3.9 compara os fatores-B experimentais e os fatores-B teóricos que obtivemos pelo GNM e wGNM para esta proteína de 104 resíduos. Os fatores-B teóricos do wGNM foram determinados utilizando os pesos avaliados a partir de séries de desenovelamento parcial quando a fração média atingiu 0.6. Esta é uma regra razoável que se aplica a outras proteínas. Para tal desenovelamento parcial, o wGNM produz o coeficiente de correlação $\rho = 0.52$, correspondendo a apenas 25 passos de MC para o modelo AB.

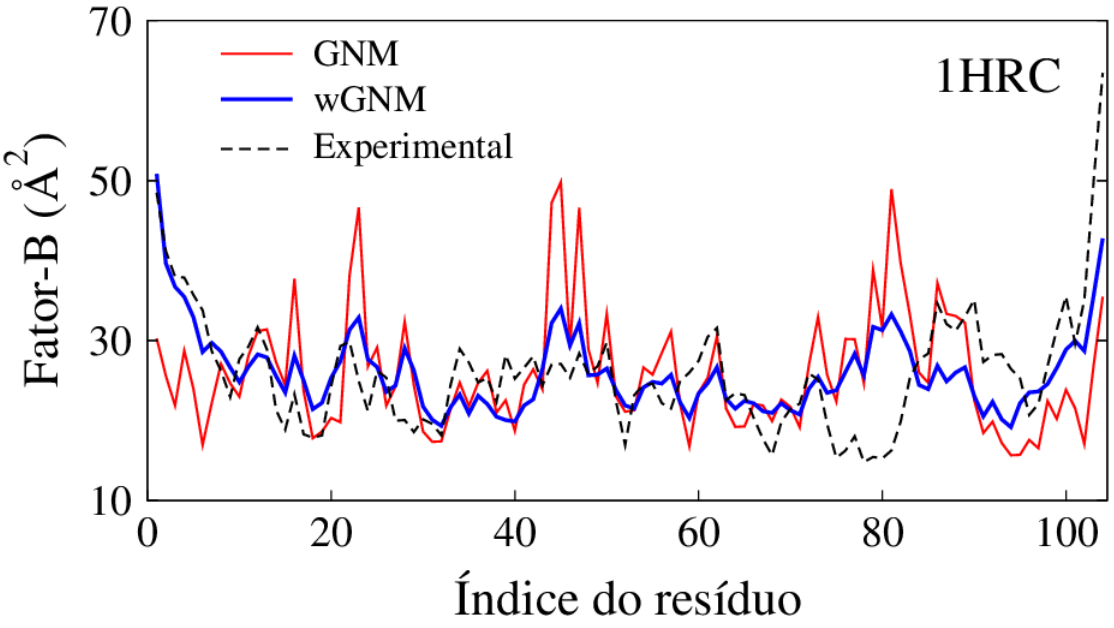


Figura 3.9: *Fatores-B teóricos e experimentais para proteína 1HRC. O perfil dos fatores-B calculado pelo wGNM foi obtido no tempo correpondente a 25 passos de MC.*

Os processos de desenovelamento induzidos pela temperatura (*temperature-induced unfolding processes*) foram analisados em um contexto geral usando um simples modelo de rede [56], mas características gerais podem ser obtidas. Simulações em temperaturas de desnaturação suaves (*mild denaturating temperatures*), isto é, logo acima da temperatura de transição, não desnatura a proteína imediatamente, mas produz conformações em torno do estado nativo ou ainda em estados parcialmente enovelados. Portanto, mesmo nestas temperaturas as conformações ainda forneceriam contatos nativos para uma avaliação

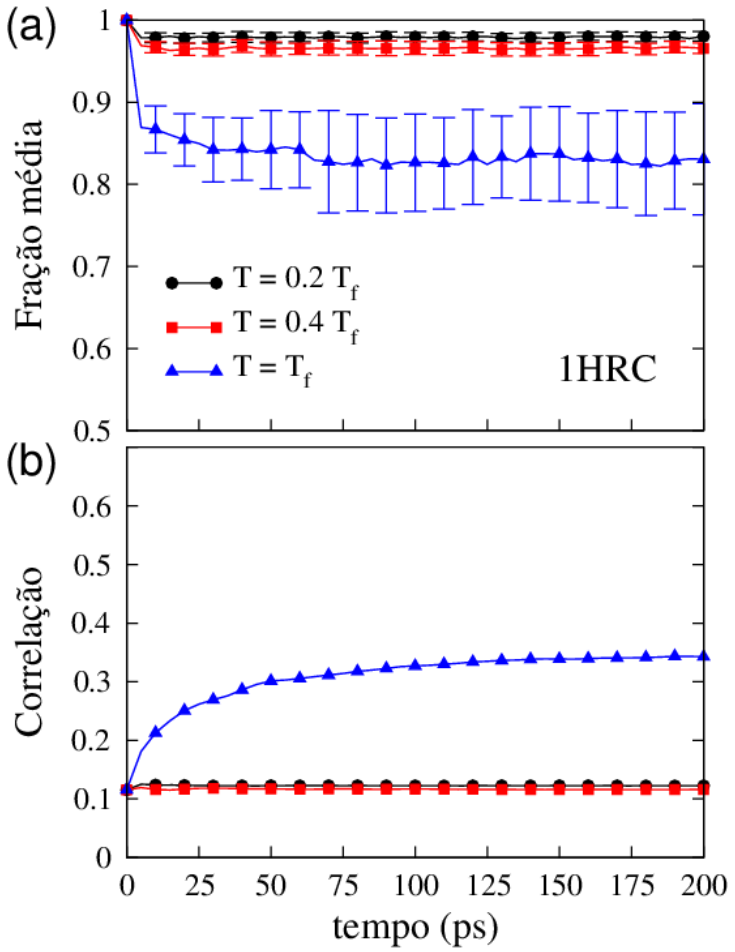


Figura 3.10: *Fração média de contatos nativos Q e o coeficiente de correlação ρ em função do tempo de simulação em picosegundos para proteína 1HRC. SBM governa a dinâmica de desenovelamento nas temperaturas $0.2T_f, 0.4T_f$, e T_f.*

adequada dos pesos w_{ij}. Depois de um acréscimo adicional na temperatura, o processo de desenovelamento deveria diminuir rapidamente a quantidade de ligações que formam a rede. Consequentemente, as trajetórias não iriam exibir a importância relativa das ligações nativas, produzindo um viés significativo na avaliação do pesos w_{ij}. As figuras 3.10(a) e 3.10(b) mostram os resultados para a fração média e o coeficiente de correlação, respectivamente, como uma função do tempo de simulação usando a dinâmica molecular para o SBM.

Baixas temperaturas não induzem um movimento apreciável em direção às novas conformações e não melhoram o coeficiente de correlação obtido pelo GNM para a rede construída a partir da estrutura nativa. Em baixas temperaturas, temos somente estru-

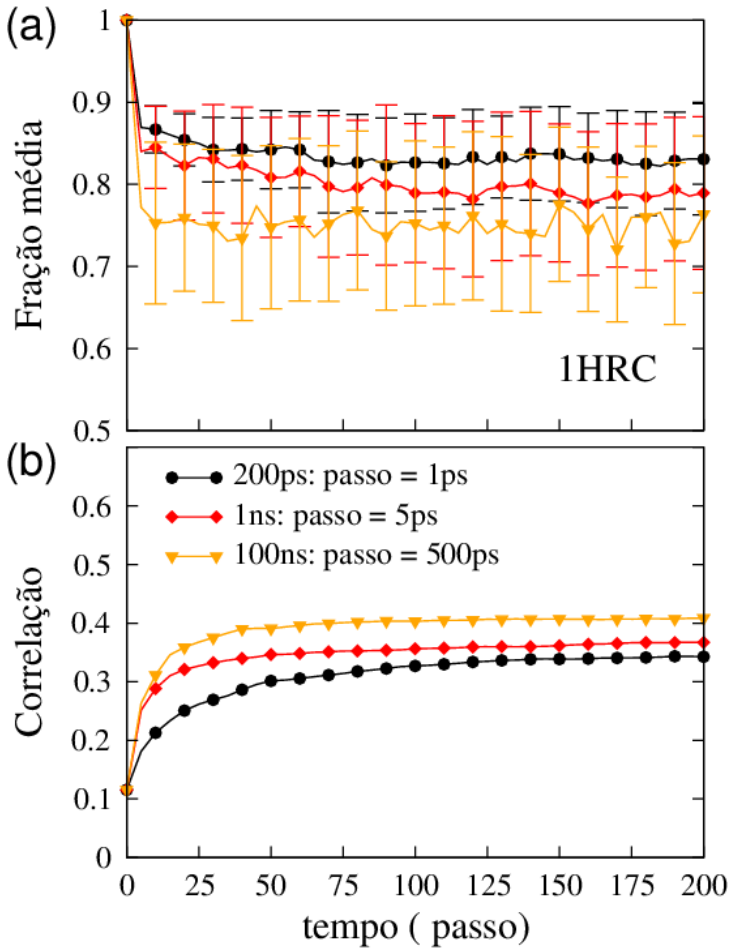


Figura 3.11: *Fração média de contatos nativos Q e coeficiente de correlação ρ em função de diferentes tempos de simulações para a proteína 1HRC com dinâmica molecular e SBM executada em $T = T_f$.*

turas que flutuam em torno do estado nativo conforme evidenciado pela fração média de contatos que sobrevivem em função do tempo. As simulações nestas temperaturas baixas oferecem pouca ou nenhuma melhora nos cálculos em comparação com a estimativa de GNM. Aqui, o algoritmo de dinâmica molecular desenovela lentamente o sistema, em comparação com a simulação de MC com o modelo AB. As simulações em T_f para 200 ps ainda mantem a maioria dos contatos que formam a rede na conformação nativa. Mesmo para essas trajetórias curtas, obtemos $\rho_{\mathrm{wGNM}} \sim 0.34$, um valor ainda muito melhor do que a estimativa $\rho_{\mathrm{GNM}} = 0.12$ ou as estimativas do pfGNM e WCN. A figura 3.11 compara as frações médias e os correspodentes coeficientes de correlação com o valor experimental para longas trajetórias obtidas de dinâmica molecular em T_f. Aqui, a simulação mais

longa levou 100 ns e forneceu um coeficiente de correlação melhor com o valor experimental. Para uma simulação em $T = T_f$ a fração de contatos nativos deveria variar em torno de 0.5. Isto porque, na temperatura de transição, aproximadamente 50% do tempo a proteína está enovelada e nos outros 50 % a proteína está desenovelada, então $Q \approx 0.5$. Entretanto, de acordo com a Figura 3.11(a), $Q > 0.7$ e isto é mostrado para as outras proteínas sob a mesma condição ($T = T_f$). Isso acontece porque essas figuras segue o critério do GNM, que usa um único limite de distância para calcular a ocorrência de contatos nativos, enquanto o SBM tem diferentes limites de distâncias para cada contato de aminoácido, que depende da estrutura nativa. Apesar do SBM ter características mais detalhadas que o modelo AB, estes detalhes finos fornecem uma melhoria limitada no desempenho de um modelo minimalista como o wGNM.

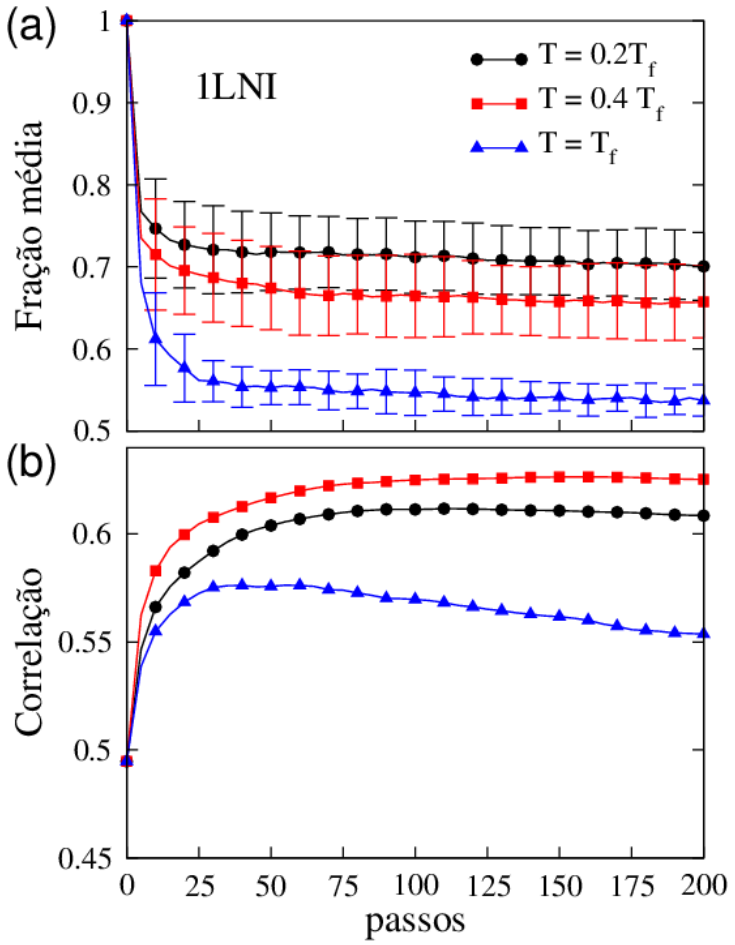


Figura 3.12: *Fração média de contatos nativos Q e o coeficiente de correlação ρ como umma função dos passos de MC para proteína 1LNI. A dinâmica do desenovelamento nas 3 temperaturas é governada pelo modelo AB.*

3.2.3 RNase SA

A proteína RNase SA com código PDB 1LNI apresenta o segundo menor coeficiente de correlação para o modelo GNM, $\rho_{\mathrm{GNM}} = 0.49$ (Tabela 3.5). Esta é uma proteína com dois domínios, possui uma única hélice α e seis fitas β.

A figura 3.12 apresenta a análise de 50 trajetórias de MC com o campo de forças AB e revela um comportamento da fração média semelhante ao da proteína 1HRC. A figura 3.12 ilustra os fatores-B teóricos obtidos com o wGNM, e novamente observamos melhoras nos cálculos do fatores-B. Em contraste a 1HRC, os pesos estimados das simulações à

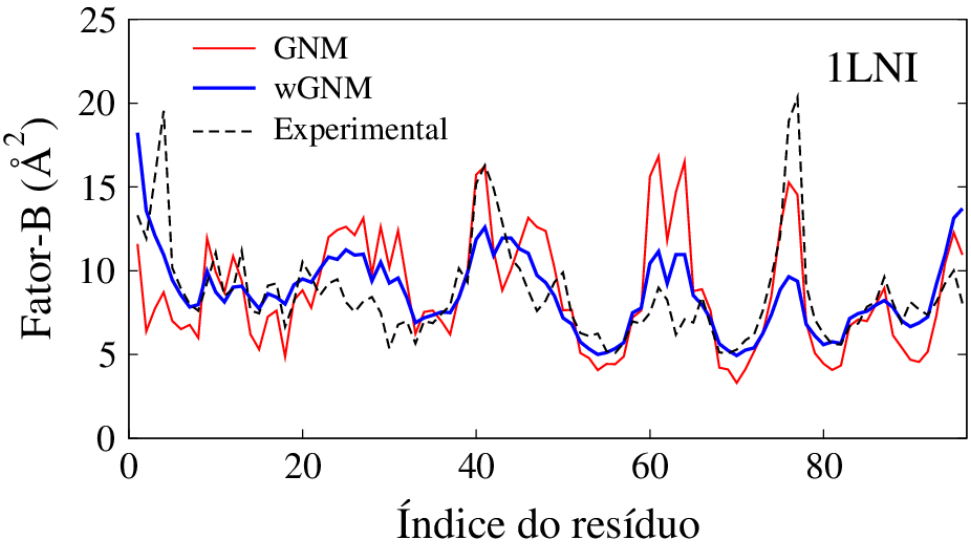


Figura 3.13: *Fatores-B teóricos e experimentais para a proteína 1LNI. O perfil dos fators-B do wGNM foi obtido com 10 passos de MC.*

baixas temperaturas parecem reproduzir melhor os fatores-B. Este sistema rapidamente quebra os *links* frágeis que formam a rede correspondente na estrutura nativa. Agora, em T_c, a correlação dos fatores-B diminui se consideramos a avaliação dos w_{ij} nos estados de equilíbrio. Isto também é verdadeiro para outras proteínas e significa que a importância relativa dos *links* não pode ser medida quando o sistema atinge estes estados.

Novamente, se assumirmos que a avaliação dos pesos w_{ij} deveria parar antes do sistema alcançar o equilíbrio, obtemos o coeficiente de correlação $\rho_{\mathrm{wGNM}} = 0.55$, o qual é melhor que as estimativas do GNM, pfGNM e WCN. Aqui, consideramos que a fração média de contatos nativos conectados é 0.6. A figura 3.13 reproduz os fatores-B experimentais e os fatores teóricos calculados pelo GNM e wGNM para esta proteína com 96 resíduos. O perfil dos fatores-B do wGNM foi calculado com 10 passos de MC.

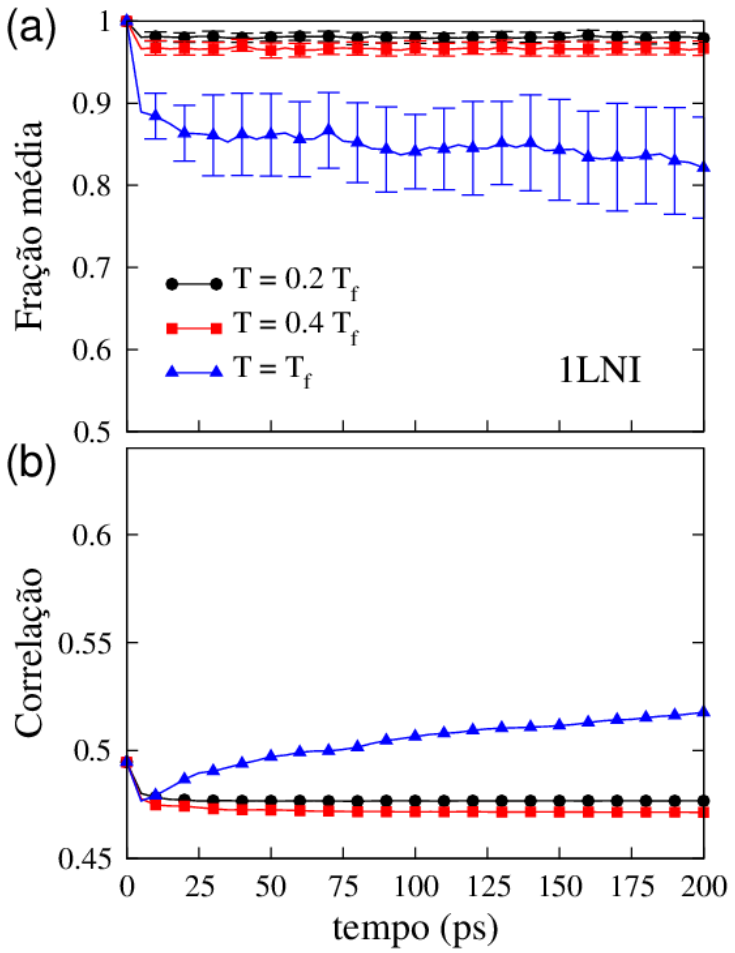


Figura 3.14: *Fração média de contatos Q e o coeficiente de correlação ρ como uma função do tempo de simulação em picosegundo para a proteína 1LNI. A dinâmica de desenovelamento nas 3 temperaturas é governada pelo modelo SBM.*

A figura 3.14 exibe o desempenho comparativo com o SBM. As conformações obtidas em T_f fornecem um aumento nos coeficientes de correlação dos fatores-B em função do tempo das simulações de dinâmica molecular. Ainda usando os dados obtidos de simulação curtas - apenas 200 ps - obtemos $\rho_{\text{wGNM}} = 0.52$ da estimativas de w_{ij} na temperatura T_f. Este resultado também é melhor que os obtidos com os modelos usuais GNM, pfGNM e WCN (ver tabela 3.5).

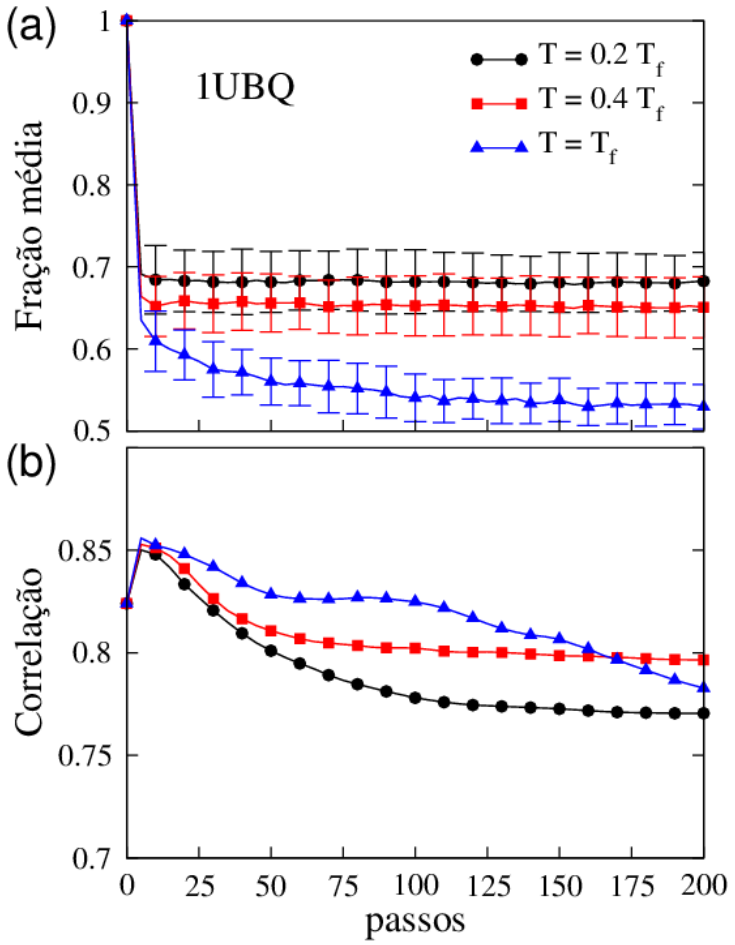


Figura 3.15: *Fração média de contatos Q e o coeficiente de correlação ρ como uma função dos passos de MC para proteína 1UBQ. A dinâmica do desenovelamento na 3 temperaturas é governada pelo o modelo AB.*

3.2.4 Ubiquitina

A ubiquitina tem recebido considerável atenção em estudos sobre o desenovelamento dependente da temperatura [57, 58] e em experimentos de estiramento [59, 60, 61]. A proteína ubiquitina com código PDB 1UBQ é uma estrutura com resolução de 1.8 Å, possui um único domínio e exibe um núcleo hidrofóbico formado pela face de uma fita β e uma hélice α [62]. A figura 3.15 retrata o desenovelamento parcial com o modelo AB. O modelo wGNM prediz muito bem os fatores-B, com um coeficiente de correlação $\rho_{\mathrm{GNM}} = 0.82$. Isto significa que um único valor para todas as constantes de mola já representa a força para a maioria dos *links* que constituem a rede, deixando pouco espaço para melhoras (ver valores obtidos pelo pfGNM e WCN na 3.5). A figura 3.15(b) mostra tam-

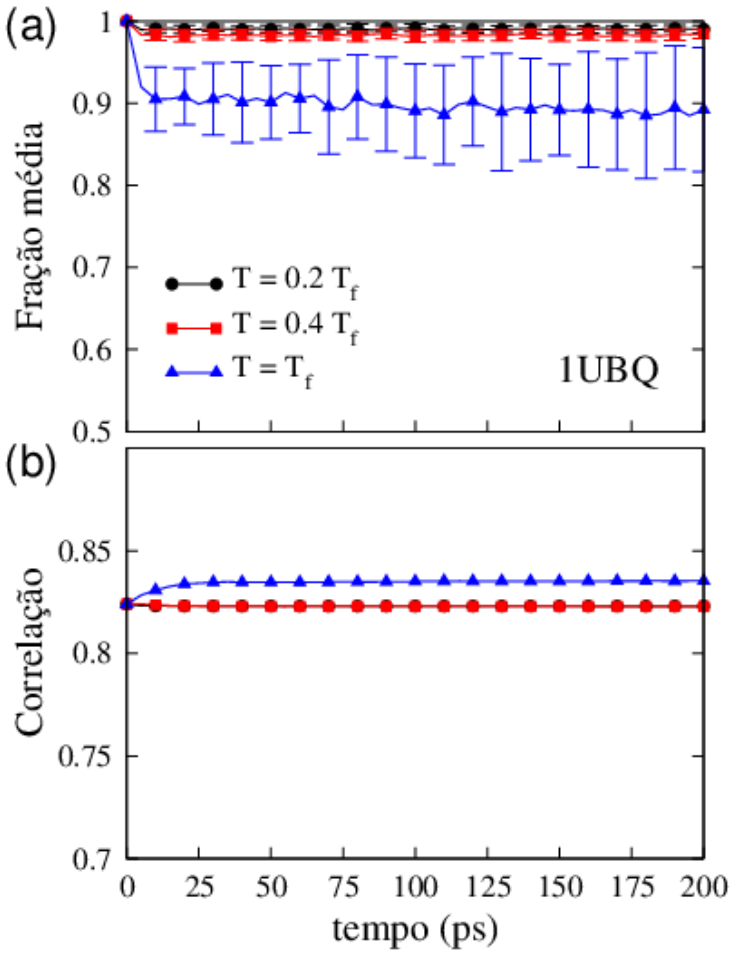


Figura 3.16: *Fração média de contatos nativos Q e coeficiente de correlação ρ com uma função do tempo de simulação em piscosegundos para proteína 1UBQ. A dinâmica do desenovelamento nas 3 temperaturas é governada pelo modelo SBM.*

bém que as estimativas para os pesos podem produzir um menor coeficiente de correlação se estendermos as simulações para a região de equilíbrio. Por outro lado, as estimativas do modelo SBM (figura 3.16(b)) segue a tendência usual, com uma melhoria consistente nas simulações de MD curtas. Estas simulações curtas preservam a maioria dos contactos nativos, conforme mostrado na figura 3.10(a), 3.14(a) e 3.16(a).

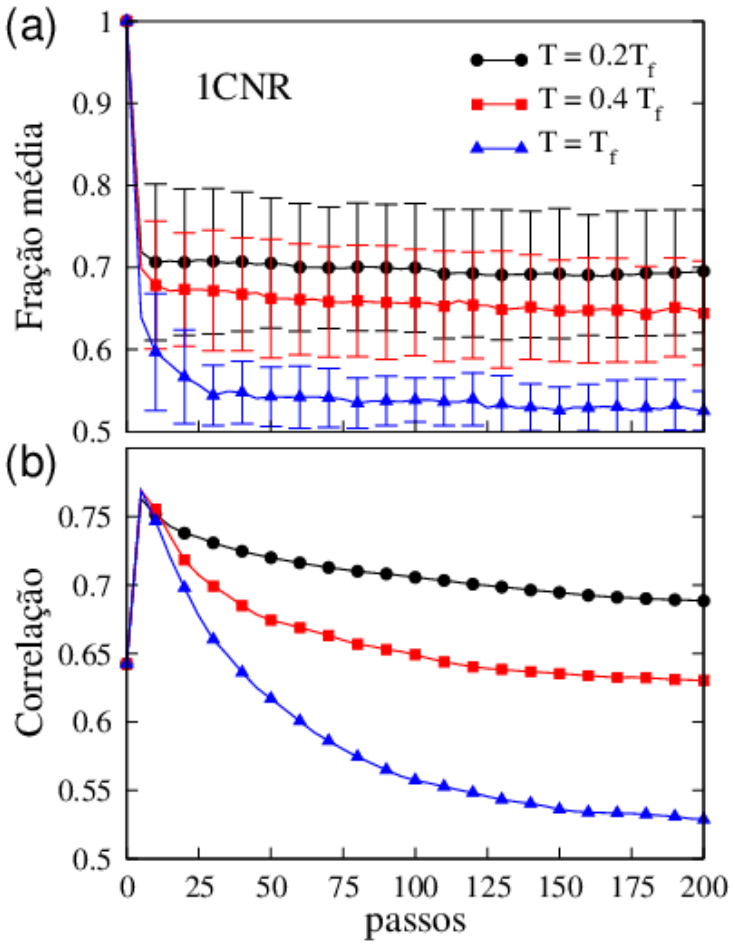


Figura 3.17: *Fração média de contatos nativos Q e o coeficiente de correlação ρ com uma função dos passoa de MC para 1CNR com o modelo AB.*

3.2.5 Crambina e CI-2

A Crambina é uma proteína pequena (presente em plantas) de alta resolução, 1.05 Å, composta por 3 hélices α e seis fitas β. Para esta proteína, o GNM já descreve bem os fatores-B, pois $\rho_{\mathrm{GNM}} = 0.64$.

A figura 3.17 mostra a fração média e o coeficiente de correlação em função do tempo de simulação obtidos a partir das simulações do modelo AB. Considerando os pesos antes do sistema atingir os estados de equilíbrio e fixando a fração média em 0.6, estimamos o coeficiente de correlação $\rho_{\mathrm{wGNM}} = 0.75$ na temperatura de simulação T_f. A figura 3.18 apresenta os resultados do SBM para a proteína 1CNR. Notamos que a dinâmica molecular melhora consistentemente as estimativas dos fatores-B quando a simulação prossegue e

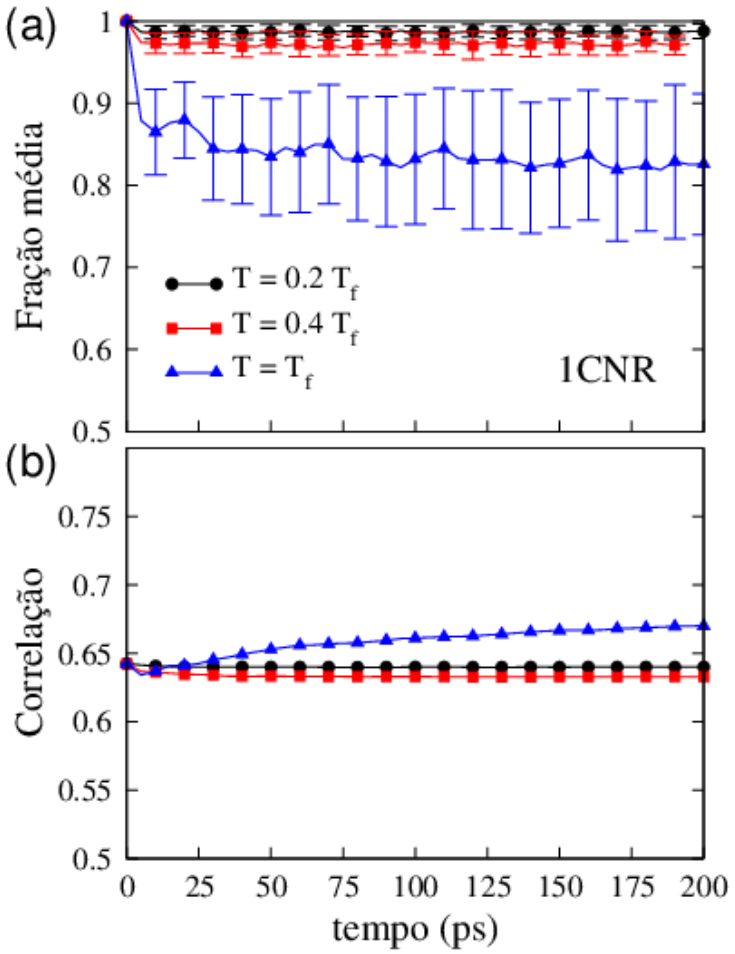


Figura 3.18: *Fração média de contatos nativos Q e o coeficiente de correlação ρ em função do tempo de simulação em picosegundos para a proteína 1CNR com o modelo SBM.*

em 200 ps obtemos $\rho_{\mathrm{wGNM}} = 0.67$.

A 1YPA é uma proteína de 2.0 Å de resolução com 2 hélices e seis fitas-β. A proteína com código PDB 1YPA apresenta o maior coeficiente de correlação, $\rho_{\mathrm{GNM}} = 0.83$ (ver tabela 3.5). Apresentamos os resultados obtidos com as simulações utilizando os modelos AB e e SBM na figura 3.19 e 3.20, respectivamente. Se consideramos a avaliação de w_{ij} até 15 passos de MC ($Q = 0.6$), o coeficiente de correlação usando o modelo AB é $\rho_{\mathrm{wGNM}} = 0.88$ na temperatura T_f. As simulações curtas da dinâmica SBM pode ainda melhorar os resultados mesmo para uma proteína com um alto coeficiente de correlação ($\rho_{\mathrm{wGNM}} = 0.85$).

As figuras para a fração média e o coeficiente de correlação entre os fatores-B em

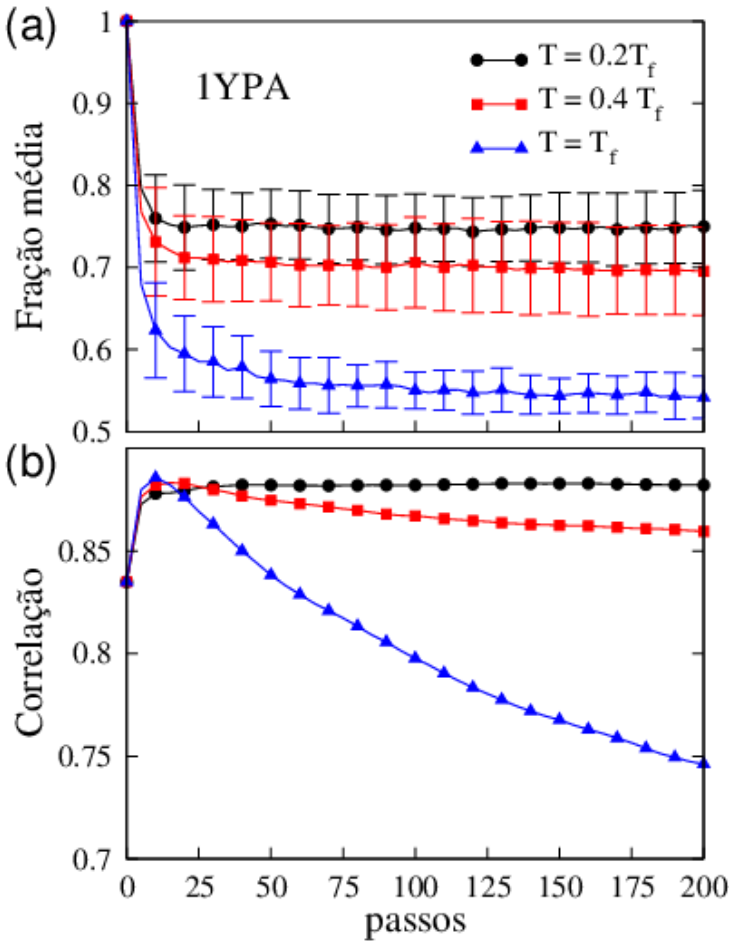


Figura 3.19: *Fração média de contatos nativos Q e coeficiente de correlação ρ em função dos passos de MC para a proteína 1YPA com o modelo AB.*

função do tempo de simulação para a proteína 1CNR são semelhantes à aquelas obtidas para a Ubiquitina (ver figuras 3.15 e 3.16). Embora as simulações com o modelo AB forneçam predições tão boas quanto ou melhores do que os modelos ENM, é importante ressaltar que as estimativas correspondentes ao modelo SBM apresentam melhoras consistentes, mesmo para proteínas com um alto coeficiente de correlação entre as predições do GNM e as experimentais. Este comportamento é evidente na figura 3.16 para a proteína 1UBQ, e nas figuras 3.17 e 3.20 para as proteínas 1CNR e 1YPA, respectivamente. Esta melhora global, pelo menos para este conjunto de dados, é atribuída ao campo de forças usado para descrever as interações entre os átomos de C_α; este campo de forças contêm informações detalhadas sobre as interações C_α-C_α.

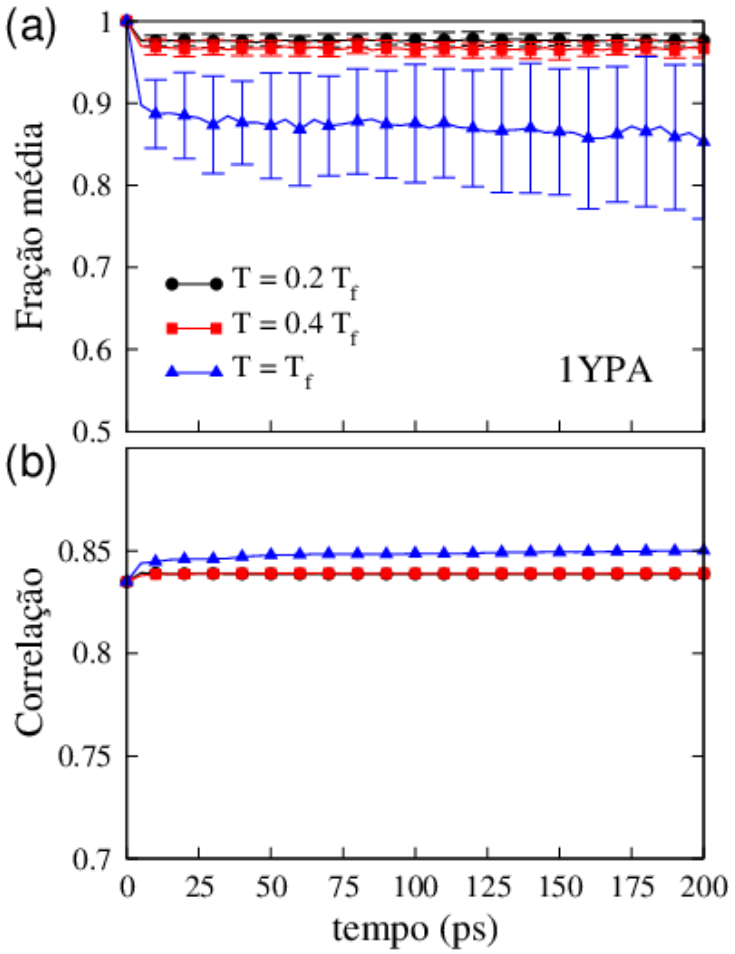


Figura 3.20: *Fração média de contatos nativos Q e o coeficiente de correlação ρ em função do tempo de simulação em picosegundos para proteína 1YPA com SBM.*

3.2.6 Azurina e Cicrofilina-A

A proteína azurina da espécie das bactérias Pseudomonas aeruginosa (código PDB 1E65 e 128 resíduos) é uma proteína de ligação de cobre [63]. Curiosamente, a sua desnaturação de equilíbrio revela um processo de desenovelamento bastante incomum. [64]. Tal desnaturação foi monitorada experimentalmente e parece ocorrer via um processo de transição de dois estados, e é provável que seja uma consequência de diferentes domínios de enovelamento [64]. Isto pode acarretar uma dificuldade adicional na avaliação adequada dos pesos a partir dos nossos modelos de simulação minimalista. Os resultados do wGNM a partir do modelo AB apresenta um comportamento mais complexo em função do tempo

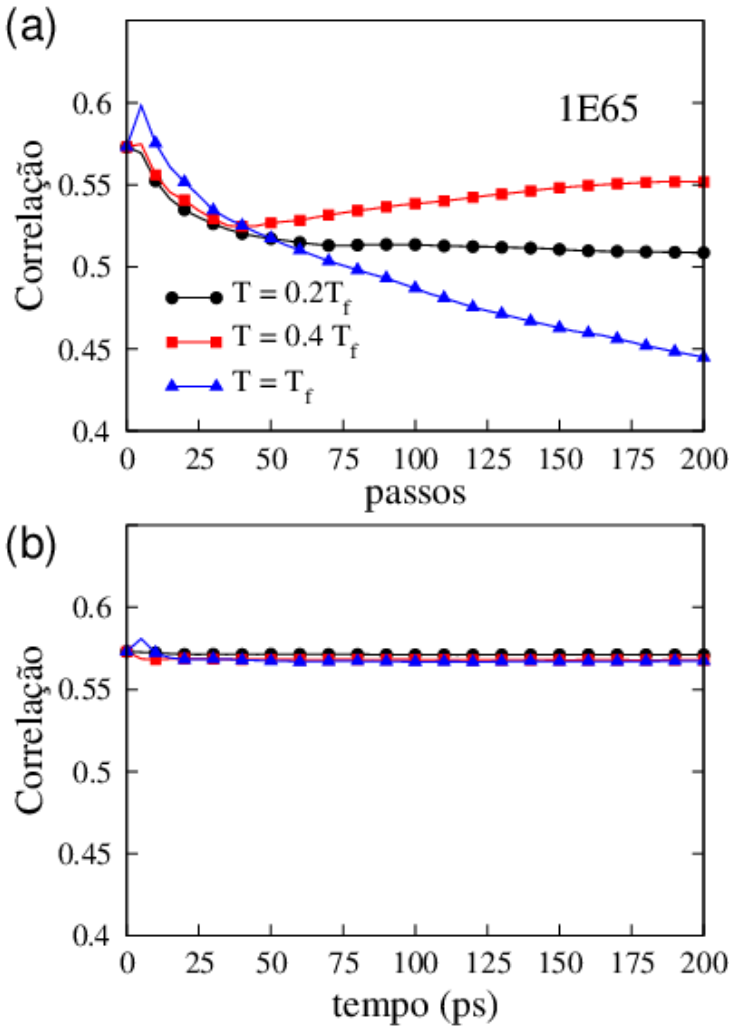


Figura 3.21: *Coeficiente de correlação ρ a partir dos modelos AB (a) e SBM (b) para proteína 1E65*

de simulação para a correlação entre os fatores-B teóricos e experimentais. As simulações curtas de desenovelamento via SBM produzem o mesmo coeficiente de correlação 0.57 do GNM (Ver figura 3.21).

A ciclofilina A (com código PDB 1LOP e 164 resíduos) é uma proteína de um único domínio do tipo β com uma arquitetura barril beta. Os resultados obtidos das simulações AB ou SBM aumenta as correlações entre os fatores-B (0.66 (GNM) e 0.68 (WCN)) para 0.83 e 0.71, respectivamente. A figura 3.22 mostra estes resultados. Aqui, esta proteína relativamente grande apresenta uma transição de dois estados totalmente reversível [65]. Isso pode ter ajudado a estabelecer os pesos adequados, pois foi possível aumentar o valor

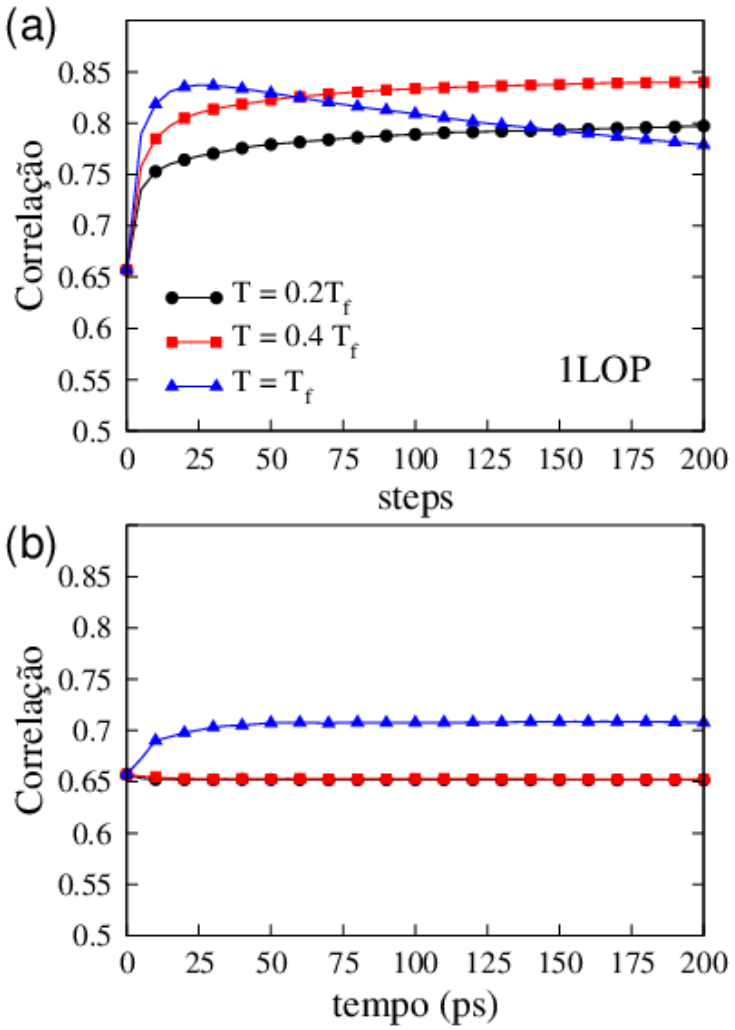


Figura 3.22: *Coeficiente de correlação ρ a partir dos modelos AB (a) e SBM (b) para proteína 1LOP.*

do coeficiente de correlação de 0.66 para 0.83.

3.3 wGNM aplicado às estruturas de NMR

Para avaliar a qualidade dos modelos de rede elástica, é usual e conveniente comparar as predições das flutuações quadráticas médias com os fatores-B experimentais. Entretanto, os fatores-B experimentais podem apresentar imprecisões na descrição das flutuações quadráticas experimentais reais dos átomos. Como os fatores-B experimentais são subprodutos da determinação da estrutura, eles podem conter ruídos como de desordem da rede cristalina, de empacotamento de cristal e assim por diante. Então, para avaliar a qualidade de um modelo de rede elástica, faz se necessário compará-lo com outras quantidades experimentais, por exemplo, com o perfil do *rmsd* para estruturas de NMR, além do fator-B para estruturas de raios-X.

A tabela 3.7 compara os valores da correlação ρ entre as flutuações teóricas calculadas pelo GNM e as flutuações quadráticas experimentais obtidas do *ensemble* de NMR.

Tabela 3.6: *Correlação ρ para estruturas de NMR calculada pelos modelos GNM e wGNM para $R_c = 7.5$ Å. N representa número de resíduos e N_{Model} o número de modelos de NMR.*

Protein	PDB	N	N_{Model}	AB (T_f)	GNM (ρ)	wGNM (ρ)
Rubredoxina	1ZRP	53	40	0.397	0.34	0.67
Proteína G	3GB1	56	32	0.456	0.38	0.47
Proteína L	2PTL	78	21	0.437	0.84	0.84
ADAh2	1O6X	20	81	0.480	0.92	0.91
ACBP	1NTI	86	20	0.424	0.79	0.72
Citocromo C3	1QN0	112	20	0.431	0.53	0.47
Lisozyma	1E8L	129	50	0.574	0.50	0.37
Colagenase	1AYK	169	30	0.511	0.82	0.82

Para as duas proteínas menores, 1ZRP e 3GB1, nota-se que o wGNM apresenta um desempenho melhor que o GNM na predição dessas flutuações. Entretanto, para as demais proteínas o wGNM fornece resultados equivalentes ou ligeiramente inferiores ao GNM.

Para a proteína 106X, o GNM prediz muito bem as flutuações quadráticas experimentais, pois $\rho_{\mathrm{GNM}} = 0.92$. Isso significa que uma única constante de força para todas as interações entre pares de C_α já é suficiente para amostrar satisfatoriamente as flutuações do *ensemble* de NMR; fornecendo poucas margens para melhora. Observa-se o mesmo para as proteínas 2PLT, 1AYK e 1NTI, para as quais $\rho_{\mathrm{GNM}} = 0.84$, 0.82 e 0.79, respectivamente. Tem-se mostrado que o GNM descreve melhor as flutuações para estruturas de raios-X do que para estruturas de NMR.

De modo geral, os resultados do wGNM apresentados na tabela 3.3 utilizando o modelo AB mostram que a melhor opção de modelo para este conjunto de proteínas é o GNM tradicional. Pois em termos de preço que se paga para obter o resultado final, o GNM é mais vantajoso. Como a liberdade do sistema de se movimentar é maior para estruturas de NMR, acreditamos que o potencial AB não consegue amostrar adequadamente as flutuações dos resíduos por causa da simplicidade deste potencial, o que pode ter prejudicado à avaliação dos pesos w_{ij}.

Em aplicações dos modelos de rede elásticas envolvendo estruturas de NMR, para saber qual modelo é o mais recomendado, basta calcular o coeficiente de correlação entre as flutuações quadráticas médias teóricas e experimentais ($rmsd$). O modelo que fornecer o maior coeficiente de correlação será o mais recomendado. Dentre os modelos já existentes, o wGNM pode ser a opção, por exemplo, para a proteína 1ZRP, para qual $\rho_{\mathrm{GNM}} = 0.34$ e $\rho_{\mathrm{wGNM}} = 0.67$.

Capítulo 4

Conclusões e perspectivas

A introdução de diferentes pesos para o potencial de Hooke no modelo GNM é um passo em direção a uma melhor descrição das interações relativas entre átomos de C_α nesse modelo mínimo. O desempenho comparativo do wGNM com o GNM, bem como com outros modelos simples, pfGNM e WCN, estabece como o wGNM pode fornecer estimativas teóricas melhores dos fatores-B experimentais. No estudo comparativo entre os modelos de rede elástica, para as proteínas selecionadas, as quais tinham baixos e altos valores dos coeficientes de correlação entre os fatores-B téoricos preditos pelo GNM e o fator-B experimental, mostramos como o wGNM pode melhorar a predição teórica dos fatores-B. A forma como atribuímos pesos à rede é importante. Nossa simples proposta leva em consideração a frequência relativa com que os contatos nativos permanecem conectados durante os experimentos de desenovelamento. A flexibilidade da proteína depende fortemente das posições dos átomos C_α na estrutura 3D; uma avaliação adequada das constantes de força conectando esses átomos é a base do nosso método. Duas dinâmicas de desenovelamento e dois campos de forças foram comparados.

Um rápido desenovelamento é alcançado com o algoritmo de Monte Carlo no modelo AB, onde as forças hidrofóbicas governam as interações. Este método produz trajetórias com alto número de ligações quebradas. Para explorar melhor o papel desempenhado pelos modelos mínimos em determinar os pesos w_{ij}, efetuamos também simulações de dinâmica molecular com um campo de forças minimalista mais detalhado. Este modelo baseado em estrutura alcança o seu mínimo quando a proteína simulada está em seu estado nativo e a intensidade da interação entre átomos de C_α é especificada pela distância de cada par ij. Resultados baseados sobre trajetórias curtas obtidas com o campo de forças SBM são consistentemente melhores quando comparados com os modelos de rede

elástica tradicionais. No que diz respeito à distância de corte constante, o SBM inspira uma nova definição de rede complexas de aminoácidos onde a constante de corte deveria ser substituída pela distância específica para cada pares de átomos de C_α.

Uma vez que a avaliação de modos de vibração é muito importante para a funcionalidade, qualquer esforço extra em tempo computacional é pago porque wGNM tem um melhor poder preditivo. Além disso, uma boa representação da rede de resíduos de aminoácidos pode levar a uma melhor caracterização das principais propriedades topológicas das conformações da proteína [66, 67, 68, 69, 70].

No que diz respeito à distância de corte constante, parece que uma distância específica para cada par de átomos C_α pode levar a uma melhor definição de tais redes complexas de aminoácidos. Isso pode ajudar a identificar etapas importantes do processo do desenovelamento via análise das alterações topológicas sofridas pela rede de aminoácidos [71].

Tendo-se a estimativa da temperatura de *folding*, o tempo de processamento necessário para obter as figuras como a figura 3.8 (a) é de cerca de 2 min para 50 séries temporais com comprimento de 200 passos de MC com o modelo AB. Este tempo computacional equivale a 30 min para uma simulação de 200 ps com SBM. Aplicações importantes que necessitam de uma melhor caracterização da dinâmica conformacional, como em *protein binding*, podem se beneficiar de um método que já inclui algumas informações fornecidas pelas conformações que coexistem em equilíbrio. Estudos da interação *ligand-binding* exigem uma melhor descrição das mudanças conformacionais de proteínas não ligadas que induzem o ligante. Neste sentido, um método que leva em consideração diferentes conformações parece ir na direção do que é necessário para descrever as dinâmicas coletivas melhor. Além disso, uma boa representação da rede de resíduos de aminoácidos pode ajudar a caracterizar as principais propriedades topológicas de conformações de proteína. Assim, o uso de uma rede ponderada para descrever a dinâmica das proteínas, não só em torno da energia mínima, mas em torno de qualquer mínimo de energia local, pode ajudar a identificar etapas importantes do processo de enovelamento via análise das alterações topológicas sofridas pela rede de aminoácidos.

Temos como perspectiva aplicar o wGNM em problemas específicos como identificação de resíduos chaves responsáveis pela transição conformacional entre os estados "aberto" e "fechado" de proteínas. Termodinamicamente, os resíduos funcionalmente im-

portantes podem ser identificados como aqueles cujas perturbações alteram significativamente a diferença de energia livre entre os estados “aberto” e “fechado". Para calcular esta diferença de energia livre, utilizaremos a funçao de partição configuracional obtida via wGNM associada a estes estados.

Referências Bibliográficas*

[1] SHAKHNOVICH, E. I. Theoretical studies of proteins-folding thermodynamics and kinetics. *Curr. Opin. Struct. Biol.*, v. 7, p. 29, 1997.

[2] HENZLER-WILDMAN, K.; KERN, D. Dynamic personalities of proteins. *NATURE*, v. 450, p. 964, 2007.

[3] MCCAMMON, J. A.; GELIN, B. R.; KARPLUS, M. Dynamic of folded proteins. *NATURE*, v. 267, p. 585, 1997.

[4] ORENGO, C. A.; THORNTON, J. Protein families and their evolution - a structural perspective. *Annu. Rev. Biochem*, v. 74, p. 867, 2005.

[5] WILSON, E. B.; DECIUS, J.; CROSS, P. *New York*. McGraw-Hill: Molecular Vibrations, 1955.

[6] GŌ, N.; NOGUTI, T.; NISHIKAWA, T. Dynamics of a small globular protein in terms of low-frequency. *Proc. Natl. Acad. Sci. USA*, v. 80, p. 3696, 1993.

[7] TIRION, M. M. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Phys. Rev. Lett.*, v. 77, p. 1905, 1996.

[8] FLORY, P. J. Statistical thermodynamics of random networks. *Proc. R. Soc. Lond. A*, v. 351, p. 351, 1976.

[9] HALILOGLU, T.; BAHAR, I.; ERMAN, B. Gaussian dynamics of folded proteins. *Phys. Rev. Lett.*, v. 79, p. 3090, 1997.

*De acordo com a Associação Brasileira de Normas Técnicas. NBR 6023.

[10] MUWAD, L.; PERAHIA, D. Motions in hemoglobin studied by normal mode analysis and energy minimization: evidence for the existence of tertiary T-like, quaternary R-like intermediate structures. *J. Mol. Biol.*, v. 258, p. 392, 1996.

[11] BAHAR, I. et al. Collective motions in HIV-1 reverse transcriptase: examination of flexibility and ensyme function. *J. Mol. Biol.*, v. 80, p. 1023, 1999.

[12] TEMIS, N. A.; BAHAR, I. V. Inhibitor binding alters the directions of domains motions in HIV-1 reverse transcriptase. *Proteins*, v. 49, p. 61, 2002.

[13] THOMAS, A. et al. Tertiary and quaternary conformational changes in aspartate transcarbamylase: a normal mode study. *Proteins*, v. 34, p. 96, 1999.

[14] KESKIN, O. et al. Molecular mechanisms of chaperonin of chaperonin GroEL-GroES function. *Biochemistry*, v. 41, p. 491, 2002.

[15] MING, D. et al. Simulation of F-actin filaments of several microns. *Biophys. J*, v. 85, p. 27, 2003.

[16] TAMA, F. et al. Dynamic reorganization of the functionally active ribosome explored by normal mode analysis and cryo-electron microscopy. *Proc. Natl. Acad. Sci. USA*, v. 100, p. 9319, 2003.

[17] WANG, Y. et al. Global ribosome motions relealed with elastic network model. *J. Struct. Biol*, v. 147, p. 302, 2004.

[18] WYNSBERGHE, A. V.; LI, G.; CUI, Q. Normal-mode analysis suggests protein flexibility modulation throughout rna polymerase' s functional cycle. *Biochemistry*, v. 43, p. 13083, 2004.

[19] CUI, Q. et al. A normal mode analysis of sctructural plasticity in the biomolecular motor F1-ATPase. *J. Mol. Biol.*, v. 340, p. 345, 2004.

[20] RADER, A. J.; VLAD, D. H.; BAHAR, I. Maturation dynamics of bacteriophage HK97 capsid. *Structure*, v. 13, p. 413, 2005.

[21] YANG, L.; SONG, G.; JERNIGAM, R. L. Protein elastic network models and the ranges of cooperativity. *Proc. Natl. Acad. Sci. USA*, v. 106, p. 12347, 2009.

[22] LIN, C. et al. Deriving protein dynamical properties from weighted protein contact number. *Proteins*, v. 72, p. 929, 2008.

[23] YANG, L.-W. et al. Insights into equilibrium dynamics of proteins from comparison of NMR and X-ray data with computational predictions. *Structure*, v. 15, n. 6, p. 741, 2007.

[24] DEMIREL, M. C.; ATILGAN, A. R.; JERNIGAN, R. L. Identification of kinetically hot residues in proteins. *Protein Science*, v. 7, p. 2522, 1998.

[25] TUZMEN, C.; ERMAN, B. Identification of ligand binding site of proteins using the gaussian network model. *PLoS ONE*, v. 6, p. e16474, 2011.

[26] HALILOGLU, T. et al. How similar are protein folding and protein binding nuclei? examination of vibrational motions of energy hot spots and conserved residues. *Biophys. J*, v. 88, p. 1552, 2005.

[27] HALILOGLU, T.; ERMAN, B. Analysis of correlation between energy and residue fluctuations in native proteins and determination of specific sites for binding. *Phys. Rev. Lett.*, v. 102, p. 088103, 2009.

[28] HALILOGLU, T.; GUL, A.; ERMAN, B. Predicting important residues and interactions pathways in proteins using gaussian network model: Binding and stability of HLA proteins. *PLoS Computacional Biology*, v. 6, p. e10000845, 2010.

[29] CHEN, C.; LIN, L.; XIAO, Y. Identification of key residues in proteins by using their physical characters. *Phys. Rev. E*, v. 67, p. 041926, 2006.

[30] GALZITSKAYA, O. V.; IVANKOV, D. N.; FINKELSTEIN, A. V. Folding nuclei in proteins. *Mol. Biol.*, v. 35, p. 708, 2004.

[31] SHMYGELSKA, A. Search for folding nuclei in native protein structures. *Bioinformatics*, v. 21, p. 394, 2005.

[32] SHAKHNOVICH, E.; ABKEVICH, V.; PTITSYN, O. Conserved residues and the mechanism of protein folding. *Letters to Nature*, v. 379, p. 96, 1996.

[33] NOLTING, B.; AGARD, D. A. How general is the nucleation-condensation mechanism? *Proteins*, v. 73, p. 754, 2008.

[34] BOGAN, A. A.; THORN, K. S. Anatomy of hot spots in protein interfaces. *J. Mol. Biol.*, v. 280, p. 1, 1998.

[35] MOREIRA, I. S.; FERNANDES, P. A.; RAMOS, M. J. Hot spots-a review of the protein-protein interface determinant amino-acid residues. *Proteins*, v. 68, p. 803, 2007.

[36] STILLINGER, F. H.; HEAD-GORDON, T.; HIRSHFELD, C. L. Toy model for protein folding. *Phys. Rev. E*, v. 48, p. 1469, 1193.

[37] BACHMANN, M.; ARKIN, H.; JANKE, W. Multicanonical study of coarse-grained off-lattice models for folding heteropolymers. *Phys. Rev. E*, v. 71, p. 031906, 2005.

[38] KYTE, J.; DOOLITTLE, R. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, v. 157, p. 105, 1982.

[39] OOI, T. et al. Accessible surface areas as a measure of the thermodynamic parameters of hydration of peptides. *Proc. Natl. Acad. Sci. USA*, v. 84, p. 3086, 1987.

[40] ROSEMAN, M. Hydrophilicity of polar amino acid side-chains is markedly reduced by flanking peptide bonds. *J. Mol. Biol.*, v. 200, p. 513, 1988.

[41] EISENBERG, D.; MCLACHLAN, A. Solvation energy in protein folding and binding. *Nature*, v. 319, p. 199, 1986.

[42] CLEMENTI, C.; NYMEYER, H.; ONUCHIC, J. N. Topological and energetic factors: what determines the structural details of the transition state ensemble and "enroute" intermediates for protein folding? an investigation for small globular proteins. *J. Mol. Biol.*, v. 298, p. 937, 2000.

[43] KOGA, N.; TAKADA, S. Roles of native topology and chain-length scaling in protein folding: A simulation study with a g-like model. *J. Mol. Biol.*, v. 313, p. 171, 2001.

[44] CHAVEZ, L. L.; ONUCHIC, J. N.; CLEMENTI, C. Quantifying the roughness on the free energy landscape: entropic bottlenecks and protein folding rates. *J. Am. Chem. Soc.*, v. 126, n. 27, p. 8426, 2004.

[45] GOSAVI, S. et al. Topological frustration and the folding of interleukin-1. *J. Mol. Biol.*, v. 357, p. 986, 2006.

[46] NOEL, J. K. et al. Smog@ctbp: simplified deployment of structure-based models in gromacs. *Nucleic Acids Research*, v. 38, p. W657, 2010.

[47] SOBOLEV, V. et al. Automated analysis of interatomic contacts in proteins. *Bioinformatics*, v. 15, p. 327, 1999.

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

[49] KIM, J.; KEYES, T.; STRAUB, J. E. Iteration-free, weighted histogram analysis method in terms of intensive variables. *J. Chem. Phys.*, v. 135, p. 061103, 2011.

[50] DE MENDONçA, M. R. et al. Inferring a weighted elastic network from partial unfolding with coarse-grained simulations. *Proteins*, v. 82, p. 119, 2014.

[51] NEMOTO. Exchange monte carlo method and application to spin glass simulations. *J. Phys. Soc. Japan*, v. 65, p. 1064, 1996.

[52] RIZZI, L. G.; ALVES, N. A. Multicanonical entropy-like solution of statistical temperature weighted histogram analysis method. *J. Chem. Phys.*, v. 135, p. 141101, 2011.

[53] SPOEL, D. V. D. et al. Gromacs: Fast, flexible, and free. *J. Comp. Chem.*, v. 26, p. 1701, 2005.

[54] FERRENBERG, A. M.; SWENDSEN, R. H. Optimized monte carlo data analysis. *Phys. Rev. Lett.*, v. 63, p. 1195, 1989.

[55] BUSHNELL, G. W.; LOUIE, G. V.; BRAYER, G. D. High-resolution three-dimensional structure of horse heart cytochrome c. *J. Mol. Biol.*, v. 214, p. 585, 1990.

[56] DINNER, A. R.; KARPLUS, M. Is protein unfolding the reverse of protein folding? a lattice simulation analysis. *J. Mol. Biol.*, v. 292, p. 403, 1999.

[57] CHUNG, H. S.; TOKMAKOFF, A. Temperature-dependent downhill unfolding of ubiquitin. i. nanosecond-to-millisecond resolved nonlinear infrared spectroscopy. *Proteins*, v. 72, p. 474, 2008.

[58] DASTIDAR, S. G.; MUKHOPADHYAY, C. Unfolding dynamics of the protein ubiquitin: Insight from simulation. *Phys. Rev. E*, v. 72, p. 051928, 2005.

[59] SUŁKOWSKA, J. I. et al. Predicting the order in which contacts are broken during single molecule protein stretching experiments. *Proteins*, v. 71, p. 45, 2008.

[60] DAS, A.; MUKHOPADHYAY, C. Mechanical unfolding pathway and origin of mechanical stability of proteins of ubiquitin family: An investigation by steered molecular dynamics simulation. *Proteins*, v. 75, p. 1024, 2009.

[61] IMPARATO, A.; PELIZZOLA, A. Mechanical unfolding and refolding pathways of ubiquitin. *Phys. Rev. Lett.*, v. 100, p. 158104, 2008.

[62] VIJAY-KUMAR, S.; BUGG, C. E.; COOK, W. J. Structure of ubiquitin refined at 1.8Å resolution. *J. Mol. Biol.*, v. 194, p. 531, 1987.

[63] NAR, H. et al. Crystal structure of pseudomonas aeruginosa apo-azurin at 1.85 A resolution. *FEBS Lett*, v. 360, p. 119, 1992.

[64] LECKNER, J. et al. The effect of the metal ion on the folding energetics of azurin: a comparison of the native, zinc and apoprotein. *Biochem. Biophys. Acta*, v. 1342, p. 19, 1997.

[65] IKURA, T. et al. Fast folding of escherichia coli cyclophilin a: a hypothesis of a unique hydrophobic core with a phenylalanine cluster. *J. Mol. Biol.*, v. 297, p. 791, 2000.

[66] GREENE, L. H.; HIGMAN, V. Uncovering network systems within protein structures. *J. Mol. Biol.*, v. 334, p. 781, 2003.

[67] ALVES, N. A.; MARTINEZ, A. S. Inferring topological features of proteins from amino acid residue networks. *Physica A*, v. 375, p. 336, 2007.

[68] VENDRUSCOLO, M. et al. Small-world view of the amino acids that play a key role in protein folding. *Phys. Rev. E*, v. 65, p. 061910, 2002.

[69] DOKHOLYAN, N. V. et al. Topological determinants of protein folding. v. 99, n. 13, p. 8637, 2002.

[70] ATILGAN, A. R. et al. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys. J.*, v. 80, p. 505, 2001.

[71] KRISHNAN, A. et al. Implications from a network-based topological analysis of ubiquitin unfolding simulations. *PLoS ONE*, v. 3, p. e2149, 2008.

[72] PETSKO, G. A.; RINGE, D. *Protein Structure and Function*. USA: New Science Press Ltd, 2004.

Apêndice A

Demonstrações

A.1 Cálculo do deslocamento quadrático médio no GNM

No GNM, a distribuição de probabilidade de todas as flutuações $P(\Delta\mathbf{R})$ é *isotrópica* e *gaussiana*,

$$P(\mathbf{\Delta R}) = P(\mathbf{\Delta X}, \mathbf{\Delta Y}, \mathbf{\Delta Z}) = p(\mathbf{\Delta X})p(\mathbf{\Delta Y})p(\mathbf{\Delta Z}),$$

onde

$$\begin{aligned} p(\mathbf{\Delta X}) &\propto \exp\left\{-\frac{\gamma}{2k_BT}\mathbf{\Delta X^T \Gamma \Delta X}\right\} \\ &\propto \exp\left\{-\frac{1}{2}\left[\mathbf{\Delta X^T}\left(\frac{k_BT}{\gamma}\mathbf{\Gamma}^{-1}\right)^{-1}\mathbf{\Delta X}\right]\right\}, \end{aligned}$$

k_B é a constante de Boltzmann e T representa a temperatura absoluta. Analogamente, temos expressões semelhantes para $p(\mathbf{\Delta Y})$ e $p(\mathbf{\Delta Z})$. A notação $\mathbf{\Delta X}$ corresponde a $[\Delta X_1, \Delta X_2, ..., \Delta X_i, ..., \Delta X_N]$ e é também uma variável aleatória gaussiana multidimensional com média zero e covariância $(\frac{k_BT}{\gamma})\mathbf{\Gamma}^{-1}$ conforme a definição geral a seguir,

$$W(\mathbf{x}, \mu, \mathbf{\Xi}) = \frac{1}{(2\pi)^N|\Xi|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^T\Xi^{-1}(\mathbf{x}-\mu)\right\},$$

para a função distribuição de probabilidade gaussiana multidimensional associada com um vetor $\mathbf{x}$ N-dimensional tendo valor médio μ e matriz convariância $\mathbf{\Xi}$. Aqui, o termo no denominador, $(2\pi)^N|\Xi|^{1/2}$, é a função de partição que garante a normalização de $W(\mathbf{x}, \mu, \Xi)$ e

$|\Xi|$ é o determinante de Ξ. Semelhantemente, a distribuição de probabilidade normalizada $p(\Delta\mathbf{X})$ é

$$p(\mathbf{\Delta X}) = \frac{1}{Z_{\mathbf{x}}}\exp\left\{-\frac{1}{2}\left[\mathbf{\Delta X^T}\left(\frac{k_BT}{\gamma}\mathbf{\Gamma}^{-1}\right)^{-1}\mathbf{\Delta X}\right]\right\},$$

onde $Z_{\mathbf{X}}$ é a função de partição dada por

$$Z_{\mathbf{X}} = \int \exp\left\{-\frac{1}{2}\left[\mathbf{\Delta X^T}\left(\frac{k_BT}{\gamma}\mathbf{\Gamma}^{-1}\right)^{-1}\mathbf{\Delta X}\right]\right\} d\mathbf{\Delta X} = (2\pi)^{N/2}\left|\frac{k_BT}{\gamma}\mathbf{\Gamma}^{-1}\right|^{1/2}.$$

Sendo as flutuações isotrópicas, podemos escrever a função de partição configuracional total como

$$Z = Z_{\mathbf{x}}Z_{\mathbf{y}}Z_{\mathbf{z}} = (2\pi)^{3N/2}\left|\frac{k_BT}{\gamma}\mathbf{\Gamma}^{-1}\right|^{3/2}.$$

Uma vez determinada a função de partição, temos então toda a informação necessária para calcularmos as grandezas médias deste sistema,

$$\left\langle \mathbf{\Delta X_i^T \Delta X_i}\right\rangle = \frac{\int \mathbf{\Delta X_i^T \Delta X_i}\,\exp\left\{-\frac{1}{2}\left[\mathbf{\Delta X^T}\left(\frac{k_BT}{\gamma}\mathbf{\Gamma}^{-1}\right)^{-1}\mathbf{\Delta X}\right]\right\} d\mathbf{\Delta X}}{\int \exp\left\{-\frac{1}{2}\left[\mathbf{\Delta X^T}\left(\frac{k_BT}{\gamma}\mathbf{\Gamma}^{-1}\right)^{-1}\mathbf{\Delta X}\right]\right\} d\mathbf{\Delta X}}.$$

Utilizando a identidade

$$\frac{\partial}{\partial\gamma}\ln Z = \frac{1}{Z}\frac{\partial Z}{\partial\gamma},$$

obtemos

$$\left\langle \mathbf{\Delta X_i^T \Delta X_i}\right\rangle = -\frac{\partial}{\partial\gamma}\ln Z.$$

Como as flutuações são isotrópicas, podemos escrever ainda

$$\left\langle \mathbf{\Delta X_i^T \Delta X_i}\right\rangle = \left\langle \mathbf{\Delta Y_i^T \Delta Y_i}\right\rangle = \left\langle \mathbf{\Delta Z_i^T \Delta Z_i}\right\rangle = \frac{1}{3}\left\langle \mathbf{\Delta R_i^T \Delta R_i}\right\rangle.$$

Apêndice B

Fator-B

B.1 Determinação da estrutura da proteína por cristalografia de raios-X

A cristalografia de raios-X tem mostrado ser uma técnica poderosa na elucidação de estruturas de moléculas como proteínas. A figura B.1 ilustra a determinação da estrutura da proteína por meio dessa técnica. Um cristal de proteína é bombardeado por um feixe de raios-X, o qual excita os elétrons dos átomos que compõem o cristal, fazendo com que os elétrons oscilem e emitam radiação. A radiação espalhada pelos planos da rede cristalina forma um padrão de difração, o qual é observado por um detector ou por uma folha de filme. Tal padrão é composto por um conjunto de pontos claros e escuros que indicam a intensidade da radiação

$$I \propto |F(hkl)|^2, \tag{B.1}$$

onde $|F(hkl)|$ é módulo do fator de estrutura para uma reflexão h, k e l, que representa a amplitude do espalhamento. Da intensidade da radição, calcula-se o mapa de densidade eletrônica do cristal

$$\rho(x, y, z) = \frac{1}{V} \sum_h \sum_k \sum_l |F_{hkl}| e^{-2\pi i(hx+ky+lk)}.$$

Entretanto, para efetuar tal cálculo é necessário ter os valores das quantidades amplitude do espalhamento e da fase. A amplitude do espalhamento é obtida da intensidade medida, conforme a equação B.1, enquanto a fase não é obtida do padrão de difração. Este consiste em um problema central em cristalografia conhecido como **"problema das fases"**. Para tentar contorná-lo, os cristalógrafos recorrem a outros métodos como o de substituição

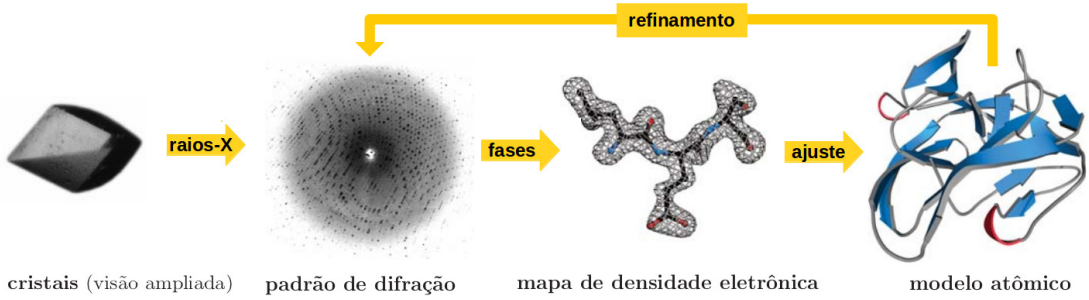


Figura B.1: *(a) O primeiro passo na determinação da estrutura por cristalografia de raios-X é a cristalização da proteína. A fonte de raios-X é frequentemente um síncrotron. Os cristais são bombardeados com raios-X os quais são espalhados pelos planos da rede cristalina e são capturados como um padrão de difração sobre um detector tais como um filme ou um dispositivo eletrônico. Deste padrão e com o uso de uma referência ou fase, informações sobre átomos marcados no cristal, mapas de densidade eletrônica são calculados para as diferentes partes do cristal. Um modelo da proteína é construído do mapa de densidade eletrônica e o padrão de difração para a proteína modelada é calculado e comparado com o padrão de difração atual. O modelo é então ajustado ou refinado para reduzir a diferença entre o padrão de difração calculado e o padrão de difração obtido do cristal, até que a diferença entre o modelo e a realidade seja tão boa quanto possível. Figura adaptada da referência [72].*

molecular para inferir as fases. Do mapa da densidade eletrônica, com ajuda de um programa de computador, um modelo inicial da proteína é construido e o padrão de difração calculado. Este modelo é refinado, variando-se todas as coordenadas dos átomos e o fator-B de modo que a diferença entre o padrão de difração calculado e o medido experimentalmente seja minimizada.

B.2 Fator-B

Apresentamos aqui a demostração da equação do fator-B,

$$B_j = \frac{8}{3}\pi^2\langle u^2\rangle,$$

também conhecido como fator de temperatura, a partir da definição clássica do fator de estrutura considerando o efeito da vibração térmica dos átomos. Nesta equação, B_j representa o fator-B do átomo j e u o seu deslocamento da posição de equilíbrio. Veremos que o fator-B é um termo de correção que surge no fator de estrutura (de espalhamento) para incluir o efeito do movimento dos átomos.

Para uma reflexão h, k, l o fator de estrutura é dado pela expressão

$$F(hkl) = \sum_j f_j(hkl) e^{2\pi i(hx_j + ky_j + lz_j)} \tag{B.2}$$

$$= \sum_j f_j(\vec{h}) e^{2\pi i \vec{h}.\vec{r}_j}, \tag{B.3}$$

onde $\vec{h} = (h, k, l)$, $\vec{r} = (x, y, z)$ e f_j o fator de espalhamento do átomo j.

Em um cristal real, os átomos oscilam em torno das suas posições médias de modo que podemos escrever $\vec{r}_j$ em função do seu deslocamento $\vec{u}_j(t)$, $\vec{r}_j = \vec{r}_j + \vec{u}_j(t)$, e o fator de estrutura como

$$F(\vec{h}) = \sum_j f_j(\vec{h}) e^{2\pi i \vec{h}.\vec{r}_j} e^{2\pi i \vec{h}.\vec{u}_j}. \tag{B.4}$$

A amplitude do espalhamento em uma direção correpondente a $|\vec{h}|$ será uma média no espaço e no tempo da equação B.4, desde que $\vec{u_j}$ varia de uma célula para outra e, dentro de uma célula, varia com o tempo. Podemos então escrever o fator de estrutura em uma temperatura T como:

$$F(\vec{h}) = \sum_j f_j(h) e^{2\pi i \vec{h}.\vec{r}_j} \langle e^{2\pi i \vec{h}.\vec{u}_j} \rangle \tag{B.5}$$

A parte dinâmica do fator de estrutura $\langle e^{2\pi i \vec{h}\vec{u}_j} \rangle$ é chamada de fator de Debye-Waller. Expandindo este fator em séries, temos

$$\langle e^{2\pi i \vec{h}.\vec{u}} \rangle = 1 + 2\pi i \langle \vec{h}.\vec{u} \rangle + \frac{1}{2} \langle (2\pi i \vec{h}.\vec{u})^2 \rangle + ... \tag{B.6}$$

Supondo que o deslocamento seja harmônico e independente de modo que $\langle \vec{h}.\vec{u} \rangle = 0$, temos a aproximação

$$\langle e^{2\pi i \vec{h}.\vec{u}} \rangle = 1 - \frac{1}{2} \langle (2\pi \vec{h}.\vec{u})^2 \rangle, \tag{B.7}$$

mas $\langle (2\pi \vec{h}.\vec{u})^2 \rangle = 4\pi^2 h^2 \langle u^2 \rangle \langle \cos^2 \theta \rangle = \frac{4}{3} \pi^2 h^2 \langle u^2 \rangle$. O termo $\frac{1}{3}$ surge do cálculo da média geométrica de $\cos^2 \theta$ sobre uma esfera (da suposição que o deslocamento é isotrópico). Assim

$$\langle e^{2\pi i \vec{h}.\vec{u}} \rangle = 1 - \frac{4}{6} \pi^2 h^2 \langle u^2 \rangle \approx e^{-\frac{4}{6} \pi^2 h^2 \langle u^2 \rangle}. \tag{B.8}$$

(Vide nota de rodapé)[1] Como $|\vec{h}| = \frac{1}{d} = 2\sin\theta/\lambda$, temos

$$\langle e^{2\pi i\vec{h}.\vec{u}}\rangle = e^{-\frac{8}{3}\pi^2\langle u^2\rangle \sin^2\theta/\lambda^2} = e^{-B_j \sin^2\theta/\lambda^2},$$

onde $B_j = \frac{8}{3}\pi^2\langle u^2\rangle$. Assim, podemos reescrever o fator de estrutura como

$$F(\vec{h}) = \sum_j f_j(\vec{h}) e^{2\pi i\vec{h}.\vec{r}_j} e^{-B_j \sin^2\theta/\lambda^2.} \tag{B.9}$$

Visto que a intensidade da radiação é proporcional a $|F_{hkl}|^2$, então o efeito da vibração térmica dos átomos a uma dada temperatura fará com que a intensidade da radiação decresça de forma exponencial. Quanto maior for a flutuação de um átomo, ou seja, maior o seu fator-B, maior será tal decréscimo.

Na etapa de refinamento da estrutura de raios-X, o fator-B é determinado por ajuste de mínimos quadrados

$$\sum_j w(\vec{h})[|F_{\rm obs}(\vec{h})| - |F_{\rm cal}(\vec{h})|]^2, \tag{B.10}$$

onde $w(\vec{h})$ é a função peso para cada reflexão, com $|F_{\rm obs}|$ e $|F_{\rm cal}|$ sendo os fatores de estrutura observados e calculados, respectivamente.

[1] $e^{-\frac{4}{6}\pi^2h^2\langle u^2\rangle} = 1 - \frac{4}{6}\pi^2h^2\langle u^2\rangle$