Universidade de São Paulo
Instituto de Física de São Carlos

Luciano Borges Censoni

Dinâmica molecular e redes complexas no estudo da difusão térmica em xilanases da família 11

São Carlos

2013

Luciano Borges Censoni

Dinâmica molecular e redes complexas no estudo da difusão térmica em xilanases da família 11

Dissertação apresentada ao Programa de Pós-Graduação em Física do Instituto de Física de São Carlos da Universidade de São Paulo, para obtenção do título de mestre em Ciências.

Área de Concentração: Física Aplicada - Opção Biomolecular

Orientador: Prof. Dr. Leandro Martínez

Versão Corrigida
(versão original disponível na Unidade que aloja o Programa)

São Carlos

2013

Ficha catalográfica elaborada pelo Serviço de Biblioteca e Informação do IFSC,
com os dados fornecidos pelo(a) autor(a)

Censoni, Luciano Borges
Dinâmica molecular e redes complexas no estudo da difusão térmica em xilanases da família 11 / Luciano Borges Censoni; orientador Leandro Martínez - versão corrigida -- São Carlos, 2013.
102 p.

Dissertação (Mestrado - Programa de Pós-Graduação em Física Aplicada Biomolecular) -- Instituto de Física de São Carlos, Universidade de São Paulo, 2013.

1. Proteínas. 2. Xilanases. 3. Redes complexas. 4. Difusão térmica. 5. Centralidade. I. Martínez, Leandro, orient. II. Título.

AGRADECIMENTOS

Pretendo ser sucinto. Este trabalho dificilmente teria sido possível, e certamente não teria significado, sem o apoio e a colaboração de um grande conjunto de pessoas. Cito-as aqui para que se registre o meu agradecimento.

Agradeço aos meus pais, Marcia e Olyntho, pelo apoio incondicional e sacrifício investidos na minha formação. O título de mestre, ainda que associado ao meu nome, será propriedade deles.

Agradeço ao meu orientador, Leandro, pela orientação extremamente competente, associada à amizade e paciência, e pelo seu ensinamento mais importante: *para que serve* um orientador.

Agradeço aos companheiros de sala, Mariana, Heloisa, Prof. Camila e Prof. Alessandro, pela amizade, simpátia e extensa reserva de bom humor.

Agradeço aos amigos, pela presença ao longo do desenvolvimento do trabalho e da pessoa: Rodrigo, Pedro Napoleão, Rodolfo e Milena, Bruno, Sato, Caio, Anna, Luisa, Paula, Gabriel, Pedro Ramón e todos os outros que porventura me fogem momentaneamente à lembrança.

Agradeço também aos professores do IFSC, pela altíssimo nível da formação que recebi, e aos funcionários, em particular ao Ricardo, à Patrícia e a Neusa, pelo bom humor e eficiência com que sempre resolveram tudo que levei para eles.

Certain shapes and patterns hover over different moments in time, haunting and inspiring the individuals living through those periods. The epic clash and subsequent resolution of the dialectic animated the first half of the nineteenth century; the Darwinian and social reform movements scattered web imagery throughout the second half of the century. The first few decades of the twentieth century found their ultimate expression in the exuberant anarchy of the explosion, while later decades lost themselves in the faceless regimen of the grid. You can see the last ten years or so as a return to those Victorian webs [...].

Steven Johnson, *Emergence*

Working with computers has provided us with a new metaphor for the laws of nature:
they carry as much (and as little) information as algorithms.

Daan Frenkel, Berend Smit, *Understanding Molecular Simulations*

RESUMO

CENSONI, L. B. *Dinâmica molecular e redes complexas no estudo da difusão térmica em xilanases da família 11*. 2013. 100 p. Dissertação (Mestrado em Ciências) – Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2013.

Proteínas tipicamente são capazes de manter a sua conformação funcional somente dentro de um intervalo limitado de temperaturas. A despeito do maquinário sofisticado de manutenção da homeostase celular, é sabido que uma variedade de fenômenos moleculares são capazes de induzir desequilíbrios localizados de energia vibracional, e que a eficiência com que cada proteína dissipa estas perturbações pode estar relacionada com a sua tolerância a altas temperaturas. No entanto, a transferência de energia térmica entre diferentes segmentos de uma cadeia proteica é difícil de caracterizar experimentalmente. Uma alternativa teórica para a investigação destes mecanismos é o emprego de simulações de Dinâmica Molecular, particularmente associadas à técnica de Difusão Térmica Anisotrópica (ATD).

Aqui, verificamos a possibilidade de empregar conceitos da teoria de Redes Complexas para construir modelos para estruturas de proteínas, e por meio destes identificar resíduos com capacidade significativa de dissipar perturbações térmicas. Investigamos os diversos protocolos de construção de modelos de rede para proteínas encontrados na literatura, e utilizamos dados experimentais representativos da base SCOP para calcular com rigor os parâmetros numéricos necessários. Produzimos uma definição precisa para o conceito de *contato* entre resíduos de aminoácidos, e a partir desta calculamos a *centralidade* de cada resíduo. Com isto, demonstramos que, em um conjunto de Xilanases para as quais dispomos de dados de ATD, a capacidade de difundir perturbações térmicas é fortemente correlacionada com a *centralidade de proximidade* de cada resíduo, fornecendo argumentos para o uso de modelos de rede para estudar a termoestabilidade de proteínas.

PALAVRAS-CHAVE: Proteínas. Redes complexas. Xilanases. Difusão térmica. Centralidade.

ABSTRACT

CENSONI, L. B. *Molecular dynamics and complex networks in the study of thermal diffusion in family 11 xylanases*. 2013. 100 p. Dissertação (Mestrado em Ciências) – Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2013.

Proteins are typically able to mantain a functional conformation only within a narrow range of temperatures. In spite of the complex cellular homeostatic machinery, it is known that a variety of molecular phenomena can induce localized vibrational imbalances, and that the efficiency with which each protein dissipates these perturbations may be related to its tolerance of higher temperatures. The transference of thermal energy among different sections of a protein chain is, however, hard to characterize experimentally. A theoretical alternative for the investigation of these mechanisms is the use of Molecular Dynamics simulations, particularly when associated with the Anisotropic Thermal Diffusion (ATD) technique.

In this work, we verify the possibility of using concepts from Network Theory to construct models for protein structures, and using those to reveal residues with significant ability to dissipate thermal perturbations. We investigate several protocols of network model construction for proteins present in the literature, and we study representative experimental data from the SCOP database to rigorously calculate the necessary parameters. We produce a precise definition for the concept of *contact* between amino acid residues, and from this we calculate the *centrality* of each residue. We then show that, in a set of Xylanases for which we have data from ATD experiments, the ability to dissipate thermal perturbations is highly correlated to the *closeness centrality* of each residue, providing arguments for the use of network models to study protein thermal stability.

KEYWORDS: Proteins. Xylanases. Complex networks. Thermal diffusion. Centrality.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

CAPÍTULO 1

Introdução

1.1 Modelagem Computacional de Proteínas

Proteínas* são macromoléculas biológicas encontradas em todas as células vivas, correspondendo tipicamente à metade do peso seco da maior parte dos organismos.

A abundância e variedade de exemplares encontrados é consequência de sua importância no metabolismo celular: a maioria das funções celulares são realizadas por proteínas, que atuam sobre os outros tipos, passivos, de biomoléculas, segundo as instruções armazenadas na informação genética.

Do ponto de vista bioquímico, proteínas são polímeros lineares heterogêneos constituídos de L-α-aminoácidos ligados entre si por ligações peptídicas, tal como ilustrado pela figura 1.1. Com esparsas exceções†, todas as células sintetizam proteínas, empregando um conjunto de diferentes tipos de aminoácidos (em geral vinte, com pequenas variações em alguns organismos) que diferem entre si pelo grupo substituinte do carbono central ou *carbono alfa* (C_α), denominado de cadeia lateral. Ao ser incorporado ao polímero durante a síntese, cada aminoácido forma uma ligação covalente com a extremidade carboxilato exposta do anterior, que perde um grupo hidroxila na forma de água. Denominamos então o grupo que sobra, que é incorporado à sequência, de resíduo de aminoácido ou simplesmente *resíduo*, nomenclatura que será adotada neste trabalho daqui em diante.

A cada proteína corresponde uma sequência bem definida de aminoácidos, denominada *estrutura primária*; experimentalmente, verifica-se que, após hidrólise completa de todas as ligações peptídicas, cada proteína gera uma mistura de aminoácidos com uma dada compo-

*Este capitulo é baseado principalmente no conteúdo de (13), cuja leitura é recomendada aos leitores interessados em aprofundar seu conhecimento dos tópicos apresentados. Outras referências consultadas serão citadas ao longo do texto, quando apropriado.

†O autor agradece ao Prof. Milton Sonoda por apontar os glóbulos vermelhos sangüíneos maduros como um dos raros contra-exemplos.

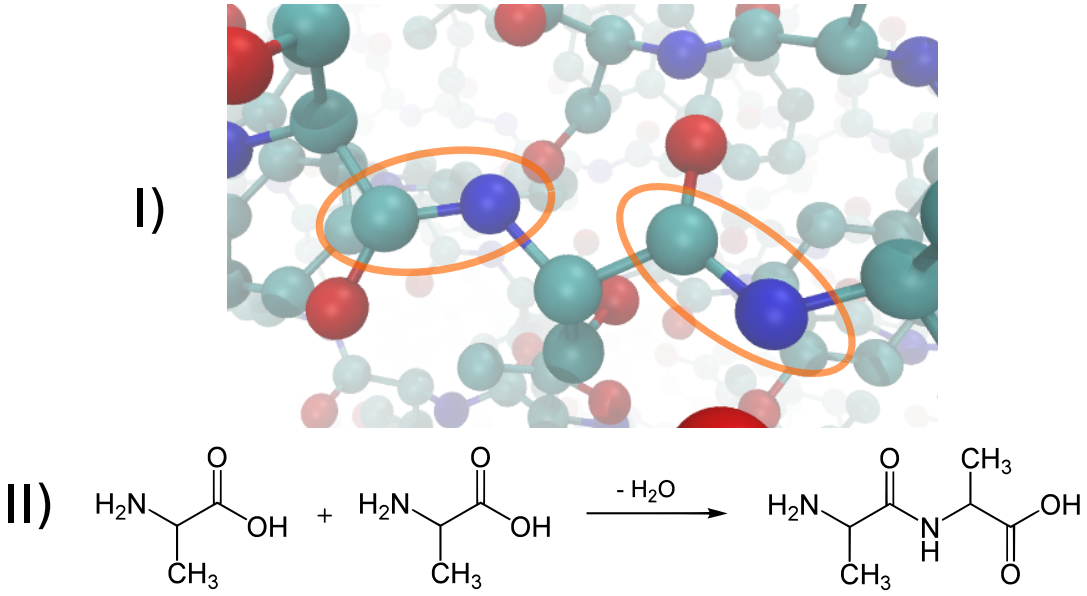

Figura 1.1 – *(I) Exemplo de ligações peptídicas (em destaque) na cadeia principal em uma proteína. Átomos de Oxigênio são representados em vermelho, Nitrogênio em azul e Carbono em ciano. Átomos de Hidrogênio não são representados. Figura gerada com o programa VMD (1, 2). (II) Formação de ligação peptídica a partir de dois aminoácidos (Alaninas, nas quais CH_3 é a cadeia lateral), com perda de uma molécula de água.*

sição, característica e invariante. O tipo e a ordem dos resíduos na cadeia, cada qual com propriedades físico-químicas distintas, determina univocamente a conformação assumida (ou pequeno número de conformações assumidas) pela mesma em solução, num processo denominado *enovelamento*. Esta conformação, que denominamos *estrutura* da proteína, é estável, e da sua reprodutibilidade depende a manutenção da homeostase, e, por consequência, da vida da célula.

A reprodutibilidade da conformação funcional é, talvez, a propriedade mais notável das proteínas, e certamente explica a sua posição de destaque no metabolismo. Experimentalmente, este fato tem uma demonstração simples. É sabido que pequenas mudanças no ambiente físico-químico de uma proteína em solução são capazes de acarretar a perda de sua estrutura, num processo denominado *desnaturação*. Estas perturbações podem incluir mudanças de pH ou temperatura, a adição de alguns tipos de solvente orgânico, de detergentes ou de pequenas moléculas tais como uréia ou cloreto de guanidínio. A perda de estrutura pode ser total, resultando em uma cadeia com conformação aleatória, ou pode ser parcial, gerando uma estrutura, ou, no caso mais geral, um conjunto de estruturas intermediárias, que retém sinais de estruturação mas já não possuem a atividade da conformação nativa da proteína. Para algumas proteínas, o simples retorno ao ambiente químico original da solução, por meio da retirada do agente desnaturante ou do reajuste da temperatura ou do pH, é suficiente para que a cadeia de aminoácidos novamente encontre sua conformação ativa e estável. Isto demonstra que a

informação necessária e suficiente para o enovelamento em solução é contida completamente na sequência de aminoácidos.

A reprodutibilidade do enovelamento de estruturas de proteína se deve, em grande medida, à manutenção de condições precisas dentro do ambiente celular. Por outro lado, mesmo sob condições controladas, a maior parte das sequências de aminoácidos não atinge nenhuma configuração particularmente bem determinada. Para explicar este comportamento, é necessário detalhar os mecanismos que levam ao enovelamento.

Toda proteína, quando sintetizada, consiste em uma cadeia linear de resíduos de aminoácidos expostos ao solvente. Cada um dos vinte aminoácidos previstos no código genético de um organismo exibe propriedades físico-químicas próprias, tais como volume/área de superfície, carga elétrica em pH fisiológico e número de aceptores/doadores de ligação de Hidrogênio; podemos dividi-los, de modo geral, entre aminoácidos apolares, polares e carregados. A princípio, poderíamos considerar que a formação de interações eletrostáticas entre aminoácidos carregados e de ligações de Hidrogênio entre grupos doadores e aceptores de aminoácidos polares é a força motriz do processo de enovelamento. Contudo, com a cadeia estendida, cada um destes resíduos faz interações do mesmo tipo com as moléculas de água do solvente, de modo que a formação de interações não-covalentes entre resíduos não é, energeticamente, um fator determinante por si só.

Considerações estatísticas, por outro lado, a princípio, não lançam luz sobre o enigma. O estado enovelado via de regra exibe alto grau de organização, sendo a posição relativa de cada átomo muitas vezes bem definida com precisão que chega a não mais do que o dobro do seu diâmetro (14). A transição de um estado desnaturado essencialmente aleatório, para o qual inúmeras configurações são acessíveis, para um estado enovelado altamente estruturado envolve, então, a obtenção de um alto grau de organização, com concomitante diminuição significativa de *entropia*, energeticamente desfavorável. Sob este enfoque, a tendência das cadeias de se enovelar torna-se verdadeiramente desconcertante.

Contudo, ainda que todos os aminoácidos possam manter ligações de hidrogênio com a água por meio de seus grupos carboxilato e amina, as cadeias laterais dos aminoácidos apolares, ou *hidrofóbicos*, não o fazem. Como consequência, as moléculas de água ao seu redor são forçadas a assumir outra configuração, organizando-se em torno das cadeias hidrofóbicas de forma a continuar mantendo ligações de hidrogênio entre si, num processo chamado *solvatação*. As novas interações formadas são em geral menos intensas, e o fato de que essa nova configuração implica em uma diminuição da entropia torna-a energeticamente desfavorável, mas, ao mesmo tempo, fornece às cadeias de aminoácidos uma estratégia chave para o enove-

lamento: ao aproximar os resíduos hidrofóbicos entre si, empacotando-os, é possível diminuir a área total das cadeias laterais hidrofóbicas que fica de fato exposta ao solvente. Ao empacotar todos os resíduos hidrofóbicos, formando o chamado *core* ou *núcleo* hidrofóbico, a cadeia de aminoácidos minimiza a área hidrofóbica em torno da qual as moléculas de água devem se organizar, sendo portanto um processo entropicamente altamente favorável. A figura 1.2 é um representação simplificada deste processo.

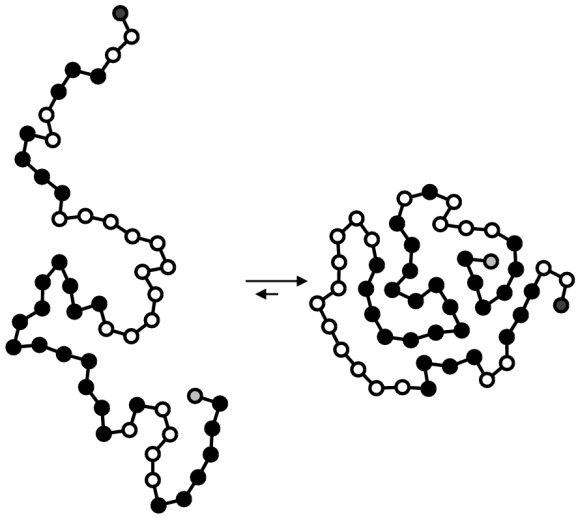

Figura 1.2 – *Ilustração do colapso do núcleo hidrofóbico durante o enovelamento de uma proteína. Círculos cheios representam aminoácidos hidrofóbicos, e círculos vazios representam aminoácidos hidrofílicos. (3).*

O colapso do núcleo hidrofóbico explica porque as cadeias de aminoácidos (pelo menos, as assim forjadas pela seleção natural) preferem procurar uma conformação globular ao invés de estendida. Contudo, não é explicação suficiente para o fato de que as estruturas nativas são sempre iguais em solução - de fato, de uma estrutura aleatória contendo aminoácidos polares e apolares, não há razão para esperar que haja apenas uma maneira de se formar o núcleo hidrofóbico, e, uma vez formado um núcleo compacto, os aminoácidos polares expostos ao solvente poderão interagir favoravelmente entre si de diversas maneiras energeticamente próximas ou equivalentes. Como é possível que uma proteína aleatória mantenha uma estrutura bem determinada, se diversas conformações estruturalmente muito distintas podem ser energeticamente equivalentes? A resposta, experimental, é que independentemente da vontade dos químicos teóricos e físicos estatísticos, ela mantém - mas a pergunta é que está mal formulada. As proteínas que adotam uma conformação nativa estável o suficiente para que possam cumprir sua atividade bioquímica não são de maneira alguma *aleatórias* - elas são o produto de grande número de gerações de seleção natural, cuja atuação é no sentido de favorecer, em cada posição da cadeia, a presença de um aminoácido que prefira fazer interações que de fato existam na conformação nativa (favoráveis, portanto, no sentido termodinâmico), e que prefira

não fazer interações que existam em conformações alternativas e/ou em armadilhas locais do processo de enovelamento (favoráveis, portanto, no sentido cinético). Ao escolher cuidadosamente o resíduo que ocupa cada posição, a seleção natural evita a existência de interações que "competem" entre si, "frustrando" a estabilidade do estado nativo, e promove o surgimento de interações "coerentes" ou "consistentes". Esta máxima, denominada *Princípio da Frustração Mínima*, é a peça que completa a explicação teórica do enovelamento de proteínas (14).

A despeito da confiabilidade do processo de enovelamento de proteínas dentro da célula e do seu bom entendimento em âmbito geral, a previsão do estado nativo de proteínas a partir de suas estruturas primárias continua não sendo, de forma alguma, corriqueira (15). Melhores resultados, por outro lado, têm sido obtidos na determinação experimental de suas estruturas, principalmente com o emprego de técnicas que exploram a interação da matéria com a radiação eletromagnética. Tal interação torna-se rapidamente mais complicada na medida em que crescem o tamanho e a complexidade do sistema estudado; no caso de macromoléculas biológicas, os primeiros avanços significativos advieram do estudo de *cristais*.

A cristalização é uma tendência espontânea de soluções supersaturadas; no caso de proteínas, as dificuldades inerentes à purificação, a coexistência de múltiplas conformações em solução e o pequeno número de contatos intermoleculares possíveis em relação à massa molecular fazem com que o nível de supersaturação necessário para a formação de cristais bem-ordenados seja alto em relação a pequenas moléculas ou sais, e que os mesmos sejam mais frágeis (16). Contudo, cristais de proteínas (tal como cristais em geral) exibem uma propriedade que os torna valiosos para estudos estruturais: sua composição uniforme e repetitiva, consistindo em inúmeras unidades idênticas (ou quase idênticas) "encaixadas" umas nas outras de maneira bem determinada, permite que estes interajam com a radiação eletromagnética de maneira previsível. Quando um cristal de proteína é exposto a um feixe de raios-X, a direção e intensidade dos raios-X refletidos depende da sua estrutura, tornando possível a determinação da mesma através de manipulação matemática do padrão de intensidades (a figura 1.3 ilustra alguns destes passos). Esta técnica, extremamente importante, é denominada *cristalografia de proteínas* associada à *difração de raios-X*, e, introduzida na década de 1950 (destacamos, por exemplo, (17)), impulsionou a Biologia Molecular numa tendência referida por Onuchic *et al.* como a "obsessão por *estrutura* do século XX" (14), em tradução livre.

Avanços subseqüentes revelaram a possibilidade de estudar a estrutura de macromoléculas *em solução*, usando pulsos de radiofreqüência em experimentos sofisticados de *Ressonância Magnética Nuclear* para sondar as interações entre *núcleos* atômicos vizinhos (18). Se por um lado tais protocolos fornecem apenas informações locais de distância entre pares específicos de átomos, gerando muitas vezes restrições ambíguas ou insuficientes para a determinação precisa

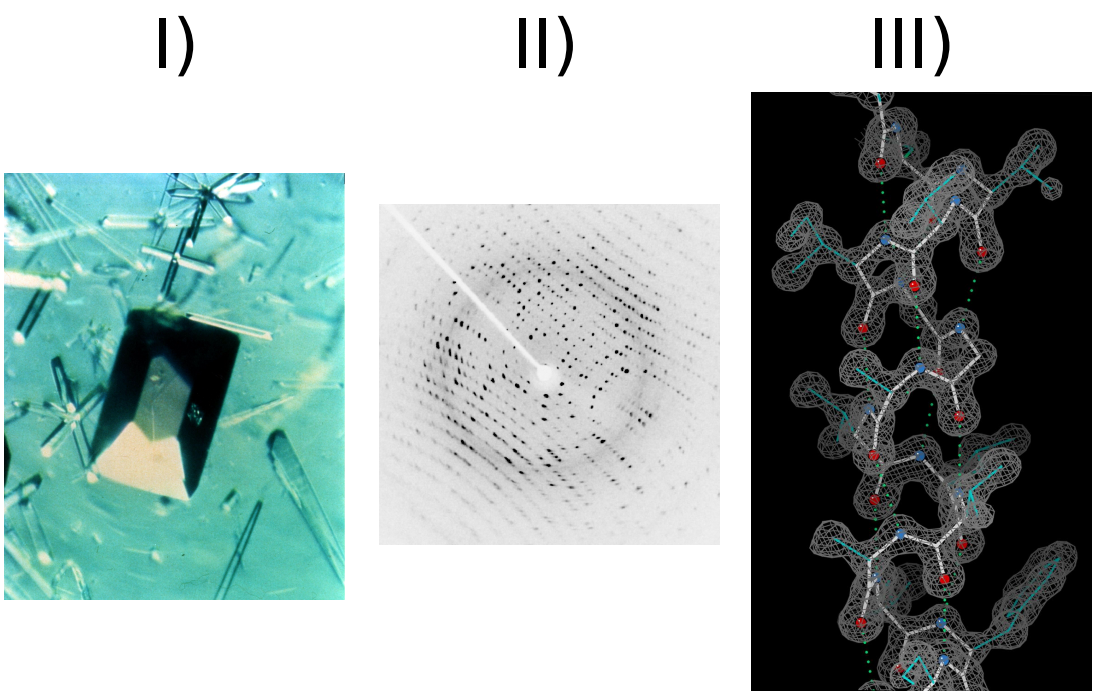

Figura 1.3 – *Ilustração do processo de determinação da estrutura de proteínas por Difração de Raios-X. (I) Exemplo de cristal de proteína. (4). (II) Exemplo de padrão de intensidades de Raios-X difratados por um cristal de proteína. (5). (III) Exemplo de densidade eletrônica calculada a partir de um padrão de difração de um cristal proteico, com modelo de cadeia principal ajustada à densidade eletrônica. (6).*

da estrutura de proteínas grandes (19), os mesmos tornam acessível a descrição do conjunto (ou *ensemble*) de conformações assumidas por cada proteína em solução, adicionando uma dimensão *dinâmica* às imagens estáticas fornecidas pela difração de raios-X.

O número de estruturas conhecidas, por sua vez, fornece subsídios para a predição computacional de estruturas ainda não determinadas. Nos melhores casos, uma dada proteína cuja seqüência de aminoácidos se conhece e cuja estrutura se quer determinar possui uma ou mais *homólogas*, cujas estruturas já são resolvidas e das quais se pode esperar alto grau de similaridade. Tipicamente, uma similaridade entre as sequências maior que 30% indica um grau de semelhança estrutural alto o suficiente para guiar a geração de um modelo baseado no alinhamento das estruturas. Esta técnica e suas variações são englobadas pelo termo "modelagem por homologia", e exibem a capacidade de resolver, em tempos computacionais razoáveis, estruturas aproximadas (de baixa resolução) para genomas inteiros (15).

Mas, no caso geral, pode não haver homólogas da proteína alvo na base de dados. Neste caso, predições podem ser feitas baseadas em conhecimento genérico abstraído do conjunto de estruturas resolvidas, tal como, por exemplo, distribuições de distâncias e ângulos ou a propensão de certos tipos de aminoácidos de formar hélices ou *loops*; predições podem também ser baseadas na exploração exaustiva do espaço de configurações, guiada por princípios físico-

químicos fundamentais. Ambas as estratégias são frequentemente utilizadas em paralelo, e englobadas sob o termo "modelagem *ab initio*". Técnicas de modelagem *ab initio* são capazes de computar, em tempo computacional razoável, estruturas nativas aproximadas para proteínas pequenas (15, 20, 21).

Uma vantagem dos métodos *ab initio*, ainda que imposta pelas características do problema, é que eles muitas vezes são, ao contrário dos métodos de modelagem por homologia, obrigados a conter hipóteses sobre a maneira através da qual proteínas exploram as configurações possíveis até encontrar o estado nativo; de fato, uma busca puramente aleatória no espaço conformacional requereria escalas de tempo muito maiores do que cosmológicas; o fato de que proteínas encontram seu estado nativo *in vivo* e *in vitro* é demonstração suficiente de que a busca não é aleatória, e evidências sugerem que a busca acontece hierarquicamente nas escalas de tamanho, com estruturas locais se formando primeiro, e velocidades de enovelamento sendo correlacionadas com o grau de localidade da topologia da estrutura (15).

A maneira mais direta de modelar a exploração conformacional de uma proteína é também a mais custosa: resolver, no tempo, as equações de movimento de cada átomo que a compõe, regidas pelas leis da Mecânica Quântica. Naturalmente, diversas aproximações devem ser empregadas para que esta estratégia se torne minimamente viável: discretização do tempo, do espaço, e a substituição das interações puramente quânticas por potenciais de interação contendo grande número de parâmetros, os quais são afinados entre si para que o potencial ou "campo de força" reproduza, sob determinadas condições, resultados experimentais. Este conjunto de condições define a técnica de Dinâmica Molecular, que receberá atenção detalhada no próximo capítulo por ser extensivamente empregada ao longo deste trabalho. Os campos de força comumente empregados em simulações de dinâmica molecular podem ser utilizados diretamente como funções de energia para avaliar a qualidade de modelos gerados para uma proteína de estrutura desconhecida, e simulações curtas são eventualmente utilizadas para refinar modelos de baixa resolução. Contudo, a técnica ainda é proibitivamente custosa para gerar modelos a partir de conformações aleatórias ou completamente estendidas (21).

Alternativas para técnicas puramente físico-químicas se fazem, então, necessárias quando o custo computacional é considerado. Muitas delas são baseadas em conhecimento adquirido pela mineração das bases de dados de estruturas, e os cientistas têm sido notavelmente criativos na aplicação da estatística para descobrir conjuntos de propriedades capazes de avaliar o grau de "natividade" de uma estrutura. Evitaremos, por brevidade, enumerar as diversas propriedades que são eventualmente empregadas na classificação de modelos; O leitor interessado é direcionado a (22–26). Mais importante é notar que estas mesmas técnicas podem ser empregadas também na avaliação de outros aspectos da bioquímica de proteínas, tal qual a

previsão de interações proteína-proteína ou a atribuição de atividade catalítica a proteínas de função desconhecida.

O presente trabalho segue este tema, empregando modelos simplificados baseados na Teoria de Redes Complexas para avaliar propriedades de proteínas que não são imediatamente evidentes pela inspeção da estrutura. O resultado mais elegante obtido consiste na utilização inovadora desta técnica de modelagem para a previsão do fluxo de energia vibracional em proteínas, com o objetivo inicial de avaliar a sua resposta a altas temperaturas. Com os resultados obtidos, estabeleceremos então uma ponte entre uma técnica baseada em física (no caso, a Dinâmica Molecular) e a técnica de modelagem empregada, baseada em redução de escala e abstração. A técnica de modelagem será detalhada no capítulo 3, precedida por uma introdução à teoria de redes complexas, e os resultados apresentados nos capítulos subsequentes.

Por hora, fecharemos este capítulo apresentando o conjunto de estruturas experimentais que foram usadas nas análises subsequentes. As proteínas estudadas, xilanases, foram escolhidas em parte por características intrigantes de suas estruturas, opacas à investigação superficial por serem demasiado semelhantes entre si e ao mesmo tempo apresentarem comportamentos fortemente diferentes, porém amigáveis ao emprego de técnicas matemáticas e computacionais por seu tamanho reduzido e globularidade e pela disponibilidade de grande quantidade de dados experimentais.

1.2 Xilanases

Xilanases são enzimas produzidas por uma grande variedade de organismos, que inclui de bactérias e protozoários a fungos e artrópodes, e tipicamente secretadas para o meio extracelular, onde atuam na hidrólise de xilanos. Xilanos são polissacarídeos cujo monômero é xilose, um açúcar de cinco carbonos representado na figura 1.4, e são encontrados na parede celular de plantas como principal componente da hemicelulose (27). Hemicelulose, celulose (um polímero linear uniforme de glicose) e lignina (um polímero fenólico heterogêneo e sem uma estrutura repetitiva bem definida) juntos são os principais componentes da parede celular secundária de plantas, encabeçando a lista dos polímeros orgânicos mais abundantes na Terra. A ubiquidade dos substratos contribui para o contínuo interesse do ponto de vista industrial no isolamento e classificação de xilanases, em particular daquelas que apresentam características extremofílicas, tais como manutenção da atividade sob temperaturas elevadas ou meios altamente ácidos.

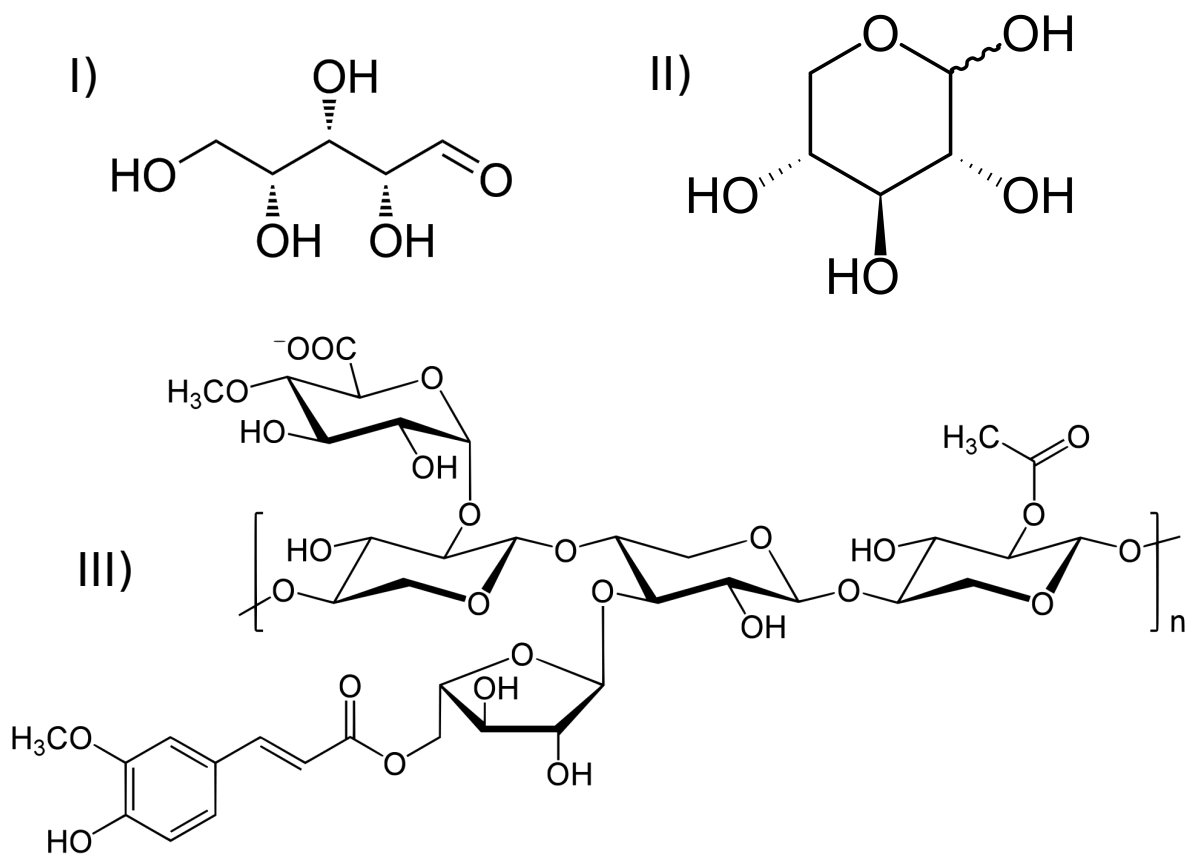

Figura 1.4 – *(I) Xilose, uma pentose, em configuração linear (II) Xilose em configuração de anel (III) Representação estrutural possível do xilano, polímero heterogêneo e ramificado de xilose.*

Xilanases catalisam a clivagem do esqueleto β-1,4 em xilanos, liberando pequenos fragmentos de composição heterogênea denominados xilo-oligômeros. A hidrólise completa de xilanos depende de uma variedade de enzimas, devido à sua complexidade, e sistemas xilanolíticos completos são tipicamente encontrados em micro-organismos que decompõem material vegetal (27).

O termo xilanase, por si, implica tão somente atividade catalítica em relação ao substrato xilano; contudo, esta capacidade é demonstrada por um grande número de enzimas, às quais está associado um número significativo de tipos de enovelamento diferentes, bem como um número significativo de mecanismos de atuação diferentes. O estudo das xilanases, então, passa necessariamente por uma etapa de classificação.

A clivagem do xilano é a quebra de uma ligação glicosídica por introdução de uma molécula de H_2O. Xilanases, são, portanto, glicosil-hidrolases. Segundo o sistema numérico EC de classificação de enzimas, baseado no tipo de reação catalisada, o grupo das glicosil hidrolases (EC 3.2.1.x) que atua na endohidrólise de ligações 1,4-β-D-xilosídicas em xilanos é o grupo EC 3.2.1.8 (28). Contudo, de um ponto de vista estrutural, existem 96 famílias de glicosil-

hidrolases, das quais seis contém enzimas com atividade xilanolítica (27). Destas seis, duas são majoritariamente relevantes: a família 10 e a família 11.

Segundo Collins *et al.* em (27), "[A família 10] consiste de endo-1,4-β-xilanases (EC 3.2.1.8), endo-1,3-β-xilanases (EC 3.2.1.32) e celobiohidrolases (EC 3.2.1.91) [...] Membros desta família tipicamente tem grande massa molecular, baixo ponto isoelétrico e apresentam um enovelamento do tipo barril $(\alpha/\beta)_8$". Em (29), reporta-se para a família 10 um tamanho médio de 316 aminoácidos, com um percentual de identidade médio de 43%, resultando em uma massa molecular média em torno de 35 KDa.

Há evidências de que as xilanases desta família apresentam atividade catalítica também em substratos celulósicos pequenos. Em contraste, a família 11 é caracterizada pela alta especificidade e pelo enovelamento do tipo *β-jelly-roll*, com baixo peso molecular; reporta-se para a mesma em (29) um tamanho médio de 185 aminoácidos e um percentual de identidade médio de 50%, com massa molecular média de 21 KDa. Um exemplar desta família pode ser encontrado na figura 1.5. Discorreremos brevemente sobre esse tipo de enovelamento e sobre as razões intrigantes pelas quais as xilanases da família 11 são os alvos deste trabalho a seguir.

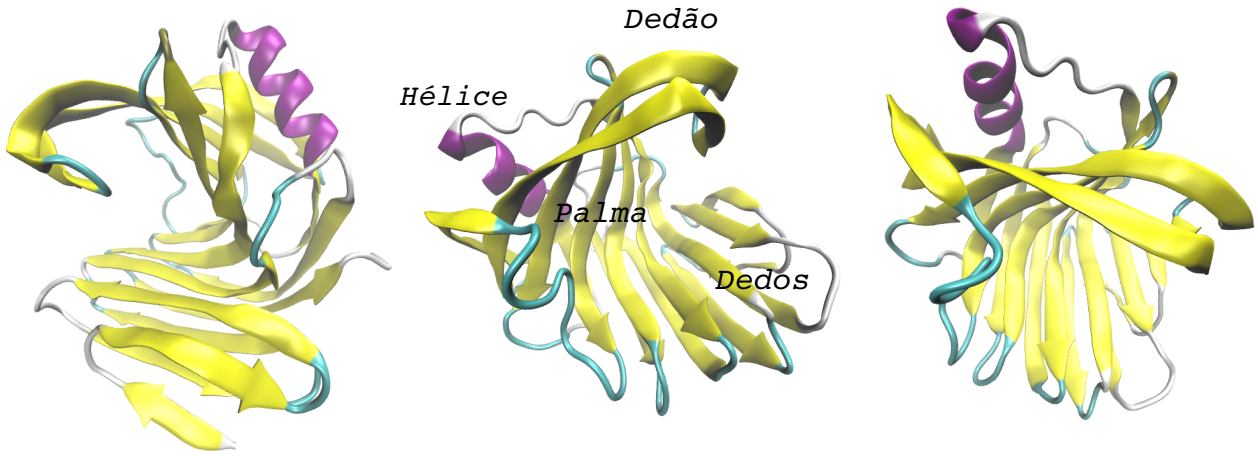

Figura 1.5 – *Visões de uma xilanase da família 11. É possível visualizar o par de fitas beta associados ao "dedão", cujo volta contém um resíduo glutamato que participa do sítio catalítico, e o par de folhas beta dobrado sobre si mesmo associado aos "dedos" e à "palma" da mão, que corresponde ao núcleo hidrofóbico. Figuras geradas com o programa VMD (1, 7).*

O enovelamento do tipo β-jelly-roll é um enovelamento predominantemente composto por fitas beta antiparalelas arranjadas em duas folhas que se empacotam e se dobram uma sobre a outra, envolvendo a região do sítio catalítico como uma mão direita parcialmente fechada. De fato, metáforas "anatômicas" são utilizadas para identificar os componentes da estrutura. Na ponta do dedão há uma volta que contém resíduos catalíticos (figura 1.6); resíduos que interagem com o substrato para mantê-lo na posição adequada também são encontrados na

palma e nos dedos. O núcleo hidrofóbico corresponde às faces internas das folhas beta, e a hélice alfa presente se empacota nas costas da palma.

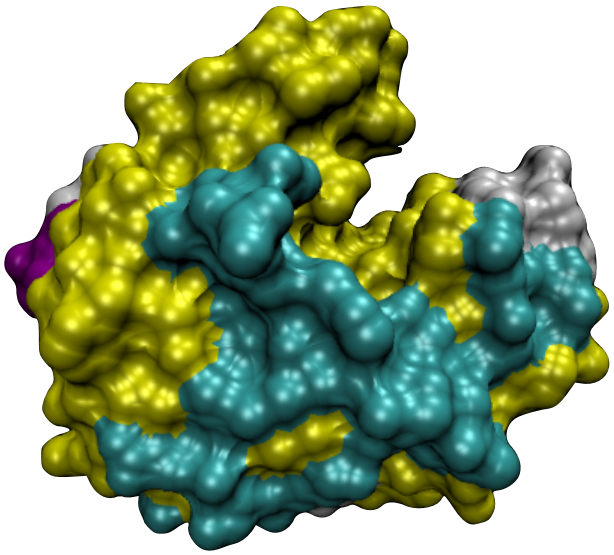

Figura 1.6 – *Visão de uma xilanase representada como sua superfície acessível ao solvente, com ênfase na fenda catalítica, paralela à palma e envolvida pelos dedos e pelo dedão, que se aproximam colocando os resíduos catalíticos na posição apropriada. Elementos de estrutura secundária estão colorizados segundo o mesmo padrão da figura 1.5, com folhas beta em amarelo, hélices alfa em roxo e* loops *em azul. Figura gerada com o programa VMD (1, 7, 8).*

Xilanases termofílicas, ou seja, capazes de manter atividade catalítica em temperaturas de 60°C a 80°C ou superiores, são encontradas principalmente nas famílias 10 e 11. Estas representam um nicho de interesse comercial[‡], pois muitos dos processos onde as mesmas são empregadas devem ocorrer em altas temperaturas, seja por razões físico-químicas tais como a viscosidade do meio reacional ou a solubilidade do substrato, ou por razões pragmáticas tais como o custo associado a etapas de resfriamento (27). Nota-se imediatamente que xilanases mesofílicas e termofílicas são estruturalmente análogas e muito parecidas; sugere-se que o aumento da estabilidade em relação à temperatura seja devido a várias modificações sutis atuando em conjunto. Em xilanases da família 10, os mecanismos moleculares responsáveis por esta adaptação são bem determinados, e correspondem aos mecanismos encontrados em proteínas termofílicas em geral. Estes incluem um aumento no número de interações não-covalentes dentro da estrutura como pontes salinas e ligações de hidrogênio, aumento no número de pontes dissulfeto e uma melhora no empacotamento do *core*, entre outros. De modo geral, estas modificações podem ser resumidas como um adensamento da rede de contatos inter-resíduos.

[‡]Na verdade, enzimas extremofílicas de toda sorte tais como termófilas, psicrófilas, acidófilas e alcalófilas, ou mesmo enzimas que combinam mais de uma destas características são frequentemente desejadas em aplicações industriais, dado que as condições fisiológicas, às quais enzimas são tipicamente adaptadas, são demasiado delicadas para a maior parte dos processos industriais.

Na família 11, de modo geral, estes mecanismos nem sempre são observados. De fato, em alguns dos exemplares termofílicos da família 11, os mecanismos responsáveis pela termoestabilidade não são elucidados, ou, quando são, induzem conclusões que vão na direção oposta do consenso da literatura (30). O conjunto de proteínas estudadas neste trabalho compreende alguns destes exemplares. Submeteremos as mesmas a simulações de Dinâmica Molecular para observar seu comportamento microscópico, e posteriormente modelaremos a rede de contatos inter-resíduos com o objetivo de evidenciar possíveis diferenças entre elas.

CAPÍTULO 2

Dinâmica Molecular em Proteínas

2.1 O Computador como Ferramenta Experimental

O autor incumbido de introduzir dinâmica molecular se vê diante de um problema mal proposto; por um lado, porque Dinâmica Molecular se refere a um conjunto grande e heterogêneo de técnicas, algoritmos e aplicações, e, por outro lado, porque o conceito e a história das simulações computacionais se confundem com a história do computador, que se inicia pouco mais de duas gerações atrás. Não por coincidência, a primeira estrutura de proteína resolvida por difração de raios-X provém da mesma época (17); tratar os dados coletados do padrão de intensidades e produzir um modelo da densidade eletrônica manualmente é um processo extremamente laborioso, ainda que possível, e o desenvolvimento e popularização da tecnologia de computação certamente impulsionaram o crescimento das bases de dados de estruturas resolvidas*. De sua origem até o presente, aquele que se propuser a relatar a história da técnica de difração de raios-X não encontrará escassez de resultados importantes.

Aquele que se propõe a relatar a história da dinâmica molecular encontra panorama semelhante. Contudo, a análise da origem e do desenvolvimento das simulações computacionais não pode ser feita exclusivamente sob o prisma da análise de uma técnica experimental, e nem somente sob o prisma de um desenvolvimento teórico/matemático. Por conta de sua natureza, a Dinâmica Molecular e as técnicas análogas ocupam uma posição intermediária entre o teórico e o experimental.

Parece injustificado atribuir a uma *simulação* o caráter de experimento. Contudo, quando olhamos para os modelos que hoje melhor descrevem o universo, as interações e o comportamento das partículas, tais como a Mecânica Quântica expressa na forma da Equação de

*Uma breve pesquisa da literatura revela que desde a década de 1960 a resolução de estruturas de proteínas por difração de raios-X é assistida por computadores, com destaque para os trabalhos pioneiros do laboratório de M. Rossman (por exemplo, (31)). É dado por certo pelo autor que a partir da década de 1970 computadores já eram considerados parte integrante do aparato experimental.

Schrödinger (ou mesmo as Leis de Newton), notamos que estes têm um péssimo costume: o costume de não apresentar solução analítica a não ser em casos muito simples. A Equação de Schrödinger prevê elegantemente os níveis de energia do átomo de Hidrogênio, mas se torna um pesadelo algébrico quando outro elétron é adicionado para caracterizar o Hélio. Nenhuma quantidade de dados experimentais pode derrubar um modelo, quando o mesmo não pode fornecer previsões para serem testadas. Uma alternativa é a resolução de versões aproximadas do modelo, como desprezar a interação elétron-elétron que torna o Hamiltoniano do átomo de Hélio complicado demais para o lápis e papel[†]. Mas muitas vezes, estas aproximações (como no caso do Hélio) produzem resultados absolutamente espúrios, os quais não podem ser atribuídos a uma falha do modelo, mas sim à matemática limitada disponível para o modelador. Com um computador, é possível obter soluções numéricas para sistemas complicados como este com precisão (em princípio) tão grande quanto necessária, o que possibilita a comparação com o experimento. Deste modo, o computador permite a aplicação de primeiros princípios a sistemas grandes, complexos, ou de outro modo intratáveis, fornecendo ao teórico ferramentas para aceitar ou descartar um modelo antes de ir ao laboratório (32).

Por outro lado, o repetido refinamento baseado em dados experimentais permite atingir um nível de considerável confiança no modelo utilizado. Um modelo computacional que, baseado em primeiros princípios, consistentemente prediga o resultado de uma reação química em água a 300K e 1 atm pode ser utilizado para prever o resultado da mesma quando submetida a altas pressões no fundo do oceano ou temperaturas extremas em fontes termais, ou sob outras condições experimentalmente inconvenientes, caras ou perigosas. Este tipo de procedimento é muitas vezes chamado de experimento *in silico*, em contraste aos já estabelecidos *in vitro* e *in vivo* (32).

Neste trabalho, utilizamos dinâmica molecular nestes dois contextos. Como ferramenta "experimental", para validar o modelo de rede proposto, e como ferramenta "teórica", para propor hipóteses que posteriormente devem ser comparadas a dados experimentais. Trataremos dos detalhes a seguir.

2.2 Dinâmica Molecular

Uma simulação de dinâmica molecular consiste na resolução iterativa das equações de movimento para um conjunto de átomos interagentes. A hipótese mestra por detrás do método

[†]Na verdade, a resolução do átomo de Hidrogênio já inclui uma aproximação, a separação dos movimentos do núcleo e do elétron ou aproximação de Born-Oppenheimer. Felizmente, graças à massa muito grande do próton em relação ao elétron, este é um exemplo de aproximação que não introduz um erro grande em relação ao valor experimental.

é a de que, dado um volume contendo um número finito e conhecido de partículas interagentes, sujeito a condições iniciais suficientemente bem caracterizadas, a evolução temporal desse sistema pode ser predita por meio da resolução das equações de movimento de cada partícula. O leitor astuto notará que esta é, de fato, a hipótese mestra por detrás da Física como um todo. A diferença reside em detalhes metodológicos, mas, em sua essência, dinâmica molecular é a simulação do mundo no computador.

No caso geral, a equação de movimento de cada partícula contém termos que dependem da posição e/ou velocidade de todas as outras partículas, de modo que a resolução analítica não é uma opção, e as equações de movimento são, ao invés disso, integradas numericamente no tempo. Esta é a primeira dentre muitas aproximações que devem ser feitas para que o método se torne capaz de tratar sistemas cujo tamanho e complexidade os aproxime daqueles estudados por químicos e biólogos.

A forma da interação que acopla os movimentos das partículas é também uma aproximação, e merece atenção especial. A rigor, a evolução temporal de um sistema de dimensões atômicas é predita pela resolução da Equação de Schrödinger dependente do tempo. Contudo, tal descrição é excessivamente complicada, e o termo "dinâmica molecular" normalmente é reservado para métodos que ignoram a descrição quântica da matéria. Ao invés disso, átomos são considerados cargas pontuais cujo movimento é regido por equações Newtonianas (33). A força que acelera cada partícula é o gradiente de uma energia potencial que depende da posição de todas as outras, tal como:

$$m\ddot{\vec{r}} = -\nabla U(\vec{r_1}, \vec{r_2}, \ldots, \vec{r_N}) \tag{2.2.1}$$

O formato da função energia potencial é um compromisso entre dois interesses primários do método: precisão e facilidade computacional. O cálculo das forças que atuam sobre as partículas é o processo mais custoso em uma simulação de dinâmica molecular (32, 33), e nem sempre uma solução ótima para uma classe de sistemas é ideal para todos os outros. A função energia potencial utilizada nas simulações deste trabalho é a soma das contribuições de várias fontes, obedecendo ao seguinte formato (33, 34):

$$U = U_{\text{ligação}} + U_{\text{ângulo}} + U_{\text{diedro}} + U_{\text{LJ}} + U_{\text{Coulomb}} \tag{2.2.2}$$

Imediatamente, podemos dividir os termos em interações correspondentes a partículas

ligadas e a partículas não-ligadas. Quando aplicada a sistemas microscópios biológicos, nos quais as moléculas de interesse são majoritariamente biopolímeros de cadeia longa, a descrição correta das interações entre átomos ligados é naturalmente significante para a reprodução dos comportamentos experimentais. Aqui, as interações entre átomos ligados são separadas em termos correspondendo a ligações covalentes, que envolvem dois átomos, ângulos, que envolvem três átomos, e dois termos para diedros, envolvendo quatro átomos numa cadeia linear ou ramificada. Tem-se:

$$U_{\text{ligação}} = \sum_{\text{ligação i}} k_i^{\text{ligação}} (r_i - r_{0i})^2 \tag{2.2.3}$$

$$U_{\text{ângulo}} = \sum_{\text{ângulo i}} k_i^{\text{ângulo}} (\theta_i - \theta_{0i})^2 \tag{2.2.4}$$

$$U_{\text{diedro}} = \sum_{\text{diedro i}} \begin{cases} k_i^{\text{diedro}} [1 + \cos(n_i \phi_i - \gamma_i)], & n_i \neq 0 \\ k_i^{\text{diedro}} (\omega_i - \omega_{0i})^2, & n_i = 0 \end{cases} \tag{2.2.5}$$

Onde, em cada soma, o índice i percorre as listas de distâncias r entre pares de átomos, de ângulos θ entre ligações e de ângulos φ e ω característicos dos diedros (n e γ indicam as posições dos possíveis múltiplos mínimos das energias de torção). As constantes k_i (*parâmetros*) caracterizam a energia destas interações.

Os outros dois termos correspondem às interações entre átomos que não são considerados ligados, por fazerem parte de moléculas diferentes ou por estarem longe demais um do outro na mesma cadeia. Contemplam-se interações de Van der Waals, modeladas por um potencial de Lennard-Jones, entre todos os pares de átomos, e interações de Coulomb (eletrostáticas) entre todos os pares de átomos carregados.

$$U_{\text{LJ}} = \sum_i \sum_{j>i} 4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \tag{2.2.6}$$

$$U_{\text{Coulomb}} = \sum_i \sum_{j>i} \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} \tag{2.2.7}$$

Onde os índices i e j percorrem a lista de átomos, r representa distância, q representa carga elétrica e ε_0 e π são constantes fundamentais. No potencial de Lennard-Jones (2.2.6), σ relaciona-se à distância de equilíbrio e ε à profundidade do mínimo de energia.

As interações não-ligadas, em particular, são muito mais numerosas e, portanto, computacionalmente exigentes. Estratégias para diminuir o tempo gasto no cálculo destas interações incluem a aplicação de *cutoffs* espaciais para o cálculo das interações de Van der Waals, efetivamente diminuindo o número de pares de partículas que "se enxergam", e a soma no espaço recíproco ou *soma de Ewald* (por meio da aplicação da transformada de Fourier) para o cálculo das interações eletrostáticas, quando a simulação é realizada com condições de contorno periódicas (33).

Cada um dos termos descritos envolve coordenadas de componentes do sistema e constantes fundamentais experimentais, mas, principalmente, envolve conjuntos de parâmetros extensivamente ajustados para reproduzir cálculos teóricos químico-quânticos em nível mais fundamental e para reproduzir resultados experimentais quando utilizados para simular sistemas já bem caracterizados. A confiabilidade dos resultados de uma simulação é consequência da precisão da parametrização empregada.

O método aplicado para integrar as equações de movimento é, naturalmente, outra aproximação. O parâmetro crucial é o tamanho do passo temporal, que influencia a ordem de grandeza do tempo de evolução do sistema que pode ser simulado no mesmo tempo de CPU, a ordem de grandeza do erro acumulado em cada passo, e os tipos de movimentos moleculares que podem ser investigados. Mais sutilmente, o método de integração pode influenciar a evolução do sistema resultando em um viés na exploração do espaço conformacional. Tipicamente, é empregado o algoritmo Velocity-Verlet, simplético, que requer apenas um cálculo de forças por passo (33).

Outros aspectos da simulação, tais como o controle de temperatura e/ou de pressão do sistema, introduzem complicações adicionais. Como consequência das aproximações empregadas, tem-se que a interpretação determinística da evolução temporal obtida deve ser realizada com cautela - cada simulação produz uma trajetória no espaço conformacional que pode não corresponder precisamente à evolução real do sistema sob as mesmas condições, ainda que a divergência das trajetórias mantenha-se pequena e limitada em tempos de simulação comuns (32, 33). Contudo, quando as trajetórias geradas são encaradas como um conjunto ou *ensemble* de conformações, as propriedades medidas aproximam-se estatisticamente das propriedades do sistema real (32), e as limitações do poder preditivo das simulações em relação à evolução do sistema são provenientes das simplificações e não inerentes ao método. O emprego de

estatística permite, deste modo, a extração de medidas confiáveis de simulações de dinâmica molecular mesmo em face das aproximações inescapáveis frente à complexidade dos sistemas estudados.

Existe, ainda, uma miríade de técnicas que estendem as simulações convencionais em direções que o experimento não pode acompanhar, mas que não fogem ao alcance das investigações teóricas. Exemplos incluem a aplicação de forças externas para guiar a evolução do sistema na direção desejada, como na técnica de dinâmica molecular dirigida (SMD), ou o estudo das consequências energéticas de transformações "alquímicas"[‡] como a mudança da natureza de átomos ou mesmo segmentos inteiros de moléculas em um sistema, como nos cálculos de perturbação de energia livre (FEP) (33). A técnica descrita na próxima seção é uma extensão das simulações convencionais cujo objetivo é acompanhar o fluxo de calor dentro de uma biomolécula, e foi extensivamente utilizada neste trabalho.

2.3 Difusão Térmica Anisotrópica

De um ponto de vista teórico, sabe-se que para proteínas globulares a existência de canais preferenciais de propagação da energia vibracional é uma consequência inevitável de sua geometria[§], e que estes canais não tem necessariamente relação com sua atividade catalítica. Experimentalmente, contudo, a transferência de energia e as correlações de movimentos entre diferentes segmentos de uma cadeia proteica obedecem a mecanismos difíceis de caracterizar. Ainda assim, é sabido que proteínas exibem, por exemplo, movimentos atômicos concertados globais, lentos e de grande amplitude, que tipicamente correspondem a mudanças conformacionais essenciais para suas funções biológicas (36). A partição da energia entre diferentes estados vibracionais e o transporte da energia ao longo de canais preferenciais são descrições alternativas de uma mesma classe de processos, denominados na literatura processos de relaxação de energia vibracional (37).

Fenômenos capazes de induzir um excesso localizado de energia vibracional incluem o curso de reações químicas no sítio catalítico, a ligação de pequenas moléculas (ligantes), choque térmico e a absorção de radiação, e a eficiência com que uma proteína dissipa a energia proveniente dessas excitações pode estar relacionada à sua capacidade de manter atividade

[‡]Para o Químico Teórico Computacional, transformar Chumbo em Ouro é absolutamente trivial. Este fato ilustra e justifica o apelo dos computadores como ferramentas de materialização da imaginação dos pesquisadores. Ilustra também o grau de diversão inerente à simulações de dinâmica molecular, as quais foram primordialmente apresentadas ao autor deste texto como "Video-games para Cientistas".

[§]A razão por detrás desta implicação reside no perfil de conectividade entre os resíduos, que para proteínas globulares se assemelha a um *cluster* percolado tridimensional (35). Evidências para esta observação são fornecidas também por este trabalho.

em altas temperaturas. Por outro lado, proteínas responsáveis pela captação de radiação eletromagnética em complexos fotossintéticos exibem a capacidade de evitar a dissipação por tempos longos o suficiente para que a energia absorvida seja direcionada para sítios ativos secundários (38).

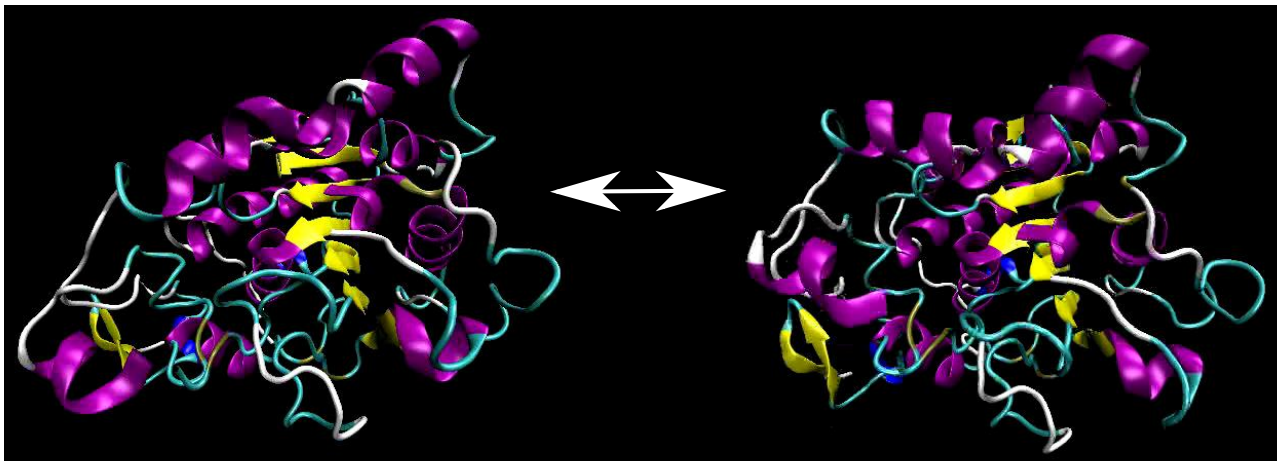

Figura 2.1 – *Ilustração de movimentos moleculares de larga escala em uma lipase, calculados por análise de modos normais. (9).*

Uma técnica não-dinâmica frequentemente utilizada para prever movimentos moleculares é a análise dos modos normais de vibração. O cálculo dos modos normais é a solução de uma equação de autovalores, que envolve as coordenadas atômicas e os potenciais de interação e produz como resultado um conjunto de funções vetoriais periódicas, que representam, para cada modo, a direção e frequência dos movimentos atômicos. Em muitos casos, o modo de mais baixa frequência captura as características dos principais movimentos moleculares e identifica as regiões mais flexíveis da proteína, ainda que esse modo possa não ser fisicamente realizável devido ao ambiente altamente amortecido do solvente em que a mesma reside (36). De fato, uma limitação comum às técnicas baseadas em modos normais é o negligenciamento da influência do solvente. Além disso, o cálculo de modos normais não fornece, por construção, hipóteses sobre como a proteína reage a um estímulo localizado, cujo efeito é caracterizado por movimentos locais, não-coordenados e corresponde a modos de alta frequência.

Aqui, utilizamos um método dinâmico de não-equilíbrio para investigar movimentos moleculares e em particular a transferência de calor em proteínas, denominado método de Difusão Térmica Anisotrópica (ATD), introduzido por Ota & Agard em (39) e estendido e sistematizado por Martínez *et al.* em (38). O ATD é uma extensão de simulações convencionais de dinâmica molecular projetado para acompanhar quantitativamente o fluxo de energia vibracional ao longo da cadeia polimérica em resposta a um estímulo local.

O protocolo geral de um experimento de ATD inicia-se com a produção de uma simulação

longa de equilíbrio, com duração da ordem de algumas dezenas de nanossegundos, e tem o seguinte formato:

1. Retiram-se, da trajetória da simulação, múltiplas estruturas, correspondendo a instantes temporais espaçados entre si de maneira a evitar correlações espúrias.

2. Para cada *frame* retirado: inicia-se uma simulação curta de equilibração (da ordem de 5 ps) a partir da estrutura retirada, obtendo-se nova termalização a 10K. De posse da estrutura artificialmente resfriada, para cada um de seus resíduos de aminoácido, faz-se:

3. O resíduo selecionado é aquecido e mantido a 300K, acoplado a um banho térmico independente do resto da estrutura, e o banho térmico frio é removido. A temperatura da estrutura como um todo e de cada resíduo em particular é monitorada durante um tempo curto (30 ps) de simulação.

4. As temperaturas finais são tabuladas e reportadas, destacando particularmente as relações *Resíduo Aquecido x Temperatura Final da Proteína* na forma de gráfico e *Resíduo Aquecido x Temperatura Final de cada Resíduo* na forma de um Mapa de Difusão Térmica. O ATD é um protocolo de não-equilíbrio, de forma que os resultados obtidos devem ser analisados sempre tendo em vista o mesmo tempo fixo de simulação.

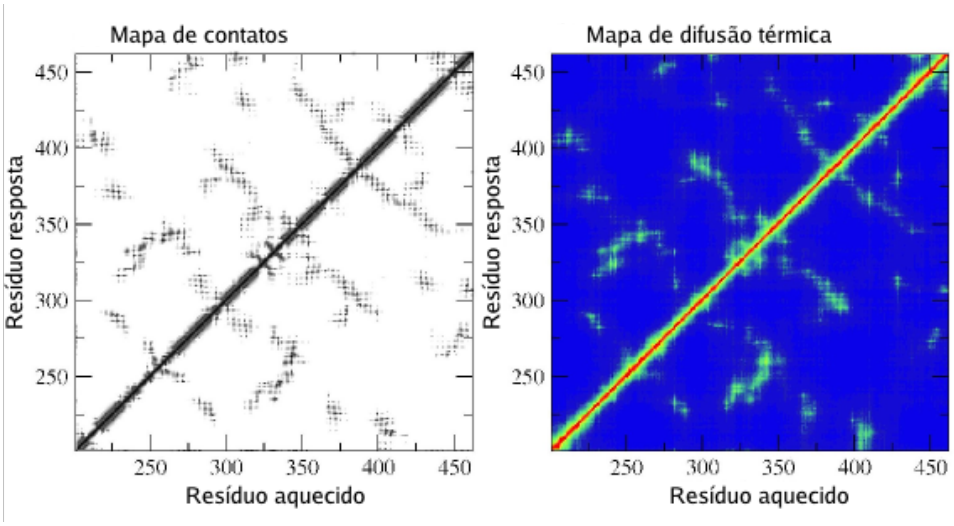

Figura 2.2 – *Exemplo de Mapa de Difusão Térmica e de mapa de contatos para um experimento de ATD. O Mapa de Difusão Térmica tabula a resposta térmica de cada resíduo em relação à identidade do resíduo aquecido. Reproduzido com autorização de (10).*

Naturalmente, o Mapa de Difusão Térmica (figura 2.2) reproduz, de maneira aproximada, o mapa de contatos da proteína. Cada resíduo aquecido transfere energia principalmente para seus vizinhos próximos, e o resultado é um aumento da temperatura dos resíduos com quem o

mesmo faz contato, tanto ao longo da cadeia principal quanto via interações não-covalentes. Observações particularmente interessantes em um experimento de ATD são, deste modo, os desvios dos caminhos esperados dado o mapa de contatos, e a conseqüente revelação de pares de resíduos cujos movimentos são correlacionados mesmo na ausência de contato próximo entre os mesmos (vide, por exemplo, (39)).

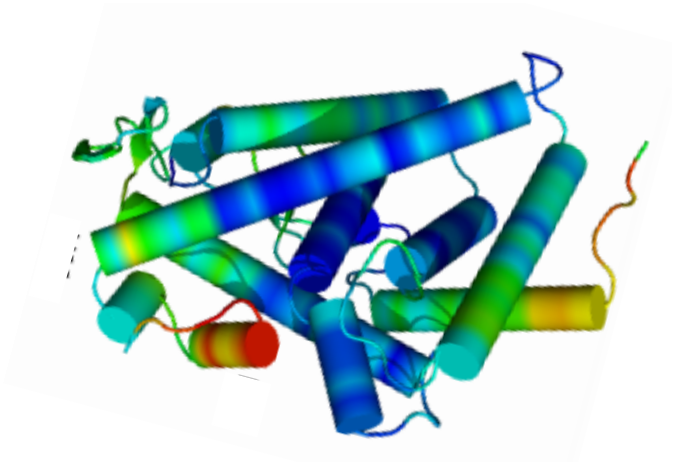

Figura 2.3 – *Exemplo de proteína com resíduos colorizados para denotar capacidade de transferir calor para a estrutura ("bons difusores"), aumentando a temperatura final da proteína num experimento de ATD. Resíduos melhores difusores estão representados em cores mais quentes. Reproduzido com autorização de (10).*

A análise das temperaturas finais da proteína (figura 2.3), analogamente, identifica resíduos com aptidão pronunciada para transferir rapidamente sua energia para o resto da estrutura, aumentando sua temperatura. Identificamos, por simplicidade, estes resíduos como "bons difusores" de calor, e Martínez *et al.* ressaltam que não se observam correlações sistemáticas óbvias entre a carga, massa ou características da estrutura química do resíduo e sua tendência a ser (ou não) bom difusor de calor. Existem, contudo, evidências experimentais de que esta característica é associada a importância funcional, pois sua substituição em experimentos de mutagênese sítio-dirigida pode resultar na anulação da atividade enzimática (38). Deste modo, predizer resíduos bons difusores a partir de dados de sequência ou estrutura é um problema teórico intrigante, e possivelmente relacionado à identificação de resíduos funcionais ou do sítio catalítico.

É hipótese do autor, frente aos fatos apresentados, de que é possível prever resíduos bons difusores analisando-se a topologia da estrutura proteica, por meio de um modelo de rede que substitui resíduos inteiros por *nós* e interações covalentes e não-covalentes por *ligações*, numa simplificação que retém e evidencia somente as relações de conectividade entre resíduos. Este modelo e sua construção são descritos no capítulo que segue.

CAPÍTULO 3

Modelagem de Proteínas como Redes Complexas

3.1 Introdução à Teoria de Redes

Em sua acepção mais simples, uma *rede* é um modelo compreendendo um conjunto de partes que interagem entre si de alguma forma (40). Por esta definição, não é imediata (ou necessária) a distinção entre rede e *sistema* ("Combinação de partes reunidas para concorrerem para um resultado, ou de modo a formarem um conjunto" (41)), apesar de uma rede não apresentar, a princípio, uma finalidade. De fato, a metáfora de rede é útil para praticamente qualquer situação (ou escala de observação) em que as interações entre as partes sejam mais relevantes que as propriedades individuais das partes em si. Objetos de estudo com tal característica são encontrados nas mais diversas disciplinas científicas - podemos enumerar os seguintes exemplos, retirados de Newman em (40), numa lista que não pretende ser exaustiva de objetos do mundo real que são adequadamente modelados por redes:

- Redes sociais de conexões entre indivíduos;
- Redes de relações de negócios entre empresas;
- Redes de citações entre artigos científicos;
- Redes de co-aparecimento ligando pessoas quando as mesmas são mencionadas num mesmo artigo de jornal;
- Redes neurais responsáveis pelo processamento cerebral;
- Redes metabólicas de inibição e ativação de genes;
- Circuitos elétricos e eletrônicos;
- Sistemas de distribuição tais como o sistema circulatório e a rede de transmissão elétrica;

- Sistemas de entrega de cartas ou encomendas;
- Malhas rodoviárias ou aeroviárias;
- Cadeias alimentares;
- A Internet, rede física de computadores ligados entre si, que é a base sobre a qual se implementa a World Wide Web;
- A World Wide Web, rede de documentos contendo *hyperlinks* que direcionam o leitor para outros documentos;
- Redes baseadas em dicionários e *thesaurus*, ligando palavras pela similaridade das idéias que indicam, ilustrando a própria estrutura da linguagem e/ou dos modelos mentais que a representam.

Felizmente, apesar da extensa aplicação de modelos de redes em áreas diversas, a terminologia empregada no seu estudo é razoavelmente uniforme. O estudo das redes é uma disciplina da matemática discreta, onde são denominadas *grafos*. Exporemos agora alguns fundamentos da teoria dos grafos.

Grafos são caracterizados por duas entidades: as partes que interagem, denominadas *nós* ou *vértices*, e as interações entre elas, denominadas *ligações* ou *arestas*. Ordinariamente, grafos são representados como conjuntos de pontos, círculos ou outras formas geométricas, os nós, ligados por linhas, retas ou não, que representam as arestas (42), e um exemplo é apresentado na figura 3.1. Esta representação é devida principalmente a Moreno*, em trabalhos publicados na década de 1930 tratando da visualização inovadora de redes de relacionamentos em grupos sociais, tais como grupos de crianças de idade pré-escolar em salas de aula (43). Além de intuitiva, a representação gráfica de pontos e linhas permite a aplicação direta de um instrumento poderoso de análise: o olho humano. A escolha adequada da posição dos nós e arestas evidencia propriedades estruturais da rede como um todo e de nós de interesse em particular, e muitos dos avanços iniciais do estudo de redes foram obtidos deste modo (40).

Análises mais profundas ou quantitativas, de modo geral, requerem o emprego da formalização matemática adequada. As seguintes definições e algumas de suas consequências são oriundas de Steen em (42), cujo excelente texto informa a maior parte desta seção (exceto quando explicitamente indicado), e foram reproduzidas *verbatim*.

*Jacob Levy Moreno (1889-1974) foi um psiquiatra e cientista social influente, nascido na Romênia mas atuante principalmente nos EUA. Moreno é considerado um dos pioneiros da análise de redes sociais, e métodos desta provenientes ainda hoje informam o avanço da teoria de grafos, num caso curioso de apropriação.

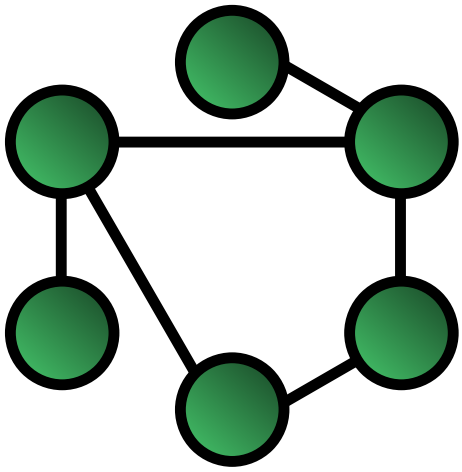

Figura 3.1 – *Exemplo de grafo, com seis nós (círculos em verde) e seis arestas. (11).*

Definição 1 *Um* ***grafo*** G *consiste num conjunto* V *de vértices e um conjunto* E *de arestas, para o qual escrevemos* $G = (V, E)$. *Diz-se que cada aresta* $e \in E$ ***une*** *dois vértices. Se* e *une* $u, v \in V$, *denota-se* $e = \langle u, v \rangle$. *Os vértices* u *e* v *são, então, ditos* ***adjacentes****.*

Em particular, $V(G)$ denota o conjunto de vértices de G e $E(G)$ denota o conjunto de arestas de G. Não se faz, aqui, distinção entre $\langle u, v \rangle$ e $\langle v, u \rangle$. Os grafos tratados neste trabalho não possuem arestas do tipo $\langle u, u \rangle$ (**loops**), e não admitem arestas múltiplas unindo os mesmos pares de vértices (pois $E(G)$ é um *conjunto*). Grafos que obedecem a estas duas condições são chamados **simples**.

Afirmamos que a metáfora de rede é aplicável sempre que a maneira como os agentes se ligam é mais relevante do que as propriedades individuais de cada agente. Neste caso, todos os agentes são representados pela mesma entidade equivalente, o nó. Contudo, a metáfora de rede seria inútil se dela não pudéssemos extrair informações além das que já fornecemos durante sua construção. Não é o caso; uma vez que a rede é estabelecida, podemos comparar os nós entre si, classificando-os, por exemplo, pela quantidade de ligações que fazem, ou pelo grau de alteração da rede que sua remoção provocaria.

A classificação do nó pelo seu número de vizinhos é um exemplo ilustrativo. Definimos:

Definição 2 *Para um grafo* G *e um vértice* $v \in V(G)$, *o conjunto* $N(v)$ *de* ***vizinhos*** *de* v *é o conjunto de vértices adjacentes a* v, *dado por:*

$$N(v) = \{u \in V(G) \mid u \neq v, \exists e \in E(G) : e = \langle u, v \rangle\}$$

O ***grau*** *de um vértice* v, *denotado* $\delta(v)$, *é o número de vizinhos de* v, *ou seja,* $|N(v)|$.

É razoável supor que o grau de um nó carregue uma medida grosseira da sua importância. Se a rede representar uma malha aeroviária, por exemplo, um nó de alto grau pode corresponder a um aeroporto do qual partem e chegam vôos de muitos outros aeroportos, com todos os custos logísticos associados. Se, por outro lado, a rede representar relações de amizade em um determinado grupo social, um nó de alto grau pode corresponder a uma pessoa com muitos amigos. Tal pessoa ocupa, por razões óbvias, posição privilegiada dentro deste grupo.

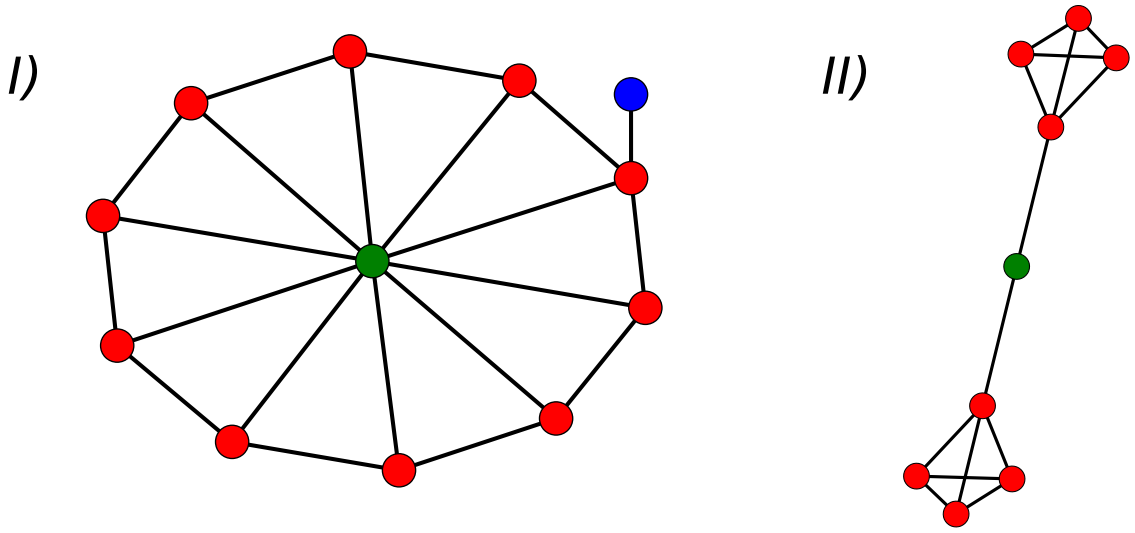

Figura 3.2 – *Ilustração de dois grafos. (I) O nó representado em verde reside numa posição privilegiada, exibindo grau muito maior que todos os outros. O nó em azul experimenta situação oposta. (II) O nó aqui representado em verde possui o menor grau da rede. Contudo, sua posição é crítica: sua remoção resultaria na obtenção de dois componentes desconectados.*

Ser altamente conectado, entretanto, não é a única propriedade que pode conferir importância a um nó. Caso o mesmo resida numa posição tal que parte significativa dos caminhos que conectam os diferentes nós da rede passem por ele, sua remoção ou inutilização poderá ser disruptiva para o transporte ao longo da rede. O conceito de *caminho* exige formalização:

Definição 3 *Seja G um grafo. Um* ***caminho*$-(v_0, vk)$** *em G é aqui definido como uma sequência alternante $[v_0, e_1, v_1, e_2, \ldots, v_{k-1}, e_k, v_k]$ de vértices e arestas em G tal que $e_i = \langle v_{i-1}, v_i \rangle$, em que todas os vértices e todas as arestas são distintos*[†]*. Dois vértices u e v em G são ditos* ***conectados*** *se existe um caminho$-(u, v)$ em G. G é dito* ***conectado*** *quando todos os pares de vértices em $V(G)$ são conectados.*

Definição 4 *Sejam u e v dois vértices conectados em um grafo G. O* ***comprimento*** *de um*

[†]Na literatura, um caminho pode, no caso geral, apresentar vértices e arestas repetidos. A definição aqui apresentada é reservada para um *caminho simples*. Todos os caminhos de interesse neste trabalho serão caminhos simples, e todos os grafos conectados.

caminho$-(u, v)$ *é o número de arestas contidas na sequência alternante que liga* u *a* v*. No caso geral,* u *e* v *são conectados por mais de um caminho. Definimos a* ***distância*** *entre* u *e* v *como o comprimento do menor caminho entre eles. Cada caminho distinto de comprimento mínimo entre* u *e* v *é denominado uma* ***geodésica*** *entre* u *e* v*.O* ***diâmetro*** *de um grafo* G *é a maior distância entre dois vértices em* G*.*

Assim, para um rede que represente o sistema de transmissão elétrica de uma cidade, um nó associado a uma torre que pertença a muitos caminhos poderá ocasionar grandes quedas no fornecimento de energia no caso de uma eventual falha. Um sistema bem projetado evita a existência de tais "gargalos" introduzindo *redundância* na rede de transmissão, isto é, fazendo com que vértices distantes sejam conectados por mais de um caminho. Uma rede que demonstra capacidade de permanecer conectada frente à remoção de vértices ou arestas é dita **robusta**. Em capítulos posteriores, investigaremos a robustez em redes que representam estruturas de proteínas, e sua possível relação com a termoestabilidade.

Convém ressaltar neste ponto que a análise "visual" de estruturas de grafos perde rapidamente sua viabilidade conforme aumenta o número de nós da rede considerada. Dado que muitos conjuntos de dados reais são adequadamente modelados por redes com milhares ou milhões de vértices, representações alternativas se fazem necessárias. Introduz-se, para tal fim, a representação de *matriz de adjacência*.

Definição 5 *Seja um grafo* G *com* N *vértices, tal que* $V(G) = \{v_1, \ldots, v_N\}$*. A* ***matriz de adjacência A*** *de* G *é uma matriz* $N \times N$ *tal que:*

$$\boldsymbol{A}_{ij} = |\{e \in E(G) \mid e = \langle v_i, v_j \rangle\}|$$

A matriz assim construída descreve completamente o grafo G*.*

Trivialmente, a cada posição i, j da matriz, associa-se o número de arestas entre v_i e v_j. Em grafos simples tais como os tratados neste trabalho, este número é sempre 1 ou 0, e a matriz assume caráter binário ou *lógico*. De posse da matriz de adjacência associada a uma rede, sua representação gráfica se torna opcional, pois suas propriedades de interesse podem ser calculadas diretamente. O grau de um nó é um exemplo trivial: é obtido imediatamente como a soma da linha ou coluna correspondente da matriz. Outras medidas, tais como o comprimento de geodésicas, requerem a aplicação de algoritmos mais complicados cuja descrição vai além do escopo deste trabalho. O leitor interessado é novamente referido a (42).

De todo modo, as propriedades relatadas até aqui com o propósito de distinguir os nós por sua importância capturam aspectos diversos de um mesmo conceito intuitivo, a idéia de

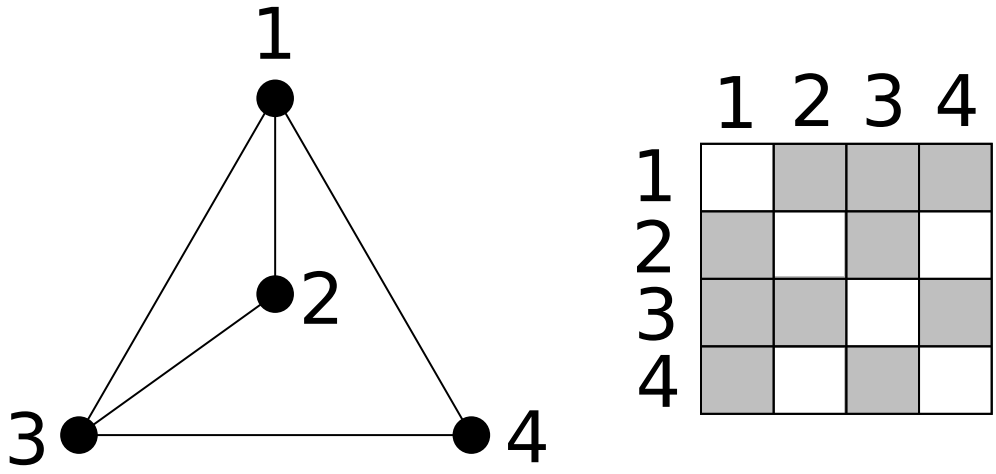

Figura 3.3 – *Exemplo de grafo e matriz de adjacência associada. (12).*

centralidade. Freeman, em um artigo de 1978, relata (em tradução livre) (44):

> "Assim, a idéia de centralidade está viva e bem, e sendo mobilizada em um largo escopo de aplicações que tende a aumentar. Todos concordam, ao que parece, que centralidade é um importante atributo estrutural de redes sociais. Todos admitem que a mesma seja fortemente relacionada a outras propriedades e processos importantes de grupos. Mas o consenso termina aí. Certamente não há unanimidade sobre exatamente o que *é* centralidade ou sobre seus fundamentos conceituais, e há pouca concordância sobre o procedimento correto para sua medida."

Em seguida, Freeman discute as diversas medidas até então reportadas na literatura, estabelecendo três definições diferentes para três medidas diferentes de centralidade, cada uma das quais capturando um aspecto distinto. As bases para duas destas já foram lançadas neste texto[‡]; a primeira remete ao conceito de *grau*:

Definição 6 *A **centralidade de grau** de um nó u em um grafo G contendo N nós é dada por:*

$$C_g(u) = \frac{\delta(u)}{N-1} \tag{3.1.1}$$

A normalização pelo número máximo possível de vizinhos de um nó em um grafo de tamanho N permite a comparação da centralidade de grau de nós provenientes de grafos

[‡]Apresentamos aqui as medidas com nomes traduzidos para o Português, segundo a tendência observada na literatura nacional (vide, por exemplo, (45)). O leitor encontrará também comumente as medidas com seus nomes originais: Degree Centrality, Betweenness Centrality e Closeness Centrality, respectivamente.

diferentes (alguns autores consideram que esta é a definição da *centralidade de grau relativa*, enquanto a centralidade de grau absoluta é simplesmente o grau do nó). A centralidade de grau relaciona o número de vizinhos de um nó ao seu potencial para atividade dentro da rede. É, entretanto, uma medida ingênua que não captura sutilezas da estrutura das ligações na rede, conforme ilustrado pela figura 3.2.

A segunda medida procura contornar essa limitação considerando não o número de vizinhos, mas o *posicionamento* do nó frente ao perfil de conexões da rede. Em particular, considerando todas as geodésicas possíveis entre todos os pares de nós da rede (vide definição 4), procuramos saber *de quantas destas* um determinado nó faz parte, ou, intuitivamente, quantos pares de nós têm sua comunicação prejudicada quando o nó de interesse é removido. Formalmente:

Definição 7 *A **centralidade de intermediação** de um nó u em um grafo G contendo N nós é dada por:*

$$C_b(u) = \frac{2}{(N^2 - 3N + 2)} \sum_{i}^{N} \sum_{j>i}^{N} \frac{g_{ij}(u)}{g_{ij}} \tag{3.1.2}$$

onde g_{ij} é o número total de geodésicas que ligam o nó i ao nó j, e $g_{ij}(u)$ é o número de geodésicas que ligam i a j passando por u.

A normalização, neste caso, é pelo máximo valor possível de centralidade de intermediação em um grafo de N nós (o leitor interessado em uma demonstração é direcionado a (44)), e permite a comparação entre nós provenientes de grafos diferentes. A centralidade de intermediação é uma medida do potencial de um dado nó de exercer *controle* sobre a comunicação dentro da rede, formalizando a associação intuitiva entre o conceito de centralidade e uma medida de *poder*.

Por fim, Freeman admite uma interpretação de centralidade como a propriedade do nó que está *próximo* a todos os outros, sendo, portanto, independente de intermediários para a sua comunicação. Formalmente:

Definição 8 *A **centralidade de proximidade** de um nó u em um grafo G contendo N nós é dada por:*

$$C_p(u) = (N - 1) \left[\sum_{i=1}^{N} d(u, i) \right]^{-1} \tag{3.1.3}$$

onde $d(u, i)$ é a distância entre o nó u e o nó i.

Assim como as definições prévias, a centralidade de proximidade é normalizada pelo máximo valor possível para um grafo de tamanho N. A centralidade de proximidade identifica

nós posicionados de tal maneira que possam atingir a todos os outros com o menor número de passos, e uma mensagem que origina-se em um nó de máxima centralidade de proximidade percorrerá a rede como um todo em um tempo (ou a um custo) mínimo (44).

Assim, tanto centralidade de intermediação quanto centralidade de proximidade são medidas que transcendem a *localidade* da medida de centralidade de grau, transmitindo aspectos diferentes do conceito intuitivo de centralidade. Veremos posteriormente exemplos da aplicação de ambas na investigação de propriedades de proteínas modeladas por redes.

Paralelamente, há situações em que as propriedades de cada nó individual não são o foco, mas sim as propriedades da rede como um todo. Como exemplo, citamos o estudo da eficiência da rede na disseminação de informação ou no transporte de bens, o estudo da topologia de cadeias alimentares ou a busca por mecanismos que, atuando localmente na formação de ligações, expliquem o perfil estrutural da rede. Para tal fim, introduzem-se medidas que levam em conta todos os nós e/ou arestas simultaneamente. Apresentamos, à guisa de exemplo, duas medidas usadas na caracterização de redes cuja topologia é particularmente interessante.

Definição 9 *Considere um vértice u pertencente a um grafo G e a um subconjunto V^* dos vértices de G tal que:*

$V^* = \{u \in V(G) \mid \delta(u) > 1\}$

e cujo conjunto de vizinhos é $N(u)$. Considere o grafo H tal que

$V(H) = N(u)$,

$E(H) = \{e \in E(G) \mid e = \langle i, j\rangle : i, j \in N(u)\}$.

Definimos o ***coeficiente de aglomeração de u****, $c(u)$, como:*

$$c(u) = \frac{2|E(H)|}{\delta(u)(\delta(u) - 1)} \tag{3.1.4}$$

Definimos também o ***coeficiente de aglomeração de G****, $C(G)$, como:*

$$C(G) = \frac{1}{|V^*|} \sum_{v \in V^*} c(v) \tag{3.1.5}$$

Apesar de sua definição complexa, o coeficiente de aglomeração (ou *clustering coefficient*, na literatura em inglês) é facilmente interpretável. Para um dado nó, o mesmo é a razão entre o número de ligações que existe *entre os vizinhos deste nó* e o máximo número de ligações que poderia existir entre eles. Calculado para um grafo, é uma medida da probabilidade de que os

vizinhos de qualquer nó sejam adjacentes entre si. É, portanto, uma medida da extensão com que os nós se agrupam em "comunidades" dentro da rede.

Definição 10 *Seja G um grafo contendo N vértices. Definimos o* comprimento de menor caminho médio *ou* ***distância geodésica média*** *L de G como:*

$$L(G) = \frac{1}{N(N-1)} \sum_{u,v \in G, u \neq v} d(u,v) \tag{3.1.6}$$

onde $d(u,v)$ é a distância entre o nó u e o nó v.

A distância geodésica média tem interpretação imediata, e fornece uma medida mais poderosa que o diâmetro para a caracterização da distribuição de distâncias em um grafo.

O coeficiente de aglomeração e a distância média são exemplos de medidas que, juntas, diferenciam quantitativamente grafos superficialmente semelhantes. Consideremos, por exemplo, a figura 3.4. O grafo (I) é uma *malha* (ou *grade*) caracterizada pela regularidade do perfil de conexões. Nós vizinhos têm, em geral, vizinhos em comum, e não há ligações "de longo alcance" que encurtem caminhos entre regiões diferentes da rede. Como resultado destas características, observa-se, para este grafo, um coeficiente de aglomeração C de valor relativamente alto, e uma distância geodésica média L idem. O grafo (II), por outro lado, é um grafo *aleatório* ou *binomial*, no qual não se observa nenhum tipo de estrutura local que se possa discernir. Os conjuntos de vizinhos de nós ligados não apresentam, por construção, correlação entre si, visto que a presença ou ausência de cada aresta possível é aleatória e independente de todas as outras. O resultado é um grafo com C e L pequenos (40). Assim, é tentador inferir que L e C juntos denotam uma medida do grau de *regularidade*[§] de um grafo, quando altos, ou de *aleatoriedade* de um grafo, quando baixos.

Redes que representam dados reais, contudo, frequentemente exibem topologias distintas e que não se encaixam adequadamente no espectro regular-aleatória. Denominamos redes com este comportamento de *redes complexas*, em alusão à sua estrutura não-trivial.

Consideremos, por exemplo, uma rede com um grau significativo de regularidade (L e C altos), submetida, repetidas vezes, ao seguinte processo: escolha uma aresta qualquer e substitua uma de suas pontas por um nó escolhido ao acaso. Após algumas iterações, a rede conserva o mesmo número de vértices e arestas, e a localidade das suas ligações não foi significativamente perturbada, resultando em um coeficiente de aglomeração C pouco menor porém ainda alto. Contudo, foram introduzidas ligações entre nós previamente distantes,

[§]"Regularidade" aqui é empregado em um sentido distinto, porém relacionado, da definição de *grafo regular*, aquele cujos vértices todos possuem o mesmo grau.

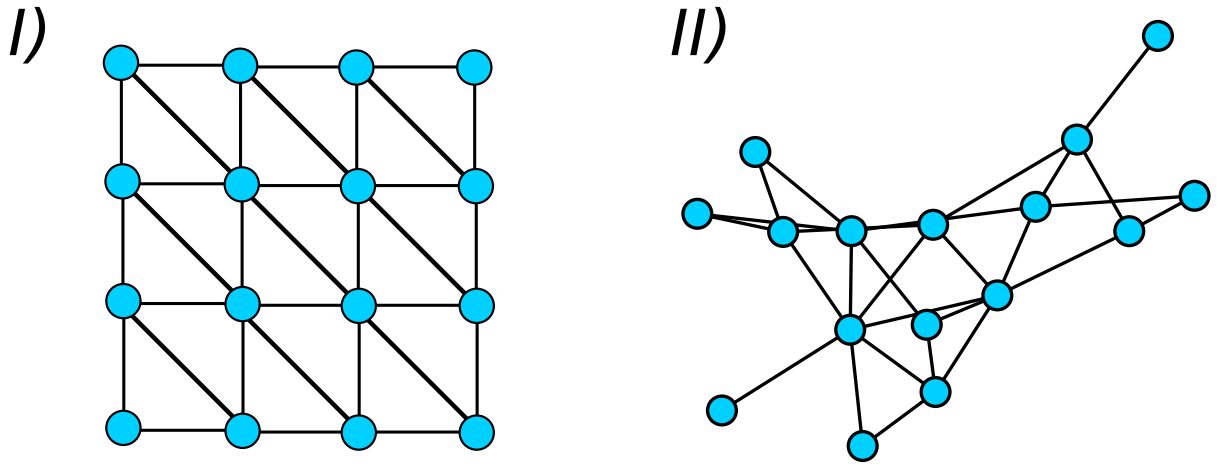

Figura 3.4 – *Exemplos de grafos com estruturas distintas. (I) Malha construída a partir de estrutura local repetitiva. (II) Grafo aleatório.*

gerando caminhos alternativos que diminuem o comprimento das geodésicas da rede como um todo. O resultado é uma distância geodésica média L significativamente mais baixa, isto é, uma rede com C alto e L baixo que não se encaixa na classificação previamente apresentada. Uma consequência dessa topologia é o fenômeno de "mundo pequeno" observado, por exemplo, em redes sociais de amizade ou colaboração, nas quais um número finito e (surpreendentemente) pequeno de passos une quaisquer dois membros da rede, a despeito de sua estrutura localmente aglomerada (40). Este número é tipicamente reportado como menor ou igual a seis, dando origem ao termo "seis graus de separação"¶. As investigações iniciais deste fenômeno são em geral atribuídas a Milgram, num experimento frequentemente citado na literatura introdutória de teoria de redes (vide (46)), ainda que o mesmo não tenha empregado o formalismo de grafos.

Outros conjuntos de dados são melhor representados por redes *livres de escala* (figura 3.5), cuja topologia é caracterizada pela presença de número significativo de vértices com grau muito acima da média, o que diminui ainda mais os comprimentos das geodésicas, gerando "mundos super pequenos". Redes livres de escala, por outro lado, em geral apresentam coeficiente de aglomeração baixo demais para modelar adequadamente sistemas em que se observam comunidades (42).

A presença de vértices altamente conectados, ou *hubs*, tem consequências ambivalentes para a vulnerabilidade de redes livres de escala: as mesmas são notavelmente robustas em relação à remoção aleatória de vértices, porém, sua funcionalidade pode ser rapidamente

¶Uma variação bem-humorada desta idéia é o conceito de *número de Erdős*, que representa a distância, em número de pesquisadores ligados pela co-autoria de artigos científicos, ao influente (e prolífico) matemático húngaro Paul Erdős (1913-1996). Mais de 90% dos pesquisadores com número de Erdős finito está a menos de 8 passos do mesmo na rede de colaborações, fornecendo uma medida aproximada do diâmetro desta. O autor deste trabalho, por ocasião da publicação destes resultados, obterá um número de Erdős 5.

degradada pela remoção sistemática dos vértices de maior grau (40, 42, 47). Redes de mundo pequeno não exibem esta mesma vulnerabilidade, dado que a maior parte de seus vértices tem grau próximo da média. O consenso da literatura parece mostrar que redes aleatórias são as mais resilientes em relação a ataques (sistemáticos ou não), mas ao mesmo tempo, a total ausência de estruturação pode torná-las inaptas para o desempenho de finalidades reais (48). Não há clara diferenciação entre redes livres de escala e redes de mundo pequeno em termos de robustez, em parte devido à escassez de trabalhos investigando redes de mundo pequeno quanto a este aspecto.

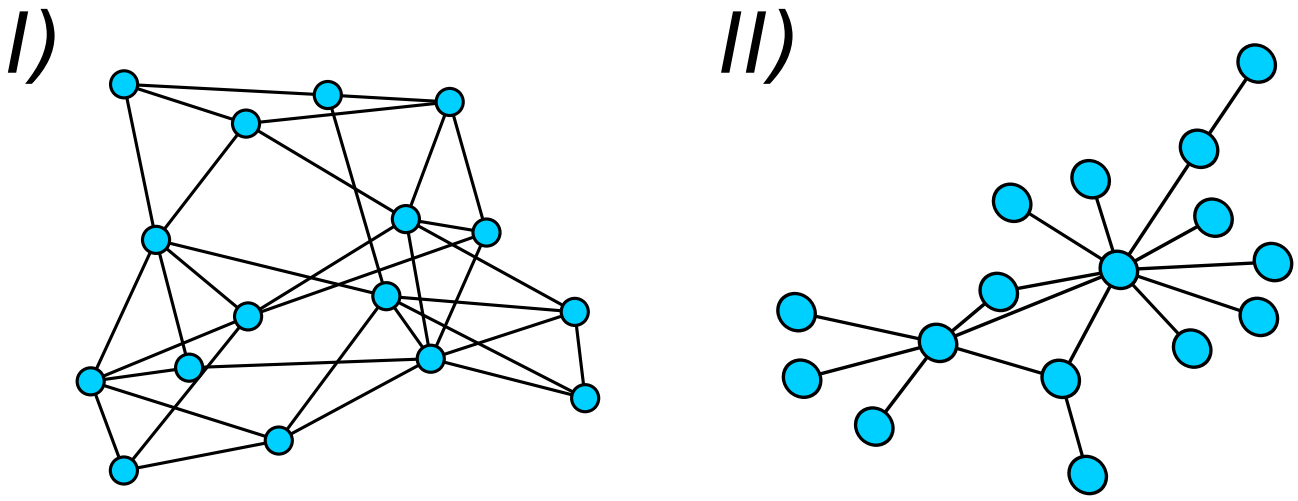

Figura 3.5 – *Exemplo de redes complexas. (I) Rede de mundo pequeno. (II) Rede livre de escala.*

Sabemos que um número importante de redes que representam dados reais exibe topologia de mundo pequeno (40, 47), incluindo, com maior relevância para o presente trabalho, redes que representam estruturas de proteínas (49–52). É opinião do autor que esta propriedade confere a proteínas redundância e robustez estrutural e eficiência na transferência energética, tanto na comunicação de sinais quanto na dissipação de flutuações térmicas, tendo sido portanto favorecida no processo evolutivo.

Naturalmente, existe uma enorme gama de descritores matemáticos utilizados na caracterização de grafos, das quais as medidas de centralidade, aglomeração e distância aqui apresentadas são apenas alguns exemplos. Para uma lista exaustiva, o leitor é direcionado a (53). Dedicaremos, à seguir, uma seção ao problema da modelagem de proteínas por redes, de particular interesse para o presente trabalho.

3.2 Proteínas como Redes Complexas

O autor espera que o leitor esteja convencido de que redes complexas são a ferramenta adequada para modelar uma ampla gama de sistemas. Doravante, apresentamos argumentos delineando como proteínas podem ser proveitosamente modeladas por meio da mesma metáfora, segundo a linha de raciocínio apresentada em (52).

O cientista que deseja entender algum aspecto da físico-química de proteínas deve, inicialmente, estabelecer uma escala de observação. De um ponto de vista nutricional, em um exemplo simplista, uma proteína é perfeitamente modelada por uma esfera, visto que seu valor calórico é, grosseiramente, função exclusiva de sua massa. Outras perguntas requerem, naturalmente, outras descrições; uma investigação da formação de estruturas fibrosas ou interações entre pares de proteínas pode requerer, no mínimo, um modelo que explicite o formato da superfície acessível ao solvente, ou talvez uma descrição das estruturas secundárias que interagem para dar à proteína seu formato. Uma investigação de um mecanismo de catálise pode não ser possível a não ser que o modelo leve em conta todos os átomos interagentes e suas posições variantes no tempo, e até mesmo as funções de onda de seus elétrons. Em todos os casos, a "escala" do modelo corresponde ao tamanho do menor elemento cujo movimento e/ou interação é relevante para o problema em questão. Muitas vezes, o comportamento do sistema como um todo é imprevisível a partir das interações simples de seus elementos; a esta característica associamos a *complexidade* de um sistema - o grau em que o mesmo é mais do que a soma de suas partes.

Pode-se argumentar, contudo, que existe uma escala "natural" para a modelagem de proteínas: a escala dos aminoácidos individuais. De fato, a troca da identidade do aminoácido em cada posição da sequência é essencialmente a única estratégia[‖] da qual a célula dispõe para provocar mudanças na estrutura ou atividade de uma proteína. Se não está ao alcance da natureza modificar um único átomo em particular de uma estrutura a fim de melhorar sua eficiência, e ainda assim os organismos apresentam vasta gama de tipos de estruturas, dobramentos e mecanismos catalíticos, cabe aos cientistas examinarem a necessidade de incluir cada átomo individual em seus modelos.

Krishnan *et al.* (52) dão continuidade a este raciocínio sugerindo que a seqüência de aminoácidos associada a cada proteína define o espaço sobre o qual cada análise deve ser projetada, adotando um sistema de coordenadas *relativo* ou *intrínseco*, por assim dizer. Para isto, ao invés de registrar a posição de cada aminoácido segundo um conjunto de eixos carte-

[‖] Naturalmente, a escolha do termo *estratégia* não implica em intenção do autor de atribuir "propósito" a uma célula, ou sugerir que a seleção natural seja de algum modo direcionada.

sianos externos, resultando numa descrição da estrutura que depende de $3N$ variáveis para N aminoácidos (ou possivelmente $4N$ caso a descrição seja dependente do tempo), registra-se a informação relativa (e biologicamente relevante) da *distância* entre cada par de aminoácidos, preenchendo uma matriz $N \times N$. Toda propriedade pode ser, em princípio, projetada em um sistema de coordenadas relativo por meio de processo análogo.

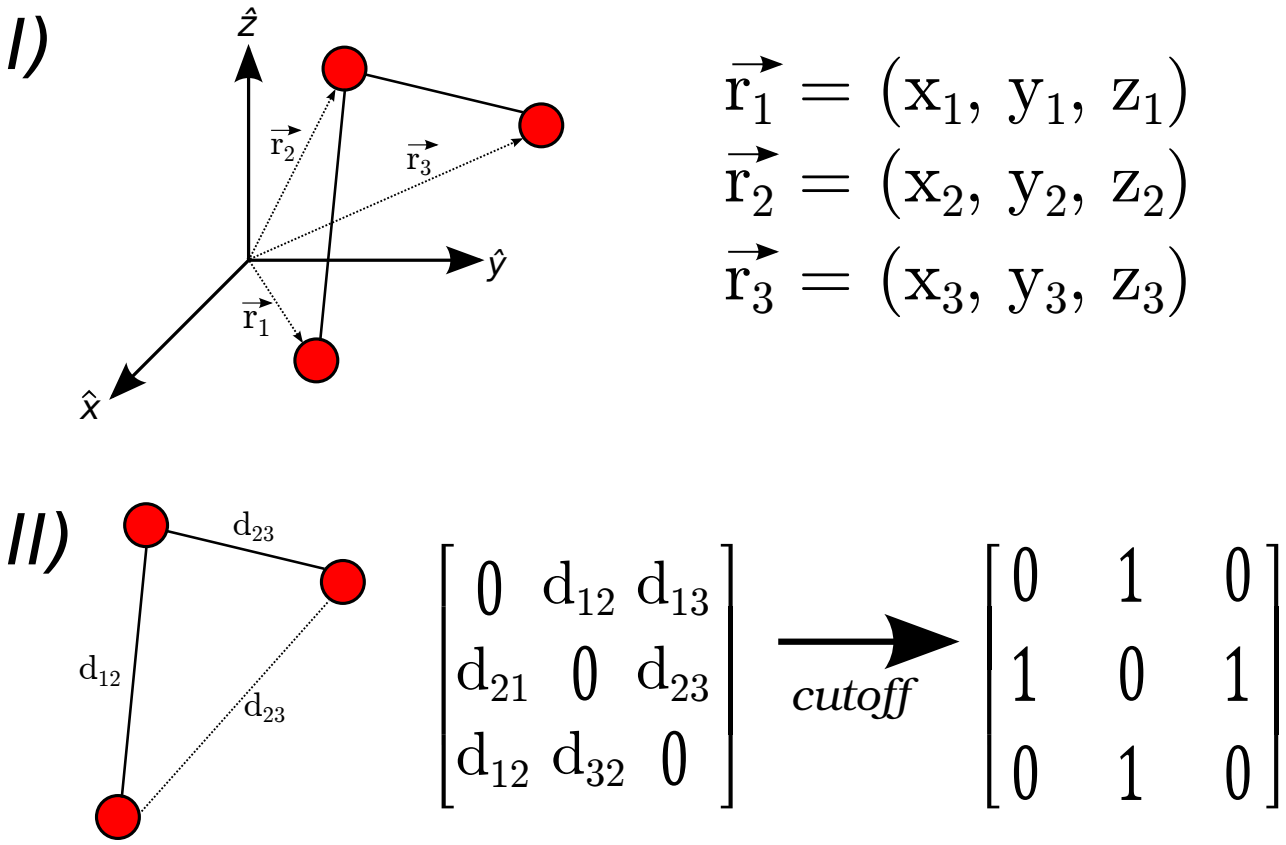

Figura 3.6 – *Ilustração das descrições absoluta e relativa, para um sistema de três entidades. (I) Descrição absoluta, em que as posições são descritas com relação a três eixos ortogonais. (II) Descrição relativa, em que as distâncias entre cada par de entidades são enumeradas, em vez de suas posições. A aplicação de um* cutoff *transforma a matriz de distâncias em uma matriz de adjacência.*

Como passo final, o espaço relativo pode ser particionado (em geral pela aplicação de um valor limite de decisão ou *cutoff*) de tal forma que a propriedade medida seja mapeada em um conjunto binário ou *sim-ou-não*. Sua matriz relativa se torna então uma matriz binária simétrica, tal que cada posição $\mathbf{A}_{ij}$ tem valor 1 quando o valor daquela propriedade, medido em relação ao par i, j de aminoácidos, atinge o limiar estabelecido, ou valor *zero*, caso contrário.

O leitor notará que esta definição ecoa a definição de matriz de adjacência - a propriedade foi mapeada em um grafo. Se, por exemplo, aplicarmos esse processo ao conjunto inicial de posições de cada aminoácido, mapeando as posições em distâncias entre pares de aminoácidos e posteriormente aplicando um *cutoff* fisicamente sensato, a matriz obtida nada mais é que o

mapa de contatos da proteína, já amplamente empregado na literatura. Krishnan *et al.* relatam que uma matriz análoga ao mapa de contatos pode ser construída para qualquer propriedade dos aminoácidos, e que essa construção é matematicamente equivalente à construção de uma rede.

Uma crítica que pode ser dirigida às descrições relativas é que sua interpretação pode se tornar confusa quando a propriedade medida não deriva de coordenadas espaciais. Por exemplo, pode-se postular que a transmissão de sinais alostéricos se dá através de uma geodésica no mapa de contatos, visto que cada passo na rede corresponde à transmissão do sinal entre aminoácidos fisicamente próximos. Contudo, para uma rede que represente, por exemplo, similaridade de hidrofobicidade, ou de outra medida físico-química, o significado de um caminho sobre a rede se torna discutível. Talvez por esta razão, a maioria dos trabalhos que empregam modelos de redes complexas para investigar proteínas se restringe à construção de mapas de contatos ou de redes de correlação de movimentos, e o presente trabalho não é exceção. Não obstante, a notação matricial traz à mão o conjunto de ferramentas de análise de grafos, das quais trata a seção prévia. Sua aplicação permite, por exemplo, a comparação entre propriedades de proteínas ou domínios de diferentes tamanhos, classes e/ou dobramentos por meio de descritores matemáticos apropriados derivados das respectivas matrizes, dos quais as medidas de centralidade apresentadas são um exemplo.

Como corolário da argumentação apresentada, observa-se na literatura um conjunto expressivo de trabalhos que empregam e demonstram essa metodologia. Podemos destacar, entre os mais importantes:

- Dokholyan *et al.* em (54) constroem grafos baseados em duas proteínas que se dobram segundo um modelo cinético de dois estados, para as quais possuem estruturas relativas ao *ensemble* do estado de transição. Os grafos são construídos considerando os aminoácidos como nós ligados por arestas quando as distâncias entre seus respectivos C_α é menor que 8.5Å. Os autores afirmam que os grafos obtidos exibem topologia de mundo pequeno, porém não livre de escala, e que a medida da distância geodésica média para cada estrutura é capaz de detectar de que lado da barreira cinética de transição a estrutura se encontra, ao contrário de medidas energéticas ou de similaridade estrutural.

- Vendruscolo *et al.* em (50) constroem grafos baseados nos *ensembles* de transição de seis proteínas, utilizando os mesmos critérios de Dokholyan *et al.* mas representando a variabilidade estrutural nos pesos das arestas. Os autores demonstram que os grafos são topologicamente de mundo pequeno, e sugerem que isso é consequência de sua geometria (i.e. da razão área/volume). Demonstram também que, no estado de transição, os

resíduos com maior centralidade de intermediação correspondem aos resíduos chave para a formação de núcleos de dobramento, previamente identificados experimentalmente em (55).

- Greene *et al.* em (49) investigam um conjunto de 65 proteínas e constroem grafos baseados nas distâncias entre todos os pares de átomos, tal que um par de átomos a uma distância menor que 5Å implica em uma aresta ligando o par de aminoácidos aos quais eles pertencem. Com isto, demonstram que os grafos gerados possuem topologia de mundo pequeno. Os autores demonstram também que, quando as ligações entre aminoácidos sequencialmente próximos são desconsideradas, o grafo então gerado tem topologia livre de escala.

- del Sol *et al.* em (51) representam 48 complexos proteína-proteína como grafos de mundo pequeno, os quais são construídos segundo as distâncias entre todos os pares de átomos, com um *cutoff* de 5Å. Demonstram que os resíduos que participam da interação proteína-proteína podem ser identificados por meio de sua centralidade de intermediação.

- Amitai *et al.* em (56) constroem grafos para 178 proteínas identificando as interações entre pares de átomos por critérios energéticos e geométricos. Demonstram que a centralidade de proximidade pode ser utilizada para identificar resíduos que pertencem ao sítio catalítico ou outros sítios funcionais.

- Um expressivo conjunto de autores emprega representações de rede para identificar padrões que permitem classificar estruturas de proteínas em termos de seus dobramentos e suas famílias, dentre os quais citamos (57–60).

O leitor interessado em revisões exaustivas do tema é direcionado a (52, 61, 62). De modo geral, observamos ausência de unanimidade em relação ao protocolo utilizado para modelar estruturas como grafos. O primeiro objetivo deste trabalho é fornecer uma justificativa, preferencialmente a partir da análise de dados experimentais de estruturas, que possa informar e embasar o protocolo aqui adotado, e será tratado no próximo capítulo.

CAPÍTULO 4

Resultados

4.1 Construção de Redes que Representam Estruturas de Proteínas

A revisão do conjunto de trabalhos que empregam a descrição de redes complexas para estruturas de proteínas revela algum grau de variação em relação ao protocolo empregado para a construção da rede que corresponde a cada estrutura. A maior parte dos trabalhos examinados não oferece nenhuma justificativa, teórica ou experimental, para o protocolo utilizado, e o parâmetro numérico de distância de *cutoff* é tipicamente apresentado sem o acompanhamento de dados de qualquer tipo que iluminem sua escolha. Procuramos, nesta seção, estabelecer um protocolo embasado por observações experimentais.

Para tal fim, utilizamos um conjunto de estruturas oriundas da base de dados *SCOP* (63). A base SCOP consiste no conjunto de proteínas e domínios descritos na literatura, a maior parte das quais proveniente do repositório *Protein Data Bank* (PDB) (64), manualmente classificadas segundo a similaridade de suas estruturas e sequências de aminoácidos em classes, superfamílias e famílias. Parcialmente derivada da base SCOP é o conjunto ASTRAL (65), que disponibiliza a edição *ASTRAL SCOP 40%* (aqui referida por SCOP40). A base SCOP40 é construída pela adição consecutiva de estruturas a partir da base SCOP: cada exemplar do SCOP é adicionado à SCOP40 desde que não apresente mais de 40% de identidade com nenhuma estrutura já presente na SCOP40. O resultado é um conjunto de estruturas que abrange todas as proteínas de estrutura conhecida sem, contudo, apresentar estruturas com dobramentos redundantes. Para os efeitos deste trabalho, consideramos a base SCOP40 como o menor conjunto de estruturas que abrange todas as famílias e dobramentos conhecidos. A edição da SCOP40 aqui utilizada, versão 1.75, foi publicada em Junho de 2009 e possui 10.569 estruturas. Todas as investigações foram repetidas para a totalidade do conjunto de estruturas, e os resultados acumulados de forma a produzir a representação mais geral possível das propriedades de interesse para uma proteína genérica.

De posse de um conjunto representativo de estruturas, realizamos experimentos para caracterizar a distribuição espacial de seus átomos e resíduos, por meio da *função de distribuição radial* ou *RDF*, tradicionalmente denotada $g(r)$. A RDF é uma grandeza adimensional definida para fluidos, gases ou conjuntos de partículas interagentes em geral, que revela, quando existem, as correlações entre as posições das mesmas, as quais são atribuídas aos efeitos das interações. A RDF é uma grandeza que pode ser obtida experimentalmente com relativa facilidade, mediante experimentos de difração de raios-X ou difração de nêutrons. Aqui, empregaremos a definição de RDF apresentada em (32)*.

A RDF é definida baseada na estrutura de um fluido de referência: um gás ideal de N partículas contido em um volume V. Para tal sistema, define-se a densidade numérica de partículas por volume, $n_{\text{gás}}$, como a razão entre o número total de partículas e o volume total por elas ocupado. Em particular, se as partículas ocupam um volume esférico de raio R, tem-se:

$$n_{\text{gás}} = \frac{N}{V} = \frac{N}{\frac{4\pi}{3}R^3} \tag{4.1.1}$$

Uma partícula qualquer dentro desse gás ideal conta um certo número de vizinhos $m_{\text{gás}}$ (figura 4.1), que depende do volume que ela "enxerga", ou seja, da máxima distância entre duas partículas que ainda permite que elas sejam consideradas vizinhas (distância de *cutoff*). Denotando tal distância por r, uma partícula em uma posição qualquer longe da superfície da esfera tem um número de vizinhos $m_{\text{gás}}$ que vale em média:

$$m_{\text{gás}}(r) = \int n_{\text{gás}} dV = n_{\text{gás}} \frac{4\pi}{3} r^3 \tag{4.1.2}$$

E o número de vizinhos situados entre uma distância r e $r + dr$ da mesma será, então:

$$dm_{\text{gás}}(r) = n_{\text{gás}} 4\pi r^2 dr \tag{4.1.3}$$

Com base nos parâmetros do gás ideal, podemos estudar uma distribuição qualquer em que as posições das partículas apresentam correlação entre si. Postulamos que as interações partícula-partícula geram posições preferenciais para cada partícula em relação às outras, re-

*O leitor interessado em uma demonstração formal é direcionado a algum livro-texto de Mecânica Estatística que contenha uma seção sobre a estrutura molecular do estado líquido, tal como (66).

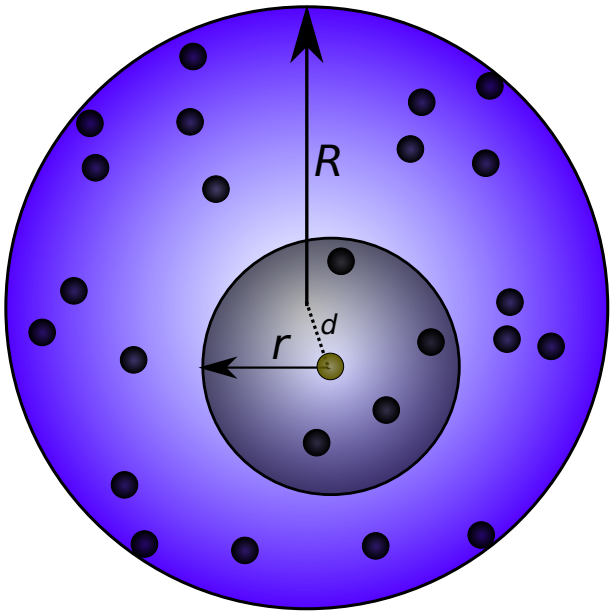

Figura 4.1 – *Distribuição de N partículas ocupando um volume esférico de raio R. A partícula destacada, localizada a uma distância d do centro da esfera, enxerga m vizinhos, contidos na esfera definida pela distância de* cutoff *r.*

sultando num efeito observável de modular a densidade de vizinhos a uma dada distância de qualquer partícula. Nos referimos a essa modulação como uma *estruturação*, e submetemos a distribuição de interesse ao mesmo tratamento que submetemos o gás ideal, só que incluindo todos os efeitos da estruturação numa função que depende da distância partícula-partícula. Essa função é a RDF, $g(r)$.

Consideremos, então, uma distribuição estruturada de N partículas ocupando um volume esférico de raio R. Para uma partícula qualquer longe da superfície, o número de vizinhos m_{est} que ela enxerga entre r e $r + dr$ é, por analogia à equação (4.1.3):

$$dm_{\text{est}}(r) = n_{\text{est}} 4\pi r^2 = n_{\text{gás}} g(r) 4\pi r^2 dr \tag{4.1.4}$$

Na qual o subscrito *"est"* se refere a uma distribuição estruturada, $n_{\text{gás}}$ corresponde à densidade numérica média de partículas para um gás ideal, e $g(r)$ modula a densidade teórica de forma que $n_{\text{est}}(r) = n_{\text{gás}} g(r)$.

Integrando entre 0 e r, obtemos o número m de vizinhos que cada partícula enxerga na distribuição estruturada:

$$m_{\text{est}}(r) = \int dm_{\text{est}}(r) = \int_0^r n_{\text{gas}} g(r) 4\pi r^2 dr$$

$$m_{\text{est}}(r) = n_{\text{gás}} 4\pi \int_0^r g(r) r^2 dr \tag{4.1.5}$$

Demonstramos, no apêndice A, seção A.1, que as equações (4.1.2) e (4.1.5) se mantém válidas em média mesmo quando se incluem todas as partículas, e não apenas aquelas distantes da superfície, desde que se obedeça $r \ll R$. Contudo, a análise do caso limite em que r é maior que o diâmetro da distribuição é informativa. Neste caso, ambas as equações devem fornecer o mesmo resultado: o número total de partículas N. Devemos ter:

$$\lim_{r \to +\infty} \frac{\int_0^r n_{\text{gás}} g(r')\, 4\pi r'^2\, dr'}{\int_0^r n_{\text{gás}}\, 4\pi r'^2\, dr'} = 1 \tag{4.1.6}$$

Que implica (vide demonstração no apêndice A, seção A.2):

$$\lim_{r \to +\infty} g(r) = 1 \tag{4.1.7}$$

A propriedade (4.1.7) pode ser usada para validar curvas $g(r)$ obtidas teoricamente ou experimentalmente; na prática, $g(r)$ se estabiliza em 1 para valores relativamente pequenos de r, que se relacionam com o alcance efetivo das interações partícula-partícula.

Em simulações, a RDF pode ser calculada trivialmente por meio de contagem normalizada de vizinhos à uma distância entre r e $r + dr$ iterada sobre cada partícula (vide (32)). Aqui, empregamos um algoritmo ligeiramente diferente, partindo do número *total* de vizinhos contado por todas as partículas para distâncias menores que r crescentes. Denotando este número, o número total de contatos contados, por $M(r)$, temos:

$$M(r) = N m(r)$$

$$\frac{M_{\text{est}}(r)}{M_{\text{gás}}(r)} = \frac{N m_{\text{est}}(r)}{N m_{\text{gás}}(r)} = \frac{n_{\text{gás}} 4\pi \int_0^r g(r') r'^2\, dr'}{n_{\text{gás}} 4\pi \frac{r^3}{3}} = \frac{3}{r^3} \int_0^r g(r') r'^2 dr'$$

Isolando $g(r)$, vem:

$$\frac{1}{3}\frac{d}{dr}\left[r^3\frac{M_{\text{est}}(r)}{M_{\text{gás}}(r)}\right] = \frac{d}{dr}\left[\int_0^r g(r')r'^2dr'\right]$$

$$g(r) = \frac{1}{3r^2}\frac{d}{dr}\left[r^3\frac{M_{\text{est}}(r)}{M_{\text{gás}}(r)}\right] \tag{4.1.8}$$

A equação (4.1.8) proporciona uma maneira de obter $g(r)$ para uma distribuição estruturada qualquer, desde que sejamos capazes de fornecer uma medida do número de contatos contados em uma distribuição aleatória de tamanho e densidade equivalentes.

Munidos deste embasamento teórico, procuramos calcular a RDF para um átomo qualquer em uma estrutura genérica de proteína. O procedimento adotado é descrito à seguir:

1. Seja uma estrutura de proteína. Calcula-se, para a mesma, a distância entre todos os pares de átomos, preenchendo-se uma matriz simétrica de distâncias. A esta matriz, aplica-se comparação com um valor dado de *cutoff*, r, obtendo-se uma matriz de contatos. Da quantidade de elementos positivos na matriz, subtrai-se o número N de átomos da estrutura, que corresponde aos contatos triviais de cada átomo consigo mesmo, e toma-se a metade do resultado, pois cada contato entre um par de átomos é contado duas vezes. O valor resultante é o número absoluto de contatos contados à distância menor ou igual a r para esta estrutura, $M_{\text{est}}(r)$. Esse procedimento é repetido para r variando de 1Å a 12Å, em passos de $0,1\text{Å}$.

2. Geram-se 5 distribuições aleatórias de N pontos contidos aproximadamente dentro da superfície que contém a estrutura em questão. Aplica-se a cada uma delas o processo descrito no item 1, e toma-se a média entre os 5 números de contatos absolutos obtidos. O resultado é o número médio de contatos à distância menor ou igual a r para a distribuição aleatória, $M_{\text{gás}}(r)$. Calcula-se então $g(r)$ para essa estrutura por meio da equação (4.1.8).

3. Os itens 1 e 2 são repetidos para cada uma das 10.569 estruturas da base SCOP40. Finalmente, toma-se a média entre todas as curvas $g(r)$ geradas, e reporta-se a curva resultante em forma de gráfico.

É importante ressaltar que, para os resultados obtidos nesta seção, átomos de Hidrogênio são ignorados em todas as estruturas. Nenhum prejuízo teórico é incorrido, mas este fato deve ser tido em mente durante a interpretação das curvas.

Para minimizar o erro devido ao formato não necessariamente esférico de cada proteína, procuramos produzir distribuições aleatórias de pontos cujo formato imite o formato da estrutura em questão. Obedecemos ao seguinte procedimento: geram-se, consecutivamente, pontos com coordenadas cartesianas aleatórias, e rejeitam-se aqueles cuja mínima distância em relação a qualquer um dos átomos da estrutura seja maior que 4Å. O processo é repetido até que tenham sido aceitos pontos em número igual ao número de átomos, e o resultado é uma distribuição cujas dimensões são ligeiramente maiores que as da estrutura original. Finalmente, as coordenadas de todos os pontos da distribuição aleatória são reescalonadas, de forma que seu volume final "encolha" para um volume semelhante ao da estrutura original. O fator de reescalonamento é a razão entre o desvio padrão das coordenadas da estrutura original e o desvio padrão das coordenadas aleatórias. Como resultado deste procedimento, quaisquer cavidades existentes no interior da estrutura original são em média preenchidas e significativamente encolhidas na distribuição aleatória, resultando numa densidade aproximadamente uniforme no interior da superfície.

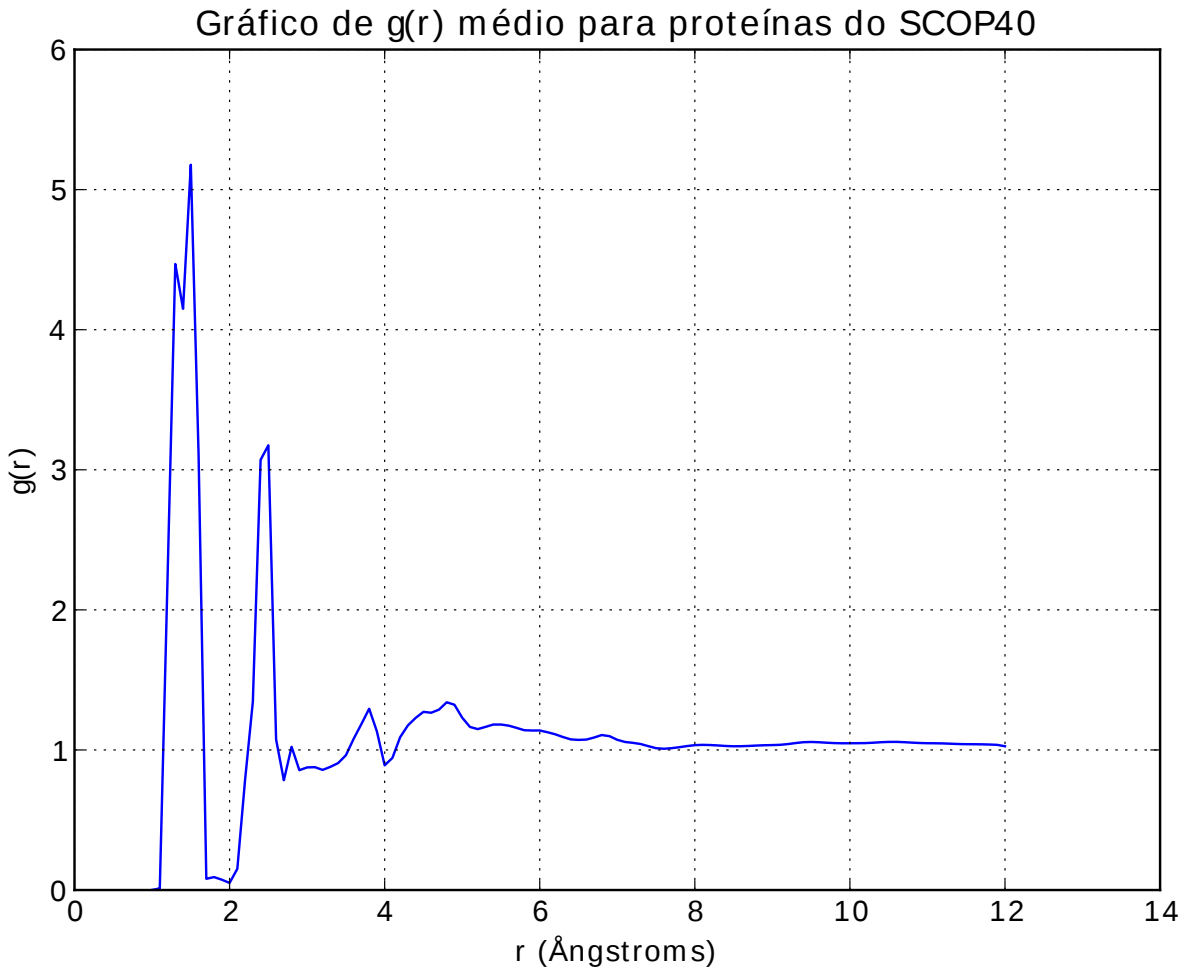

Figura 4.2 – *Gráfico de $g(r)$ médio calculado sobre todas as estruturas da base SCOP40.*

A figura 4.2 apresenta o resultado deste procedimento. Os resultados mostrados, exceto quando indicado, foram obtidos por meio de *software* desenvolvido pelo autor para este fim, escrito em linguagem *Python* (67) associada aos pacotes *NumPy* (68) e *SciPy* (69) para a obtenção dos resultados e *Matplotlib* (70) para a produção dos gráficos.

O grande número de estruturas incluídas na média tem por efeito minimizar e compensar as fontes de erro. Os picos e vales que sobrevivem representam a organização observada consistentemente em torno de todos os átomos. Contudo, a curva de modo geral apresenta regiões de pouca suavidade, o que pode ser seguramente atribuído ao efeito da derivação presente na equação (4.1.8) aplicada ao domínio discreto. Para facilitar a análise, então, aplicamos à curva obtida uma suavização por meio de média móvel com janela de 3 pontos $(n-1,\, n,\, n+1)$. O gráfico suavizado é apresentado na figura 4.3.

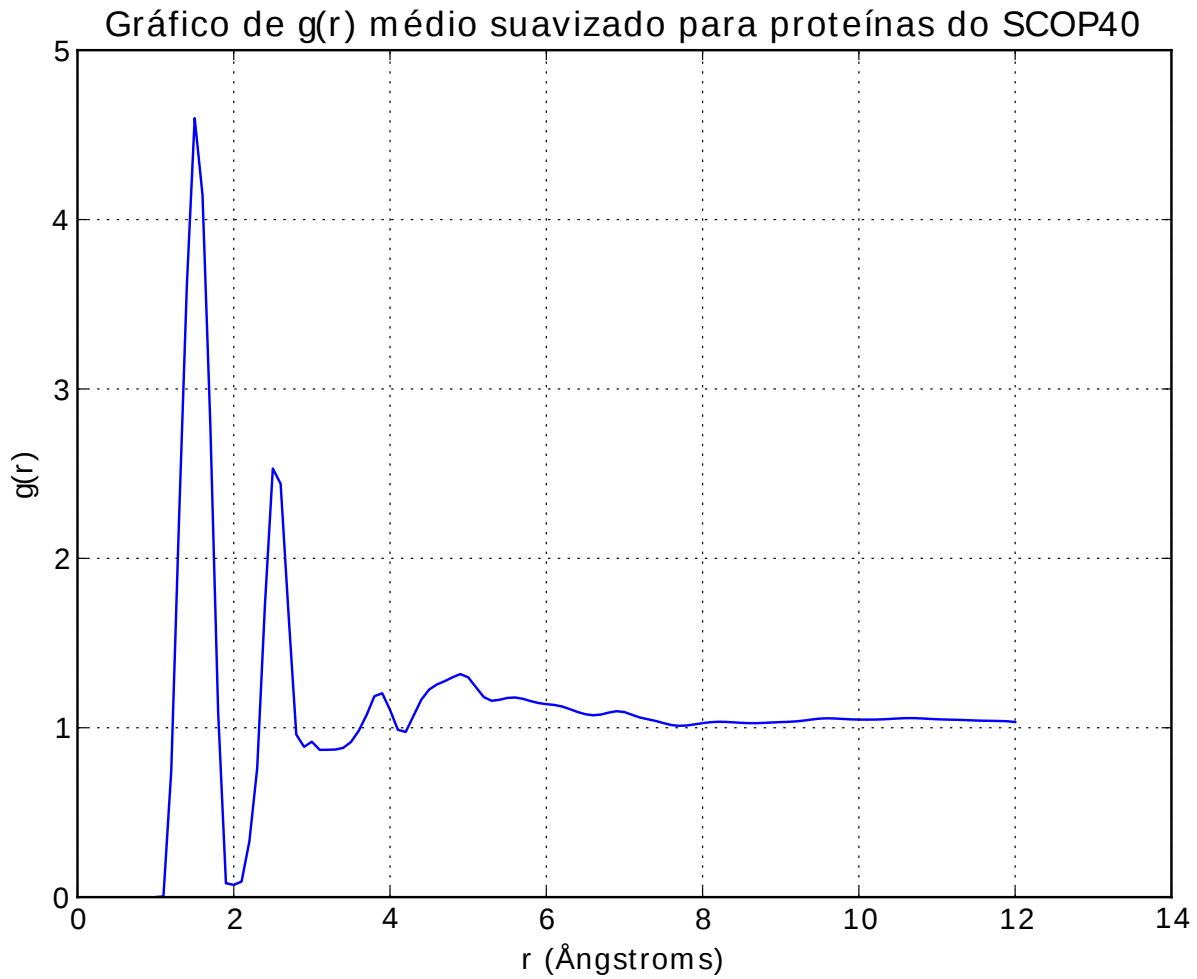

Figura 4.3 – *Gráfico de $g(r)$ médio calculado sobre todas as estruturas da base SCOP40, suavizado pela aplicação de média móvel com janela de 3 pontos.*

A inspeção do gráfico revela uma densidade nula para valores menores do que aproximadamente $r = 1$ Ångstrom, que corresponde aproximadamente ao valor médio do raio atômico, e

próximo a $r = 1,5$ Ångstrom um pico correspondente à distância média da ligação covalente. O próximo pico, na vizinhança de $r = 3$ Ångstroms, corresponde à distância típica de ligações de Hidrogênio. Para distâncias maiores, observa-se uma estruturação que não pode ser imediatamente atribuída a nenhuma causa óbvia, com densidades maiores do que a expectativa aleatória persistindo até distâncias próximas de 7 a 8 Ångstroms. Em aproximadamente $r = 8$ Ångstroms a curva se estabiliza em torno do valor $g(r) = 1$ correspondente à expectativa aleatória.

Naturalmente, a curva obtida apenas confirma as expectativas, e, neste sentido, não traz novidades. A distância média das ligações covalentes ou de Hidrogênio em polímeros biológicos em geral já é muito bem determinada, e a concordância observada deve ser interpretada como *validação* do método empregado *neste* trabalho, e não como uma contribuição proveniente do mesmo. A principal informação deste gráfico é menos conspícua: a partir de $r = 8$ Ångstroms de distância de cada centro, a distribuição enxergada é *indistinguível* de uma distribuição aleatória. Interpretamos aqui esta distância como uma medida da *escala* de observação a partir da qual estruturas de proteína apresentam comportamento de distribuição aleatória. Este é um resultado importante, e corrobora as conclusões reportadas em (49), que observa organização de mundo pequeno mas não livre de escala em um conjunto representativo de estruturas de proteína, e afirma que, quando são desconsideradas ligações "de curto alcance" (entre resíduos próximos na estrutura primária), a topologia resultante é de rede aleatória.

Seguindo esta linha de investigação, definimos e calculamos também uma variação da RDF: a função de distribuição radial para os carbonos C_α, denotada RDF$_\alpha$ ou $g_\alpha(r)$.

A RDF$_\alpha$ é definida como uma extensão direta da RDF, e calculada segundo o mesmo protocolo desta. Contudo, são considerados apenas os átomos de carbono C_α da estrutura de cada proteína, e as distribuições aleatórias são geradas com um número compatível de pontos. A curva resultante descreve a estruturação observada em relação às distâncias entre os resíduos, para os quais os carbonos C_α atuam como marcadores de posição. Apresentamos a seguir na figura 4.4 a curva obtida, também suavizada por média móvel com janela de 3 pontos:

Novamente, a curva acompanha os resultados já bem determinados da literatura. Em torno de 4 Ångstroms observa-se um pico bem pronunciado correspondente à distância média entre os C_α de resíduos consecutivos, isto é, resíduos unidos por ligação peptídica. O próximo pico, na vizinhança de 6 Ångstroms mas já com uma posição menos bem definida, corresponde à distância entre carbonos C_α de resíduos ligados a um vizinho comum, ou seja, resíduos com índices da forma n e $n + 2$ na estrutura primária. Observa-se, em contrariedade ao gráfico

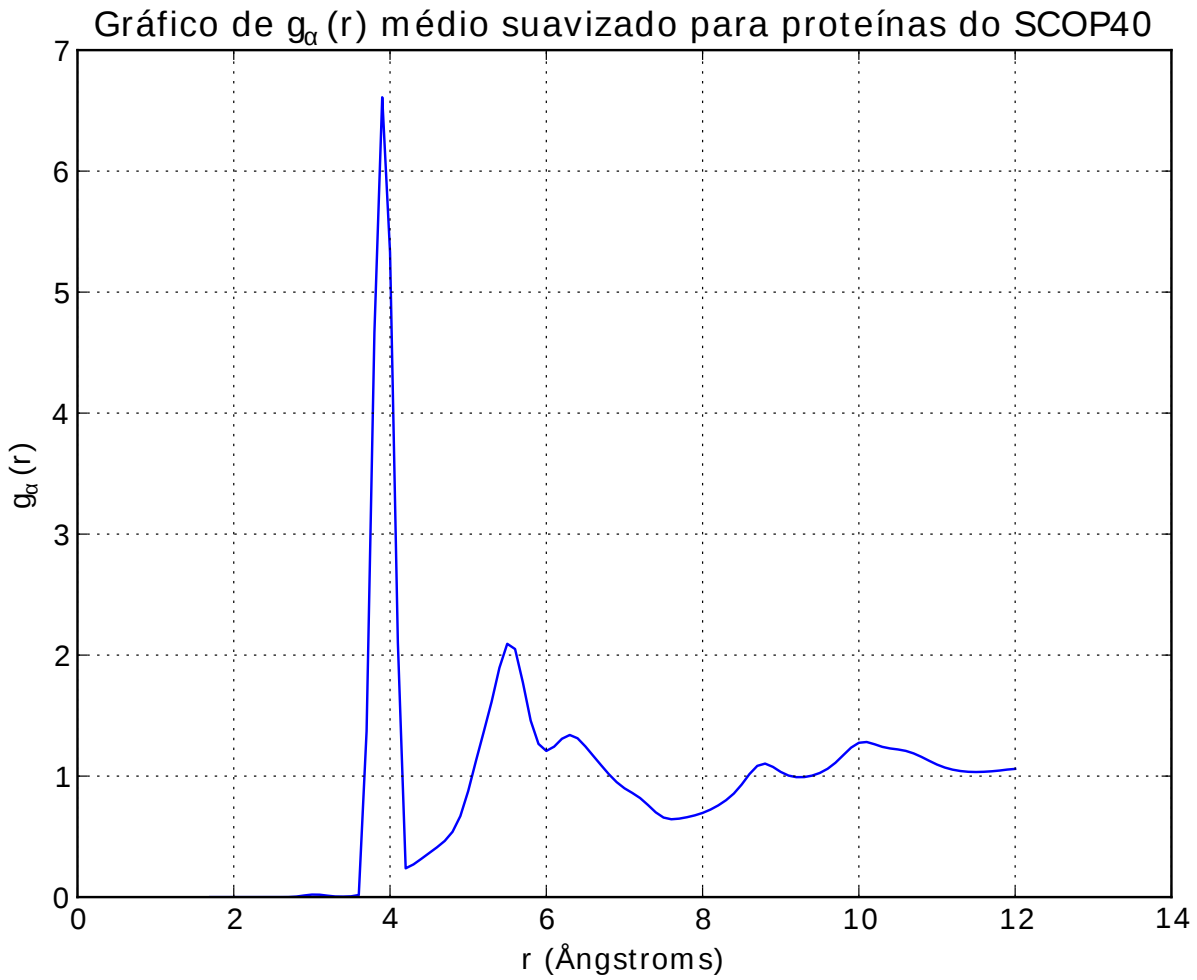

Figura 4.4 – *Gráfico de $g_\alpha(r)$ médio calculado sobre todas as estruturas da base SCOP40, suavizado pela aplicação de média móvel com janela de 3 pontos.*

da figura 4.3, que a estabilização em torno do valor $g_\alpha(r) = 1$ de densidade aleatória só é atingida nas proximidades de $r = 11$ Ångstroms. Esta aparente incoerência é possível na medida em que, no cálculo da RDF$_\alpha$, estudamos a distribuição de um subconjunto particular de pontos, em meio a um conjunto de partículas previamente consideradas indistingüíveis. Ao mover o limiar de estabilização para distâncias maiores, a curva da RDF$_\alpha$ revela um grau de estruturação previamente ignorado pela RDF, mas não altera a conclusão geral de que *existe* uma escala de observação a partir da qual a distribuição enxergada é aleatória.

Os dados até aqui apresentados lançam luz sobre protocolos descritos na literatura para a construção de redes a partir de estruturas de proteínas. Dokholyan *et al.*, em (54), e Vendruscolo *et al.*, em (50), por exemplo, constroem redes considerando cada par de aminoácidos *ligado* quando a distância entre seus C_α é menor ou igual a $8,5$Å. Investigando a figura 4.4, notamos que esta escolha é equivalente a afirmar que, para os propósitos do referido trabalho, a esfera de influência de cada resíduo de aminoácido se estende por uma distância que

é em média capaz de englobar seus vizinhos covalentes próximos *e os vizinhos de seus vizinhos*, além de quaisquer outros que, por meio de interações não-covalentes, ocupem posição suficientemente próxima. Este protocolo é ilustrado na figura 4.5.

Já Greene *et al.*, em (49), e del Sol *et al.*, em (51), consideram as distâncias entre todos os pares de átomos na construção de suas redes. Dois resíduos de aminoácidos são considerados ligados quando a menor distância entre qualquer par de átomos tal que cada átomo pertença a um dos resíduos seja menor ou igual a 5Å. A inspeção da figura 4.3 revela que, ao fazer esta escolha, os autores implicitamente contemplam interações covalentes e ligações de Hidrogênio como mecanismos por detrás da asserção de que um resíduo está *ligado* a outro. Contudo, o *cutoff* de 5Å repousa arbitrariamente sobre uma região em que se observa estruturação, a qual se estende aproximadamente desde 4Å até 8Å. Se o critério empregado pelos autores considera que tal estruturação é o efeito de interações irrelevantes do ponto de vista da análise pretendida, um *cutoff* de 4Å levaria às mesmas conclusões, possivelmente a um custo computacional menor[†]. Por outro lado, na medida em que o formato da RDF pode ser interpretado como consequência do formato do potencial de interação devido à partícula de referência[‡], ignorar a estruturação que persiste até 8Å acarreta perda de informação, com consequências não-óbvias para o resultado da análise. Este protocolo está ilustrado na figura 4.6.

De todo modo, é razoável supor que a inclusão de todos os átomos na análise ofereça vantagens em comparação ao protocolo baseado nos C_α. A inspeção dos dados gerados para proteínas individuais (não mostrados aqui) revela que a construção baseada nos C_α contabiliza um número de vizinhos menor que a construção baseada em todos os pares, para todo *cutoff* fixo, e é razoável postular que a diferença seja em parte devida a importantes contatos cadeia lateral-cadeia lateral que coloquem os respectivos carbonos alfa a distâncias maiores que o *cutoff*. Com base na hipótese que todo contato interatômico tem a capacidade de transferir energia, e em face da disponibilidade de poder computacional, favorecemos neste trabalho o protocolo que leva em consideração todos os pares de átomos, mas optamos por investigar diretamente o efeito do *cutoff*. Para isto, calculamos uma segunda variação da RDF, denominada aqui de função de distribuição radial agrupada por resíduo, ou RDF_R. Partimos da equação (4.1.5), aqui repetida, que relaciona o número de vizinhos contado por uma partícula com a RDF:

[†]O custo computacional da contagem de vizinhos só é independente do *cutoff* quando algoritmos sub-ótimos são empregados.

[‡]Tal interpretação, denominada *potencial de força média* calculado entre pares, ignora os efeitos das interações que não envolvem a partícula de referência, e é tanto melhor quanto menor for a densidade numérica do sistema, donde resulta que sua aplicabilidade para sistemas tais como estruturas de proteínas é limitada a uma primeira aproximação.

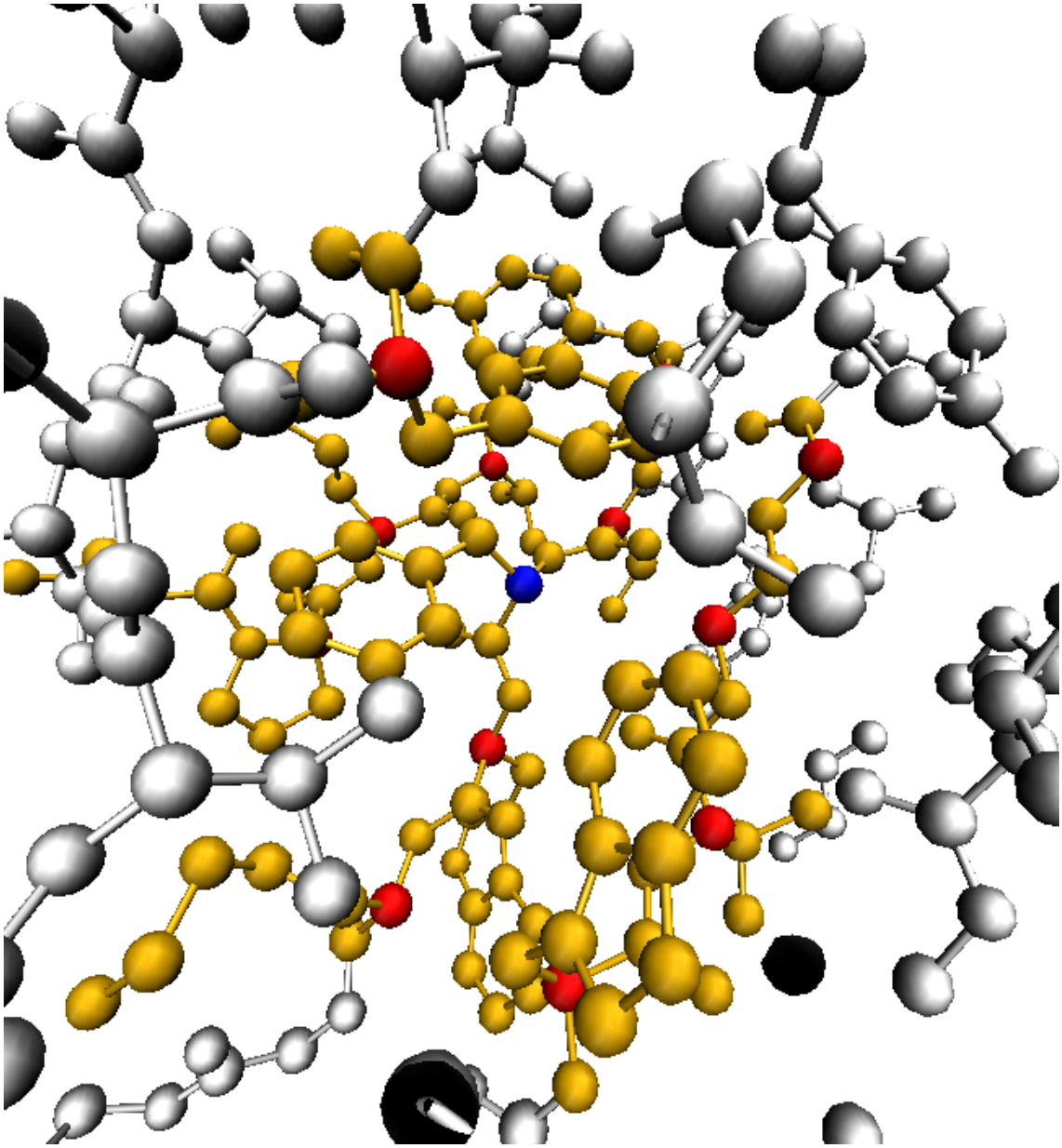

Figura 4.5 – *Ilustração do método de contagem de vizinhos baseado nas distâncias entre os carbonos C_α. O átomo ilustrado em azul é o carbono alfa do resíduo de interesse. Os carbonos alfa que estão a menos de $8\mathring{A}$ de distância do mesmo estão ilustrados em vermelho, e os resíduos aos quais eles pertencem, considerados vizinhos do resíduo de interesse, ilustrados em laranja. Figura gerada com o programa VMD (1, 2). Os vizinhos do resíduo N podem ser selecionados, segundo a sintaxe do VMD, por meio do comando* (same resid as (name CA and within 8 of (name CA and resid N))).

$$m_{\text{est}}(r) = n_{\text{gás}} 4\pi \int_0^r g(r) r^2 dr$$

Seja a equação acima aplicada a uma estrutura de proteína, e seja $m(r)$ o número de vizinhos à distância menor ou igual a r contado para um resíduo, quando são consideradas

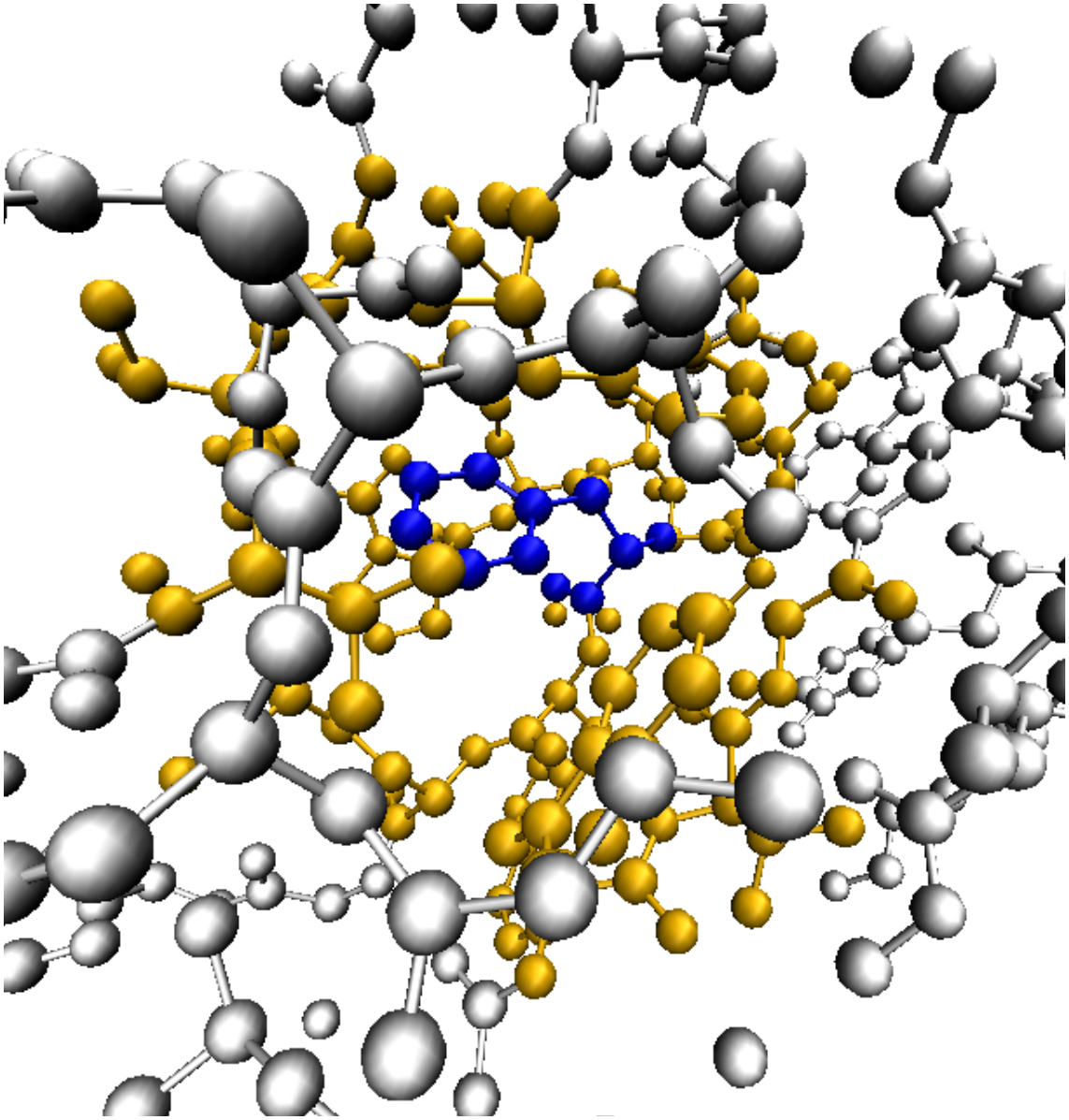

Figura 4.6 – *Ilustração do método de contagem de vizinhos baseado nas distâncias entre todos os pares de átomos. Os átomos do resíduo de interesse estão ilustrados em azul, e todos os resíduos que contém pelo menos um átomo a menos de $5\mathring{A}$ de algum átomo do resíduo de interesse estão ilustrados em laranja. Figura gerada com o programa VMD (1, 2). Os vizinhos do resíduo N podem ser selecionados, segundo a sintaxe do VMD, por meio do comando* (same resid as (within 5 of resid N)).

as distâncias entre todos os átomos (tal como na figura 4.6). Multiplicando pelo número N de resíduos, para obter o número total de vizinhos contado por todos os resíduos, e isolando $g(r)$, vem:

$$g_{\mathsf{R}}(r) = \frac{1}{r^2}\frac{d}{dr}\left[\frac{N m_{\mathsf{est}}(r)}{4\pi n_{\mathsf{gás}}}\right] = \frac{1}{r^2}\frac{d}{dr}\left[\frac{M_{\mathsf{est}}(r)}{4\pi n_{\mathsf{gás}}}\right]$$

A equação obtida depende de $n_{\text{gás}}$, que depende da estrutura sendo estudada. Absorvendo os fatores constantes:

$$g_{\text{R}}(r) = \frac{A}{r^2}\frac{dM_{\text{est}}(r)}{dr} \tag{4.1.9}$$

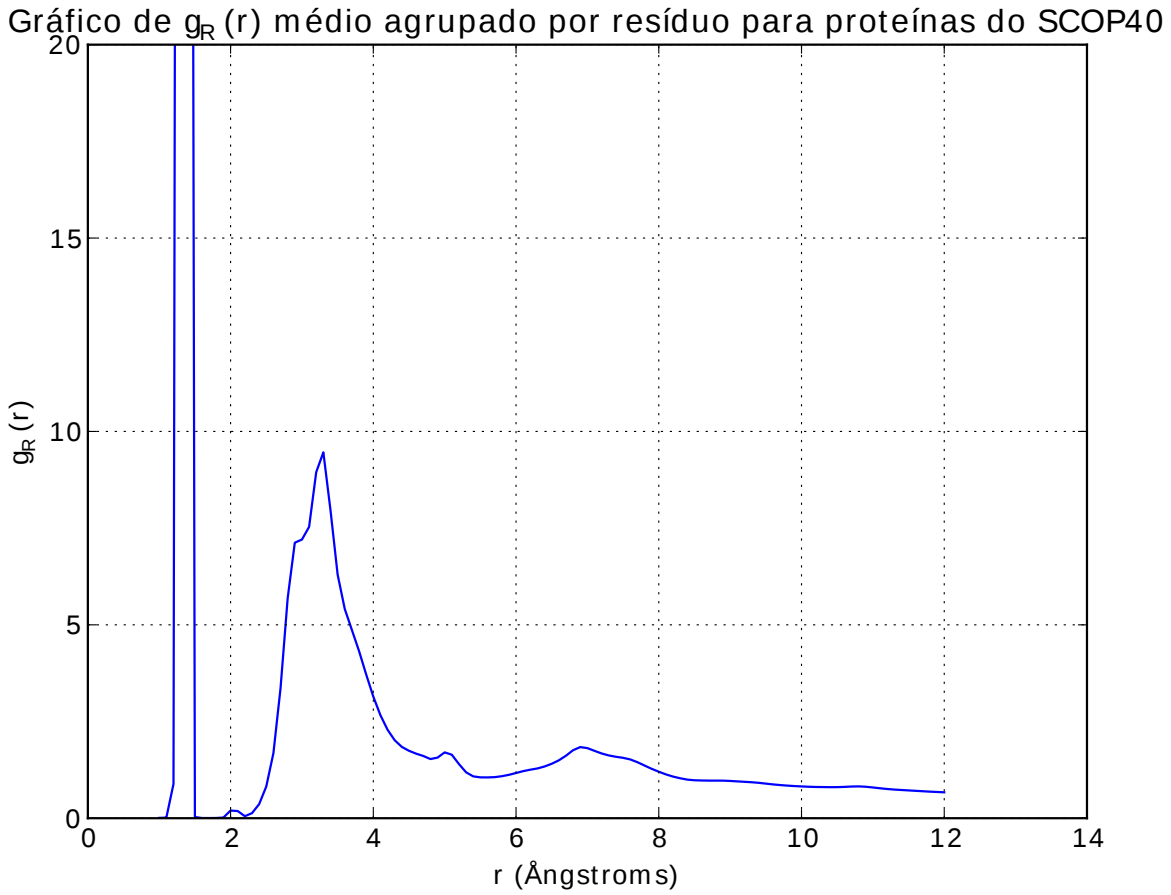

Figura 4.7 – *Gráfico de $g_R(r)$ médio contra r para todas as proteínas da base SCOP40. O pico em torno de $r = 1,5\text{Å}$, correspondente aos vizinhos covalentes, atinge um valor máximo em torno de 100, e foi suprimido para preservar a clareza da figura.*

Para permitir a comparação entre estruturas diferentes, lançamos mão da hipótese de que $g_R(r)$ deve ter limite 1 para grandes distâncias, e calculamos A para cada estrutura de modo que esta restrição seja obedecida. Tomando a média entre as curvas obtidas, obtemos um perfil geral da distribuição dos vizinhos em torno de cada resíduo, apresentado na figura 4.7:

O gráfico é coerente com os resultados já apresentados e também, em linhas gerais, com resultados da literatura tal como (71). A curva apresenta os picos de densidade nas posições esperadas: ligações covalentes em torno de $1,5\text{Å}$ e interações não-covalentes, principalmente ligações de hidrogênio, entre 2Å e 6Å aproximadamente. O *cutoff* de 5Å utilizado por Greene

et al. e del Sol *et al.*, se mostra inadequado e é suplantado pelo valor 6Å, que engloba completamente as regiões dos dois primeiros picos de estruturação. Sua escolha, todavia, acarreta em ignorar uma região de estruturação entre 6Å e 8Å, cuja densidade é diferente de 1 ainda que sua magnitude seja menos significativa do que os picos anteriores. Não está claro se existe prejuízo teórico nesta escolha, principalmente em trabalhos que dependam de uma descrição detalhada da estruturação.

Do ponto de vista estrutural, os dados fornecem uma definição quantitativa de "vizinho" de um resíduo do ponto de vista da organização da distribuição atômica em estruturas de proteína, na medida em que esta se relaciona aos potenciais interatômicos. De um ponto de vista "prático", as análises realizadas ao longo deste trabalho que se constroem sobre a definição de vizinho, em particular aquela descrita na próxima seção, se mostram razoavelmente independentes do *cutoff* adotado, dentro dos limites tipicamente encontrados na literatura. Observa-se certa variação individual para os resultados quando se adota a definição de vizinho dependente dos C_α *versus* dependente de todos os átomos, mas as tendências gerais são respeitadas. Aqui, consideraremos *vizinhos* todos os pares de resíduos tais que pelo menos um átomo do primeiro está a no máximo 6Å de algum átomo do segundo.

Munidos deste conceito, oferecemos a seguinte definição, objetivo desta seção:

Definição 11 *Seja uma proteína composta de N resíduos, para a qual dispõe-se das posições espaciais de todos os átomos que a compõem (sua* estrutura*). Esta proteína pode ser representada por uma* rede*, ou* grafo*, definido por sua matriz de adjacência* $\boldsymbol{A}$*.*

A matriz de adjacência é tal que $\boldsymbol{A}$ *é uma matriz $N \times N$, e seus elementos obedecem $\boldsymbol{A}_{ij} = 0$ a não ser que pelo menos um átomo do resíduo i esteja a no máximo 6Å de distância de algum átomo do resíduo j, caso em que vale $\boldsymbol{A}_{ij} = 1$. Diz-se então que os resíduos i e j são* vizinhos *ou que eles estão* ligados*.*

4.2 Previsão do Fluxo de Calor por meio de Descritores de Rede

Uma vez estabelecido o método de construção, investigamos a capacidade preditiva destes modelos de rede no que concerne ao fluxo de energia vibracional em proteínas. Evidentemente, *fluxo de energia vibracional* se refere a um conceito cuja medida pode ser operacionalizada através de uma miríade de maneiras, sejam elas teóricas ou experimentais. Utilizamos aqui simulações de dinâmica molecular, associadas à técnica de ATD (vide seção 2.3). O custo computacional dos experimentos de ATD torna inviável a sua realização para um conjunto

de proteínas do tamanho do SCOP40. Analisamos, alternativamente, os resultados de experimentos de ATD referentes a um conjunto de xilanases da família 11 (vide seção 1.2), fornecidos generosamente por Heloisa Muniz, colega de laboratório a quem o autor é agradecido, e publicados na sua dissertação de mestrado (72), na qual se encontram também detalhes metodológicos e análises complementares.

Nos concentramos aqui em um aspecto particular dos experimentos de ATD: a temperatura final atingida pela proteína. Podemos explorar visualmente os resultados de um experimento de ATD graficando a temperatura final contra a posição do resíduo na estrutura primária. A curva resultante se assemelha a uma série temporal, em parte porque a temperatura final atingida varia razoavelmente suavemente de cada posição para a seguinte. Certamente isso se deve ao fato de que posições consecutivas na cadeia são unidas por ligação peptídica, formando o *backbone* da estrutura, e a influência de cada posição se estende parcialmente sobre as posições vizinhas. Um exemplo é apresentado na figura 4.8.

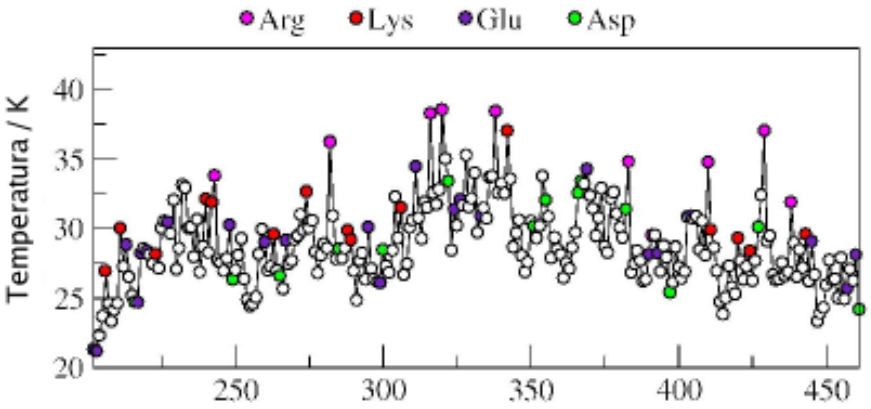

Figura 4.8 – *Exemplo de curva de temperatura final* versus *posição do resíduo aquecido na estrutura primária, com alguns resíduos destacados e suas respectivas identidades. A identidade do resíduo correlaciona-se moderadamente com a sua capacidade de aquecer a proteína como um todo, mas não explica por si só a forma da curva. Reproduzido com autorização de (10).*

Durante um experimento de ATD, um resíduo, aquecido isoladamente, transfere energia térmica para seus vizinhos mais frios, e através deles, gradativamente, para o resto da estrutura. Dada uma estrutura e condições experimentais fixas, cada resíduo particular demonstra uma capacidade distinta de transferir calor para a proteína como um todo, e o aquecimento de cada resíduo resulta, após o mesmo intervalo de tempo, numa temperatura final da proteína distinta. Este conceito é brevemente discutido na seção 2.3, na qual é introduzido o conceito de *bom difusor*. Em primeira aproximação, tal observação não é surpreendente, visto que cada tipo de aminoácido compreende um número diferente de átomos, arranjados de maneira diferente e com uma massa total diferente. Levar um conjunto de partículas a uma dada temperatura

requer uma quantidade diferente de energia dependendo das características deste conjunto, resumidas num parâmetro denominado *capacidade térmica*, e isto por si só justificaria, a princípio, as temperaturas finais diferentes. Para efeito de ilustração, uma lista das capacidades térmicas experimentais de cada aminoácido pode ser encontrada em (73).

Uma observação experimental, todavia, introduz uma complicação: resíduos idênticos ou com capacidades térmicas semelhantes mas em posições diferentes na cadeia principal geram temperaturas finais diferentes quando aquecidos. Logo, a explicação dos experimentos de ATD requer um nível maior de sofisticação.

A resposta deve residir no fato de que o ATD é uma técnica de não-equilíbrio. Um gradiente de temperatura é estabelecido e mantido durante toda a (curta) simulação, e a energia térmica se transfere localmente. O banho térmico quente, ao qual apenas o resíduo de interesse é acoplado, fornece energia para este em média na mesma taxa em que ele cede energia para seus vizinhos. A energia *total* que será cedida para a proteína, assim, dependerá da intensidade do *acoplamento* térmico entre o resíduo de interesse e seus vizinhos imediatos, e destes para com seus vizinhos, assim por diante. Sob esse ponto de vista, resíduos fortemente acoplados serão aqueles que resultarão em temperaturas finais mais altas. O foco desta seção é a investigação dessa asserção.

O conceito de "acoplado" é propositalmente deixado sem definição. A hipótese aventada pelo autor é a de que estas posições podem ser identificadas valendo-se do formalismo de grafos. Em particular, construído o grafo que representa cada estrutura, acreditamos que a *centralidade* de cada resíduo (vide seção 3.1 e definições associadas) pode fornecer uma medida deste acoplamento. Conforme previamente discutido, o conceito de centralidade pode ser operacionalizado de mais de uma maneira. Investigaremos paralelamente a possível relação das três formalizações distintas de centralidade com a temperatura final dos resíduos.

Tal hipótese pode ser testada, em primeira aproximação, pela simples e direta sobreposição das curvas de temperatura final *versus* resíduo e de centralidade *versus* resíduo. Todavia, a comparação direta entre duas grandezas de natureza distinta, uma das quais adimensional, não é matematicamente bem definida. Contornamos esta dificuldade graficando, no lugar de cada grandeza, a magnitude de sua distância em relação à média, que é um número sem dimensão. Diz-se, então, que comparamos as grandezas *normalizadas pela média*. Esta normalização, o *Z-Score*, é assim definida:

Definição 12 *Seja x uma variável aleatória, da qual se dispõe de N observações. Denote por μ a média da amostra, e por σ seu desvio padrão. Então, valerá:*

$$\mu(x) = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\sigma^2(x) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu(x))^2$$

Denote por $z(x)$ *o* Z-Score *normalizado de* x*. Então, valerá:*

$$z(x)_i = \frac{x_i - \mu(x)}{\sigma(x)}$$

Aplicamos esta transformação para os valores de temperatura final e para os valores de centralidade, e graficamos os dois na mesma escala. Produzimos tais gráficos para um conjunto de cinco xilanases, para as quais possuímos dados de experimentos de ATD. Seus identificadores na base PDB são: 1F5J, 1M4W, 1XNB, 2VUJ e 2VUL. Comparamos tais dados com dados de centralidade de grau, centralidade de intermediação e centralidade de proximidade para as respectivas estruturas, e os gráficos obtidos *para o caso da centralidade de proximidade* mostram forte evidência da correlação entre as duas propriedades. Apresentamos os gráficos de sobreposição para a centralidade de proximidade nas figuras 4.9-4.13.

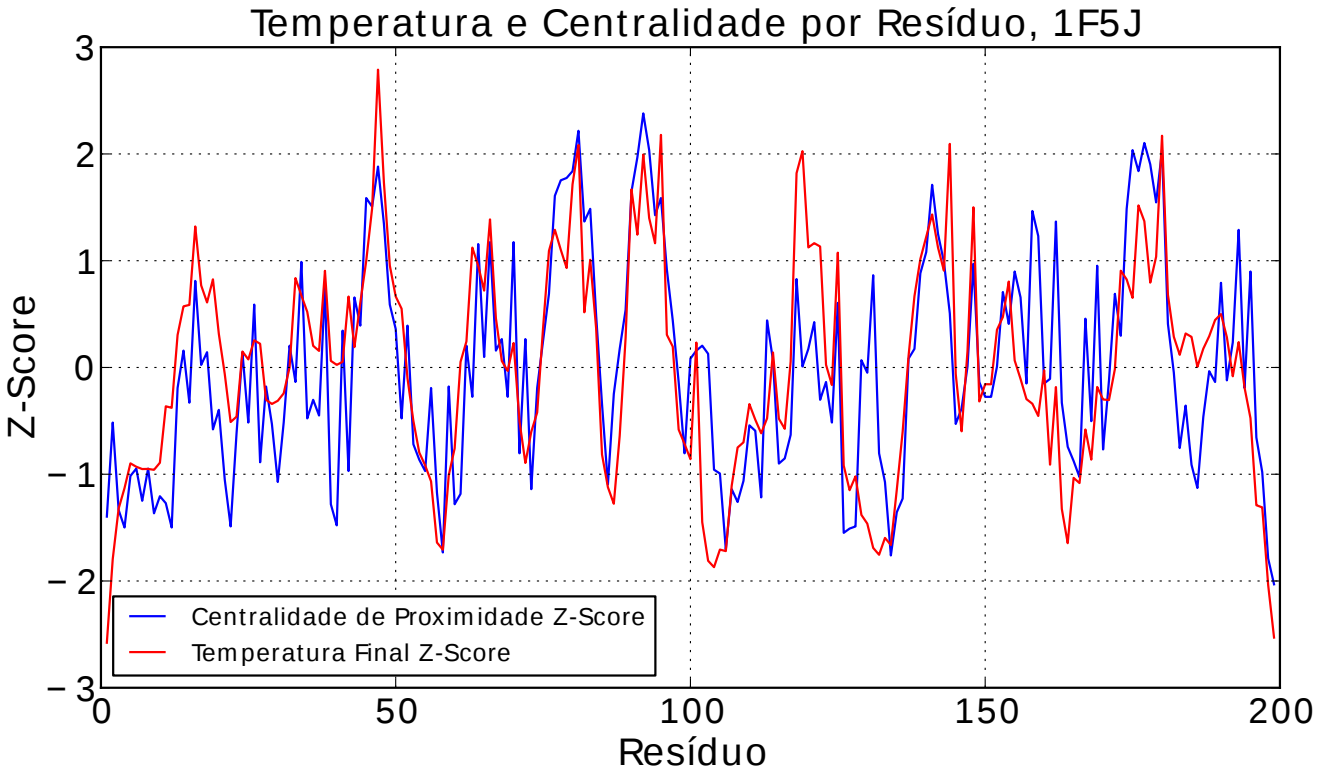

Figura 4.9 – *Sobreposição entre as curvas normalizadas de temperatura final por resíduo e centralidade de proximidade por resíduo para a xilanase de identificador 1F5J.*

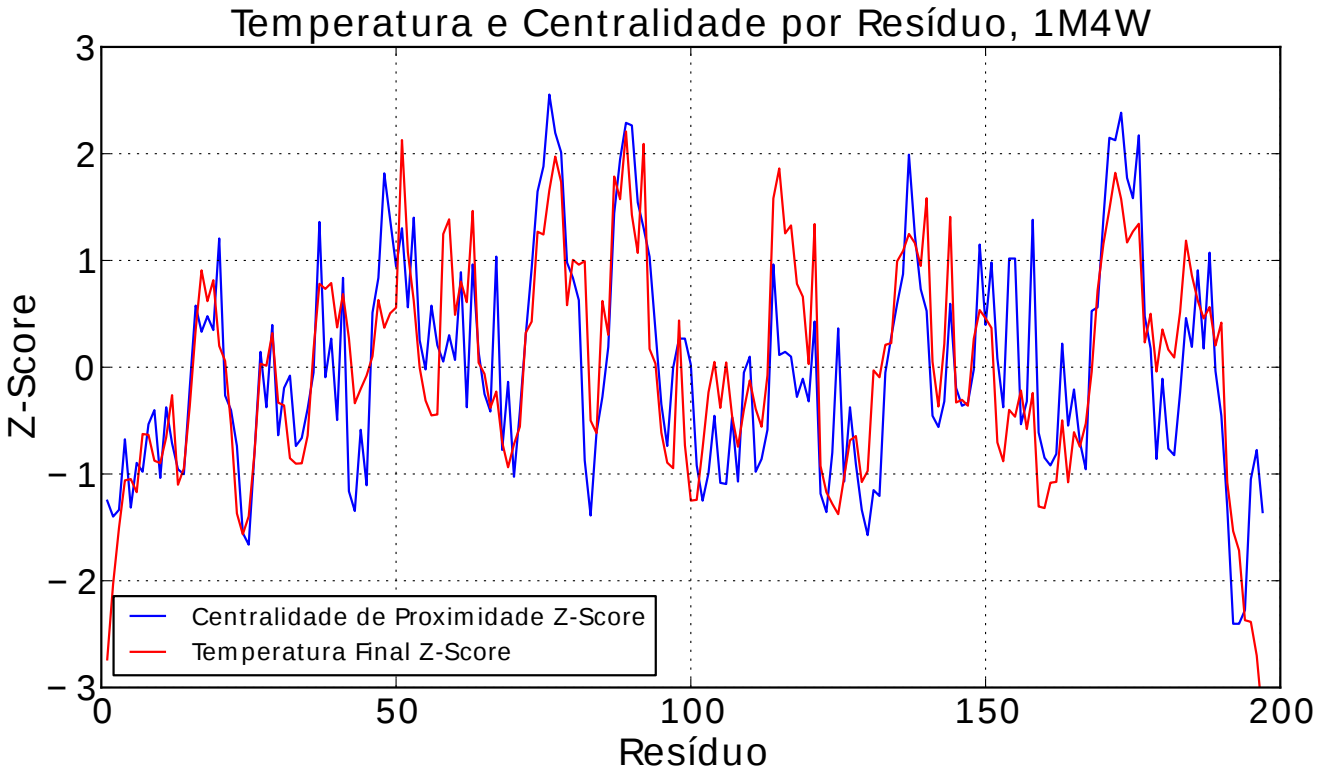

Figura 4.10 – *Sobreposição entre as curvas normalizadas de temperatura final por resíduo e centralidade de proximidade por resíduo para a xilanase de identificador 1M4W.*

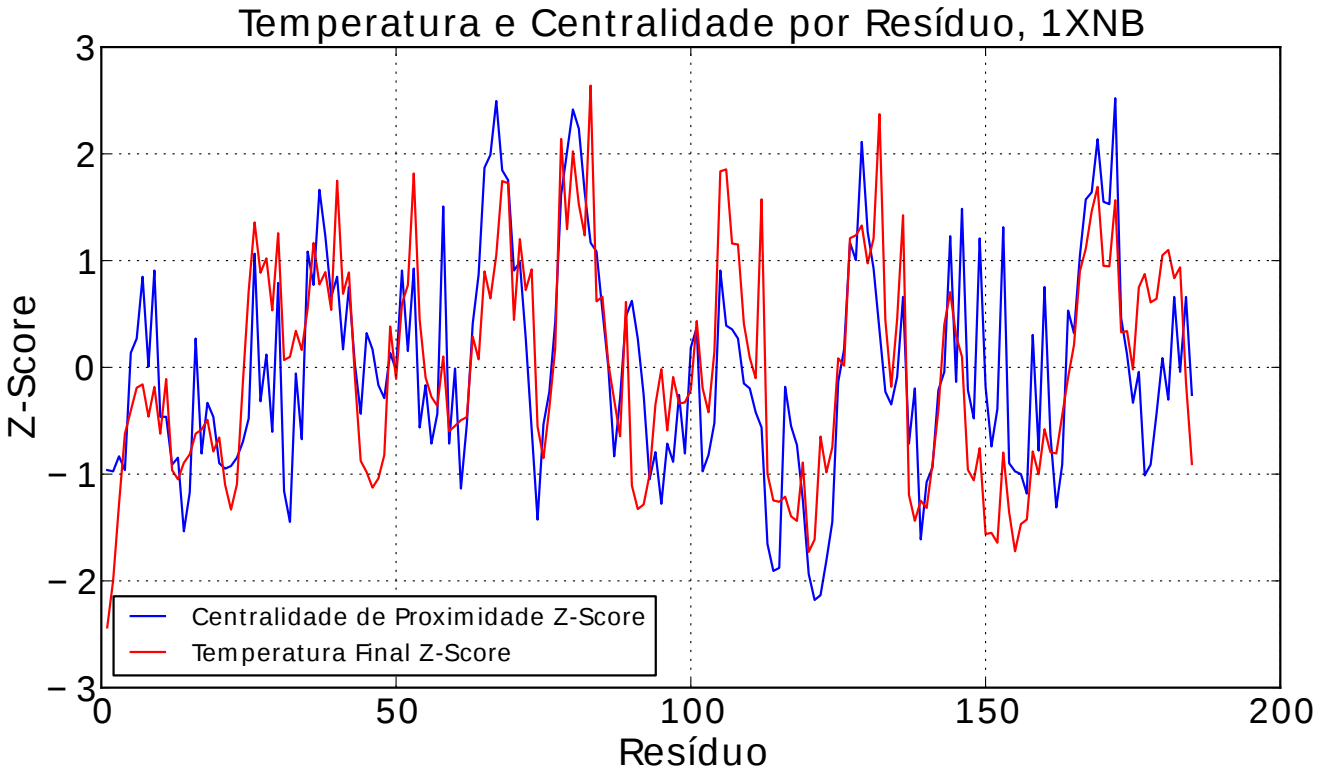

Figura 4.11 – *Sobreposição entre as curvas normalizadas de temperatura final por resíduo e centralidade de proximidade por resíduo para a xilanase de identificador 1XNB.*

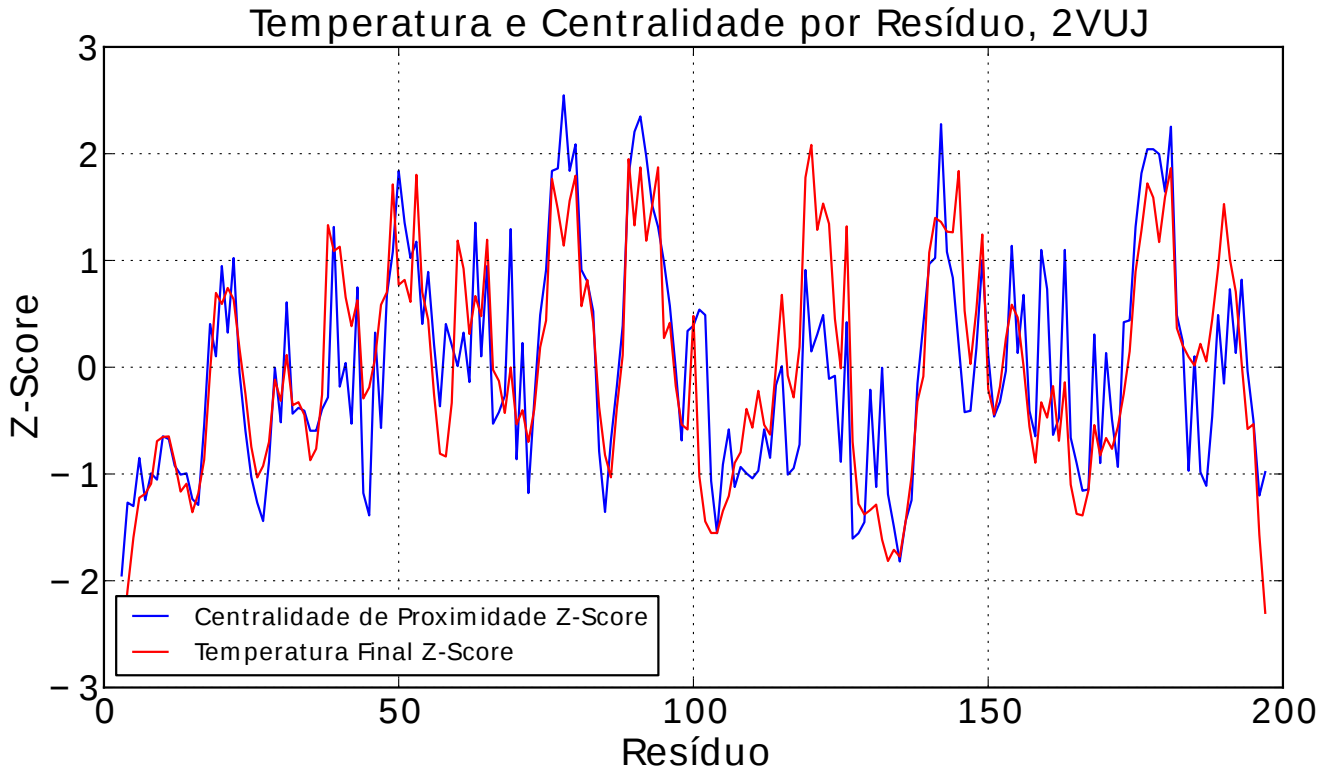

Figura 4.12 – *Sobreposição entre as curvas normalizadas de temperatura final por resíduo e centralidade de proximidade por resíduo para a xilanase de identificador 2VUJ.*

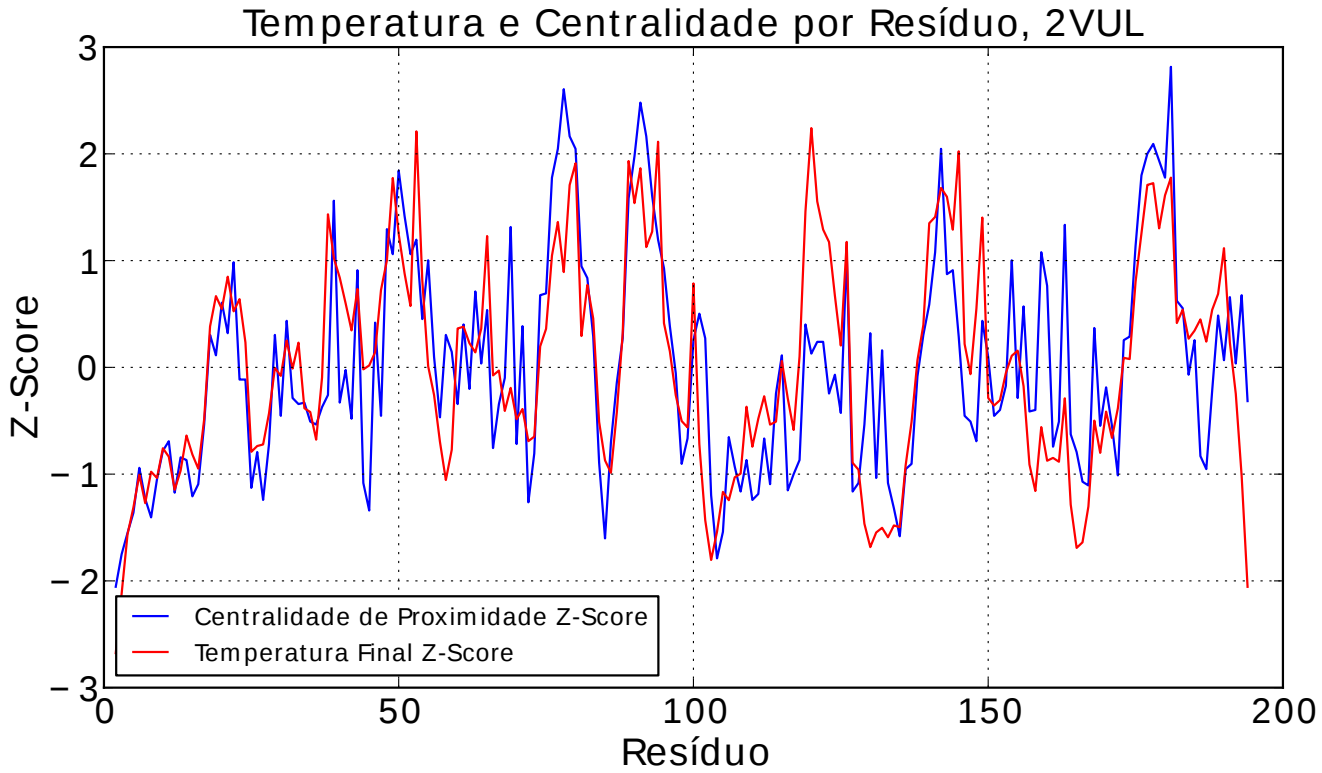

Figura 4.13 – *Sobreposição entre as curvas normalizadas de temperatura final por resíduo e centralidade de proximidade por resíduo para a xilanase de identificador 2VUL.*

Das figuras, sobressaem características gerais que são obedecidas no conjunto dos experimentos: vê-se que a variação da centralidade de proximidade é razoavelmente menos suave que a variação da temperatura, mas acompanha claramente a tendência da mesma ao longo da cadeia. Nos resíduos dos extremos N-terminais e C-terminais, tanto a temperatura final quanto a centralidade são significativamente abaixo da média.

Vê-se que os perfis de centralidade por resíduo, assim como os perfis de temperatura final por resíduo, são notavelmente semelhantes entre si, sugerindo a possibilidade da identificação de famílias de dobramentos ou possivelmente de domínios funcionais para a anotação de novas estruturas através da distribuição de centralidade por resíduo. A figura 4.14 é uma análise particularmente interessante nesta linha de investigação. Nela, destacamos os resíduos para os quais a temperatura final é expressivamente *diferente* da centralidade de proximidade. Estes resíduos são *outliers*, no que concerne à análise aqui apresentada, e não aparentam relação entre si quando as sequências são comparadas. Contudo, quando as estruturas de cada xilanase são colorizadas de forma a evidenciar os *outliers*, torna-se claro que regiões homólogas da estrutura exibem resposta térmica semelhante. Notamos também que, de modo geral, resíduos próximos à fenda catalítica tendem a demonstrar resposta térmica menor do que a prevista por sua (alta) centralidade.

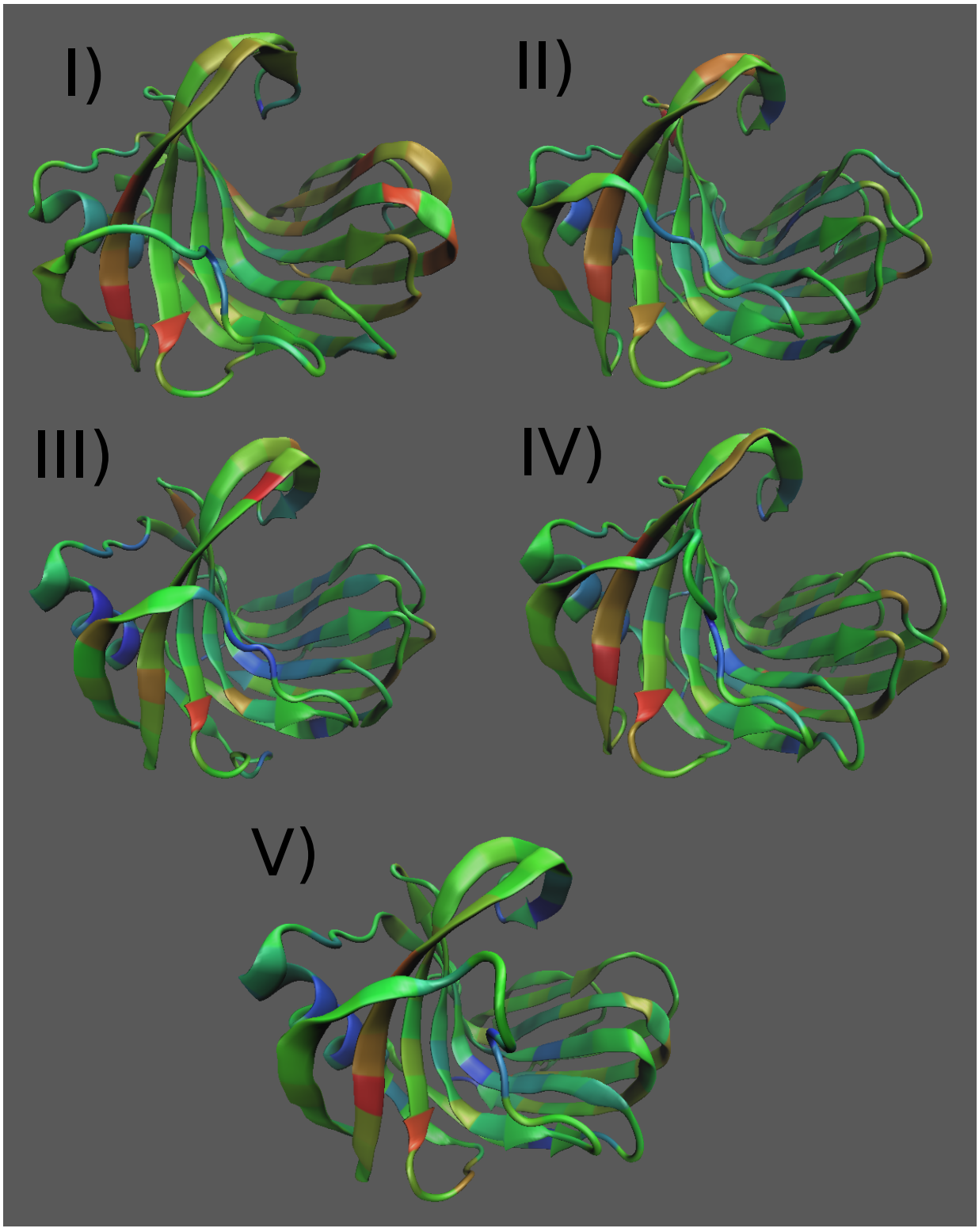

Figura 4.14 – *Comparação entre os* outliers *da correlação entre centralidade e temperatura final para o conjunto de xilanases a saber: (I) 1F5J; (II) 1M4W; (III) 1XNB; (IV) 2VUJ; (V) 2VUL. Cores quentes representam resíduos cuja capacidade de difundir calor é significativamente maior do que prevê sua centralidade de proximidade. Cores frias representam o oposto. É notável a concentração de resíduos cuja resposta é menor do que a prevista na região da fenda catalítica.*

A comparação entre as curvas de centralidade de intermediação ou centralidade de grau e de temperatura final não sugere correlação significativa, e os gráficos não são mostrados aqui. Avançamos a investigação graficando a dispersão das relação entre a centralidade e a temperatura final para cada proteína, e calculando os respectivos *coeficientes de correlação de Pearson*. O coeficiente de correlação de Pearson, ou r, é uma medida da correlação linear entre duas variáveis, e é assim definido:

Definição 13 *Sejam x e y duas variáveis aleatórias, para cada uma das quais dispomos de N observações. Seja $\mu(x)$ a média das observações de x e $\mu(y)$ a média das observações de y, e $\sigma(x)$ e $\sigma(y)$ os respectivos desvios padrão. A* **covariância** *do par x, y é assim definida:*

$$Cov(x,y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu(x))(y_i - \mu(y))$$

O **Coeficiente de Correlação de Pearson**, *r, é definido:*

$$r = \frac{Cov(x,y)}{\sigma(x)\sigma(y)}$$

O mesmo pode ser expresso em termos dos Z-Score de x e y. Absorvendo os desvios padrão no somatório, vem:

$$r = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i - \mu(x)}{\sigma(x)}\right)\left(\frac{y_i - \mu(y)}{\sigma(y)}\right)$$

Pela definição 12:

$$r = \frac{1}{N}\sum_{i=1}^{N} z(x)_i z(y)_i$$

O coeficiente assim definido tem valor contido no intervalo $[-1; 1]$, tal que valores absolutos de r próximos de 1 indicam dependência linear entre as duas variáveis, e valores próximos de zero indicam independência linear. Pode ser demonstrado[§] que o valor de r^2, denominado *coeficiente de determinação*, é uma medida da fração da variação de y que pode ser explicada pela variação de x. Calculamos o coeficiente de correlação e o coeficiente de determinação

[§] Admitimos aqui que as demonstrações dos conceitos elementares de estatística empregados podem ser dispensadas sem prejuízo para o conteúdo deste trabalho, favorecendo a clareza do texto. O leitor interessado é direcionado a (74).

para as proteínas estudadas, e apresentamos os resultados obtidos a seguir, bem como os gráficos de dispersão, nas figuras 4.15-4.19.

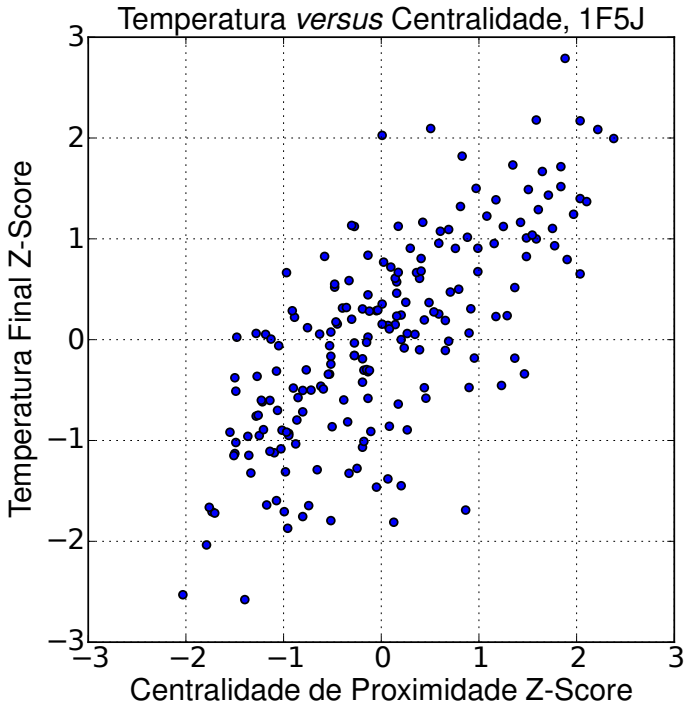

Figura 4.15 – *Gráfico de dispersão para a temperatura final* versus *centralidade para a proteína 1F5J.*

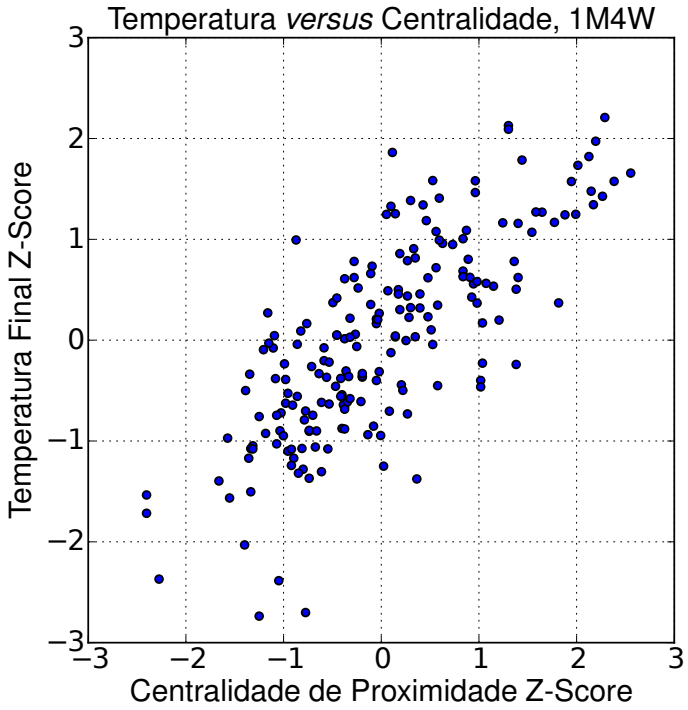

Figura 4.16 – *Gráfico de dispersão para a temperatura final* versus *centralidade para a proteína 1M4W.*

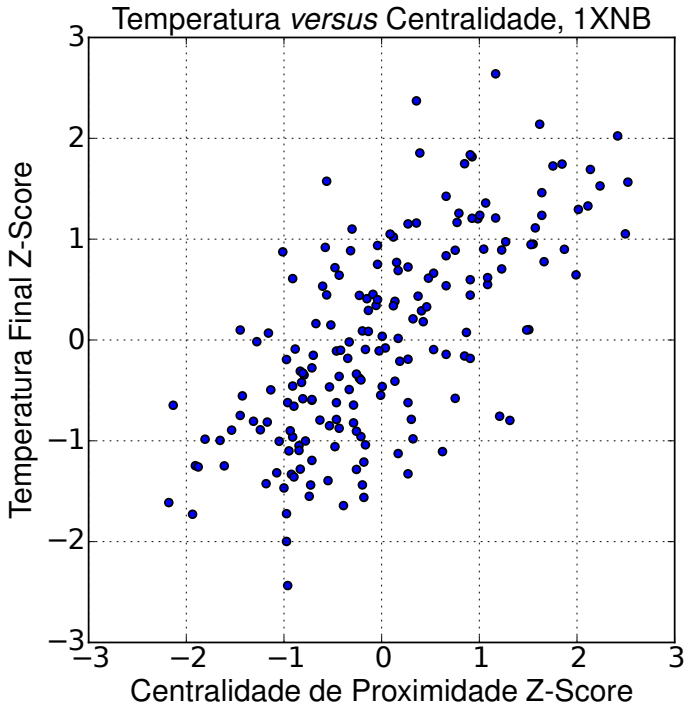

Figura 4.17 – *Gráfico de dispersão para a temperatura final* versus *centralidade para a proteína 1XNB.*

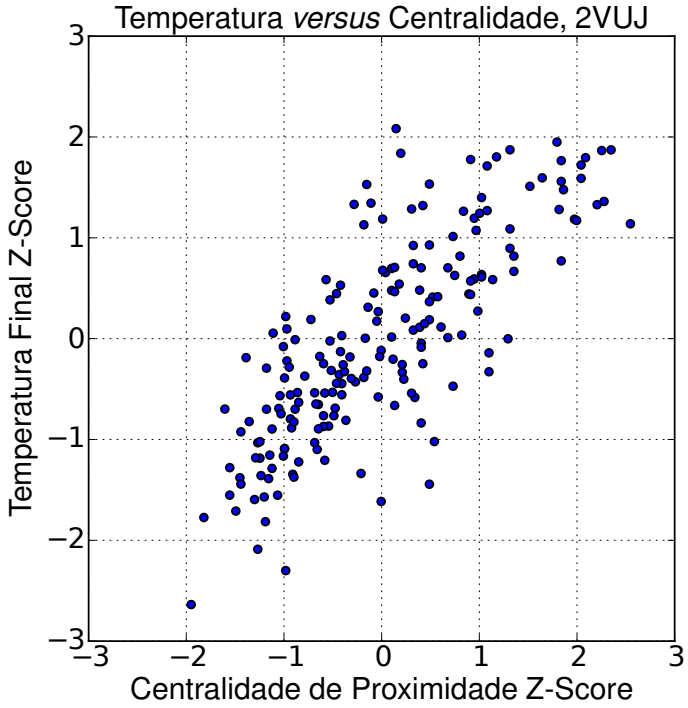

Figura 4.18 – *Gráfico de dispersão para a temperatura final* versus *centralidade para a proteína 2VUJ.*

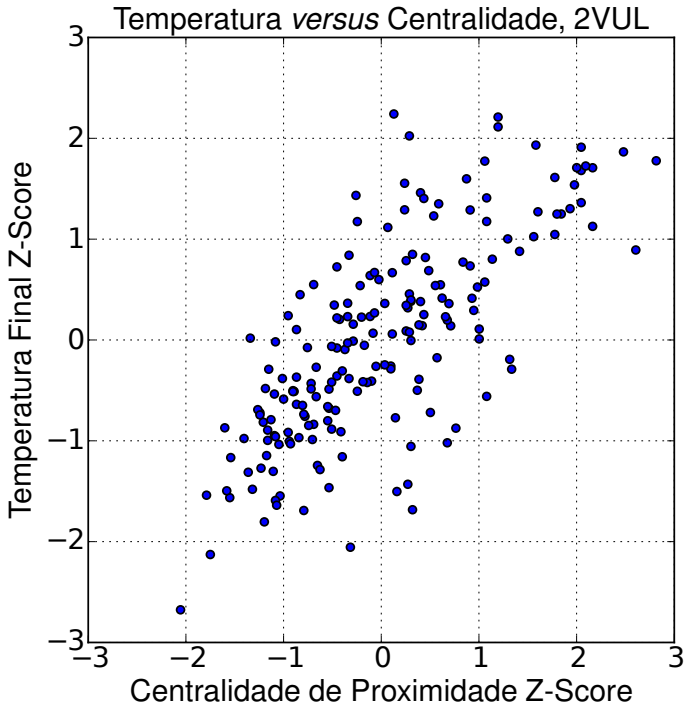

Figura 4.19 – *Gráfico de dispersão para a temperatura final* versus *centralidade para a proteína 2VUL.*

As dispersões evidenciam a relação aproximadamente linear entre a centralidade e a temperatura final. Apresentamos os coeficientes de correlação na tabela 4.1.

Tabela 4.1 – *Magnitude das correlações entre as medidas de centralidade e a temperatura final de cada resíduo, para o conjunto de xilanases estudadas.*

Medida	Proteína									
	1F5J		1M4W		1XNB		2VUJ		2VUL	
	r	r^2	r	r^2	r	r^2	r	r^2	r	r^2
Centr. de Proximidade (C_C)	**0,73**	**0,53**	**0,77**	**0,59**	**0,69**	**0,47**	**0,79**	**0,62**	**0,75**	**0,56**
Centr. de Intermediação (C_B)	0,62	0,38	0,61	0,37	0,59	0,35	0,64	0,41	0,61	0,37
Centr. de Grau (C_D)	0,69	0,48	0,73	0,54	0,71	0,50	0,73	0,54	0,70	0,50

Não é surpreendente que as três medidas de centralidade produzam correlações parecidas, dado que as medidas são sabidamente correlacionadas entre si (vide, por exemplo, (75)). Ainda assim, a centralidade de proximidade fornece a melhor estimativa entre as três medidas, explicando uma porcentagem média de 55% da variação observada na temperatura final dos resíduos para as cinco proteínas. Consideramos estes dados evidência de que a centralidade de proximidade é uma expressão adequada do conceito de *acoplamento* térmico entre cada resíduo e o resto da estrutura. Este resultado é corroborado por Sabidussi em (76), que demonstra que os nós com alta centralidade de proximidade são por definição aqueles que mais rapidamente disseminam um sinal ao longo de toda a rede, associando assim centralidade de proximidade

com eficiência na propagação de informação. Discutimos, no próximo capítulo, as implicações deste resultado.

CAPÍTULO 5

Conclusões

Oferecemos aqui algumas considerações sobre as análises apresentadas.

Em primeiro lugar, devemos reafirmar aqui, de maneira concisa, os resultados apresentados no capítulo 4, seção 4.1. Analisamos um conjunto expressivamente numeroso e representativo de estruturas experimentais, e demonstramos que a distância média dentro da qual cada átomo observa estruturação em torno de si vale pouco menos que 8 Ångstroms. Para distâncias maiores do que esta, não se observa estruturação que possa distinguir estruturas proteicas de distribuições aleatórias.

Quando são considerados apenas os carbonos C_α, revela-se estruturação que persiste até a distância de aproximadamente 11 Ångstroms, provavelmente alcançando, a partir do C_α do resíduo (n), o C_α do resíduo $(n+3)$, e valores de *cutoff* sensatos para a identificação de resíduos vizinhos são de 7Å ou possivelmente de 11Å. Considerando todos os átomos explicitamente, demonstramos que há não mais que duas escolhas sensatas para um *cutoff* que tencione identificar resíduos *vizinhos* entre si: 6Å ou 8Å, a depender do nível de detalhe estrutural pretendido.

Utilizamos os resultados descritos para construir as redes correspondentes às estruturas de um conjunto de xilanases, para as quais dispúnhamos de dados referentes a experimentos de ATD, e demonstramos capacidade preditiva dos modelos de rede em relação ao fluxo de calor.

É justo lembrar que os experimentos de ATD são modificações de simulações de dinâmica molecular, e, ainda que não possam ser classificados como esforços puramente teóricos, na medida em que simulações habitam o regime de existência ambíguo dos experimentos *in silico*, certamente não carregam o peso de resultados experimentais. Os modelos de rede, por outro lado, são abstrações cuja existência e emprego só são justificados na medida em que estes são capazes de prever resultados experimentais. A correlação obtida entre os resultados dos experimentos de ATD e as medidas de centralidade derivadas dos modelos de rede poderia, sob esse ponto de vista, ser encarada como vazia de significado. A ponte aqui demonstrada

entre as duas técnicas atua no sentido de fortalecer a justificativa para o emprego de ambas. A previsão do fluxo de energia em sistemas moleculares complexos é notoriamente complicada, e o fato de que aspectos deste problema podem ser adequadamente modelados por uma construção simples tal qual uma rede é um argumento importante a favor da abstração. Existe ampla literatura apoiando o emprego de modelos de rede no estudo de proteínas, cujos principais resultados são discutidos no capítulo 3, e a técnica de ATD exibe importantes sucessos na previsão de resultados experimentais, discutidos no capítulo 2. É certo que trabalhos futuros certamente explorarão a correlação encontrada, testando os limites de sua validade e investigando a gama de possíveis aplicações, e o autor acredita possível que a análise dos *outliers* encontre aplicação na engenharia de proteínas termoestáveis. Não há dúvidas de que trabalhos de cunho experimental trarão os resultados mais importantes.

O autor tenciona mencionar algumas das linhas de investigação seguidas ao longo deste trabalho que não produziram resultados de utilidade imediatamente óbvia, no intuito de melhor direcionar futuras investigações. Uma significativa investigação foi dedicada à hipótese de que os resultados de ATD seriam melhor explicados por meio da incorporação de informações dinâmicas. Foram produzidas simulações longas de equilíbrio para cada uma das xilanases estudadas, e grafos foram construídos a partir de cada *frame* das mesmas. Naturalmente, este conjunto de grafos apresenta um grau de variação, pois algumas interações não-covalentes não permanecem ligadas ao longo de toda a simulação. O conjunto todo pode ser codificado em um único grafo, codificando esta variação nos *pesos* das arestas, numa extensão da formalização não apresentada aqui. As medidas de centralidade podem ser redefinidas em termos das arestas com pesos, produzindo valores que em teoria melhor descrevem o acoplamento entre cada resíduo e o resto da estrutura. Contudo, tal linha de investigação não produziu resultados significativamente melhores (nem significativamente diferentes, em alguns casos), e foi abandonada.

Os resultados positivos apresentados, por outro lado, suscitam questões imediatas que irão guiar trabalhos futuros. Há de se verificar, por exemplo, se a seleção natural tende a favorecer resíduos de alta capacidade térmica em posições centrais, e investigar a relação entre a distribuição de resíduos centrais e/ou bons difusores e a termoestabilidade de uma dada estrutura, um dos objetivos iniciais deste trabalho. Um exemplo intrigante é o resultado apresentado em (56), que identifica com alta taxa de sucesso resíduos que fazem parte do sítio catalítico por meio de seus valores de centralidade de proximidade. É razoável supor, com base nesta observação, que os resíduos do sítio catalítico em geral exibam também o atributo de ser bons difusores de calor, e que tal propriedade seja importante para a manutenção da conformação ativa de cada estrutura. É opinião do autor que essa linha de investigação

pode lançar uma perspectiva interessante sobre proteínas cuja termoestabilidade não seja bem explicada de outro modo.

REFERÊNCIAS

1 HUMPHREY, W.; DALKE, A.; SCHULTEN, K. VMD – visual molecular dynamics. *Journal of Molecular Graphics*, v. 14, n. 1, p. 33–38, 1996.

2 STONE, J. An efficient library for parallel ray tracing and animation, 1998. 89 p. Thesis (Master of Science) - Computer Science Department, University of Missouri-Rolla, 1998.

3 PROTEIN folding schematic. 2008. Disponível em: <http://en.wikipedia.org/wiki/File:Protein_folding_schematic.png>. Acesso em: 05 dec. 2012.

4 PORCINE elastase crystal. 2009. Disponível em: <http://commons.wikimedia.org/wiki/File:Protein_Crystal_Growth_Porcine_Elastase.jpg>. Acesso em: 5 dec. 2012.

5 DIFFRACTION image of lysozyme crystal. 2009. Disponível em: <http://commons.wikimedia.org/wiki/File:Lysozym_diffraction.png>. Acesso em: 5 dec. 2012.

6 RICHARDSON, J. S. Crystallographic electron density. 2010. Disponível em: <http://commons.wikimedia.org/wiki/File:Helix_electron_density_myoglobin_2nrl_17-32.jpg>. Acesso em: 5 dec. 2012.

7 FRISHMAN, D.; ARGOS, P. Knowledge-based secondary structure assignment. *Proteins:* structure, function and genetics, v. 23, n. 4, p. 566–579, 1995.

8 VARSHNEY, A.; BROOKS, F. P.; WRIGHT, W. V. Linearly scalable computation of smooth molecular surfaces. *IEEE Computer Graphics and Applications*, v. 14, p. 19–25, 1994.

9 GANJALIKHANY, M.; RANJBAR, B.; TAGHAVI, A.; MOGHADAM, T. T. Functional analysis of Candida antartica lipase B, 2012. Disponível em: <http://commons.wikimedia.org/wiki/File:Functional-Motions-of-Candida-antarctica-Lipase-B-A-Survey-through-Open-Close-Conformations-pone.0040327.s014.ogv>. Acesso em: 18 dec. 2012.

10 MARTíNEZ, L. Simulações de dinâmica molecular dos receptores do hormônio tireoideano, 2007. 168 p., Tese (Doutorado em Química) - Instituto de Química, Universidade Estadual de Campinas, Campinas, 2007.

11 DIAGRAM of a partial mesh network. 2008. Disponível em: <http://commons.wikimedia.org/wiki/File:PartMeshNetwork.svg>. Acesso em: 25 dec. 2012.

12 COETZEE, D. Description of a graph isomorphism. 2007. Disponível em: <http://commons.wikimedia.org/wiki/File:Graph_isomorphism.svg>. Acesso em: 28 jan. 2013.

13 NELSON, D. L.; COX, M. M. *Lehninger principles of biochemistry*. 4th ed. New York: W. H. Freeman, 2004.

14 ONUCHIC, J.; WOLYNES, P. Theory of protein folding. *Current Opinion in Structural Biology*, v. 14, n. 1, p. 70–75, 2004.

15 DILL, K.; OZKAN, S.; WEIKL, T.; CHODERA, J.; VOELZ, V. The protein folding problem: when will it be solved? *Current Opinion in Structural Biology*, v. 17, n. 3, p. 342–246, 2007.

16 DURBIN, S.; FEHER, G. Protein crystallization. *Annual Review of Physical Chemistry*, v. 47, p. 171–204, 1996. DOI: 10.1146/annurev.physchem.47.1.171.

17 KENDREW, J. C.; BODO, G.; DINTZIS, H. M.; PARRISH, R. G.; WYCKOFF, H. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*, London, v. 181, n. 4610, p. 662–666, 1958.

18 WUTHRICH, K. Protein structure determination in solution by Nuclear Magnetic Resonance Spectroscopy. *Science*, Washington, v. 243, n. 4887, p. 45–50, 1989.

19 YU, H. Extending the size limit of protein Nuclear Magnetic Resonance. *Proceedings of the National Academy of Sciences*, v. 96, n. 2, p. 332–334, 1999.

20 PAVLOPOULOU, A.; MICHALOPOULOS, I. State-of-the-art bioinformatics protein structure prediction tools (review). *International Journal of Molecular Medicine*, v. 28, n. 3, p. 295–310, 2011.

21 LEE, J.; WU, S.; ZHANG, Y. *Ab initio protein structure prediction*. 2009. Disponível em: <http://physweb.ucsc.edu/drupal/sites/default/files/AbInitioStructure.pdf>. Acesso em: 20 dec. 2012.

22 LUTHY, R.; BOWIE, J. U.; EISENBERG, D. Assesment of protein models with three-dimensional profiles. *Nature*, London, v. 356, n. 6364, p. 283–285, 1992.

23 LASKOWSKI, R. A.; MACARTHUR, M. W.; MOSS, D. S.; THORNTON, J. M. PROCHECK - a program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography*, v. 26. pt. 2, p. 283–291, 1993.

24 HOOFT, R. W. W.; VRIEND, G.; SANDER, C.; ABOLA, E. E. Errors in protein structures. *Nature*, London, v. 381, n. 6580, p. 272–272, 1996.

25 COLOVOS, C.; YEATES, T. O. Verification of protein structures: patterns of nonbonded atomic interactions. *Protein Science*, v. 2, n. 9, p. 1511–1519, 1993.

26 PONTIUS, J.; RICHELLE, J.; WODAK, S. J. Deviations from standard atomic volumes as a quality measure for protein crystal structures. *Journal of Molecular Biology*, v. 264, n. 1, p. 121–136, 1996.

27 COLLINS, T.; GERDAY, C.; FELLER, G. Xylanases, xylanase families and extremophilic xylanases. *FEMS Microbiology Reviews*, v. 29, n. 1, p. 3–23, 2005.

28 NOMENCLATURE Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB). Enzyme nomenclature. 1992. Disponível em: <http://www.chem.qmul.ac.uk/iubmb/enzyme/>. Acesso em: 20 dec. 2012.

29 MIZUGUCHI, K.; DEANE, C. M.; BLUNDELL, T. L.; OVERINGTON, J. P. HOMSTRAD: a database of protein structure alignments for homologous families. *Protein Science*, v. 7, n. 11, p. 2469–2471, 1998.

30 DUMON, C. et al. Engineering hyperthermostability into a GH11 xylanase is mediated by subtle changes to protein structure. *The Journal of Biological Chemistry*, v. 283, n. 33, p. 22557–22564, 2008.

31 TOLLIN, P.; ROSSMAN, M. G. A description of various rotation function programs. *Acta Crystallographica*, v. 21. pt. 6, p. 872–876, 1966.

32 FRENKEL, D.; SMIT, B. *Understanding molecular simulation from algorithms to applications*. 2nd ed. New York: Academic Press, 2002.

33 PHILLIPS, J. C.; BRAUN, R.; WANG, W.; GUMBART, J.; TAJKHORSHID, E.; VILLA, E.; CHIPOT, C.; SKEEL, R. D.; KALE, L.; SCHULTEN, K. Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry*, v. 26, n. 16, p. 1781–1802, 2005.

34 BROOKS, B. R. et al. CHARMM: the biomolecular simulation program. *Journal of Computational Chemistry*, v. 30, n. 10, p. 1545–1614, 2009.

35 LIANG, J.; DILL, K. A. Are proteins well-packed? *Biophysical Journal*, v. 81, n. 10, p. 751–766, 2001.

36 ALEXANDROV, V.; LEHNERT, U.; ECHOLS, N.; ENGELMAN, D. M. D.; GERSTEIN, M. Normal modes for predicting protein motions: a comprehensive database assesment and associated Web tool. *Protein Science*, v. 14, n. 3, p. 633–643, 2005.

37 LEITNER, D. Energy flow in proteins. *Annual Review of Physical Chemistry*, v. 59, p. 233–259, 2008. DOI: 10.1146/annurev.physchem.59.032607.093606.

38 MARTíNEZ, L.; FIGUEIRA, A.; WEBB, P.; POLIKARPOV, I.; SKAF, M. Mapping the intramolecular vibrational energy flow in proteins reveals functionally important residues. *The Journal of Physical Chemistry Letters*, v. 2, n. 6, p. 2073–2078, 2011.

39 OTA, N.; AGARD, D. A. Intramolecular signaling pathways revealed by modeling anisotropic thermal diffusion. *Journal of Molecular Biology*, v. 351, n. 2, p. 345–354, 2005.

40 NEWMAN, M. E. J. The structure and function of complex networks. *SIAM Review*, v. 45, n. 2, p. 167–256, 2003.

41 DICIONÁRIO Priberam da língua portuguesa, 2012. Disponível em: <http://www.priberam.pt/dlpo/default.aspx?pal=sistema>. Acesso em: 25 dec. 2012.

42 VAN STEEN, M. *Graph theory and complex networks: an introduction*. Amsterdan: Maarten van Steen, 2010.

43 MORENO, J. L. *Who shall survive?* Beacon: Beacon House, 1934.

44 FREEMAN, L. Centrality in social networks: conceptual clarification. *Social Networks*, v. 1, n. 3, p. 215–239, 1978. DOI:10.1016/0378-8733(78)90021-7.

45 DE FREITAS, L. Q. Medidas de centralidade em grafos. 2010. 111 p. Dissertação (Mestrado em Engenharia de Produção) - Universidade Federal do Rio de Janeiro, 2010. Disponível em: <http://objdig.ufrj.br/60/teses/coppe_m/LeandroQuintanilhaDeFreitas.pdf>. Acesso em: 20 dec. 2012.

46 MILGRAM, S. The small-world problem. *Psychology Today*, v. 2, n. 1, p. 60–67, 1967.

47 STROGATZ, S. H. Exploring Complex Networks. *Nature*, London, v. 410, n. 6825, p. 268–276, 2001.

48 HOLME, P.; KIM, B. J.; YOOM, C. N.; HAN, S. K. Attack vulnerability of complex networks. *Physical Review E*, v. 65, n. 5, p. 0561091–05610914, 2002.

49 GREENE, L.; HIGMAN, V. Uncovering network systems within protein structures. *Journal of Molecular Biology*, v. 334, n. 4, p. 781–791, 2003.

50 VENDRUSCOLO, M.; DOKHOLYAN, N.; PACI, E.; KARPLUS, M. Small-world view of the amino acids that play a key role in protein folding. *Physical Review E*, v. 65, n. 6, p. 061910, 2002.

51 DEL SOL, A.; O'MEARA, P. Small-world network approach to identify key residues in protein-protein interaction. *Proteins:* structure, function, and bioinformatics, v. 58, p. 672–682, 2005.

52 KRISHNAN, A.; ZBILUT, J.; TOMITA, M.; GIULIANI, A. Proteins as networks: usefulness of graph theory in protein science. *Current Protein and Peptide Science*, v. 9, n. 1, p. 28–38, 2008.

53 COSTA, L. F.; RODRIGUES, F. A.; TRAVIESO, G.; VILLAS BOAS, P. R. Characterization of complex networks: a survey of measurements. *Advances in Physics*, London, v. 56, n. 1, p. 167–242, 2007.

54 DOKHOLYAN, N. V.; LI, L.; DING, F.; SHAKHNOVICH, E. I. Topological determinants of protein folding. *Proceedings of the National Academy of Sciences of the United States of America*, Washington, v. 99, n. 13, p. 8637–8641, 2002.

55 VENDRUSCOLO, M.; PACI, E.; DOBSON, C. M.; KARPLUS, M. Three key residues form a critical contact network in a protein folding transition state. *Nature*, London, v. 409, n. 6820, p. 641–645, 2001.

56 AMITAI, G.; SHEMESH, A.; SITBON, E.; SHKLAR, M.; NETANELY, D.; VENGER, I.; PIETROKOVSKI, S. Network analysis of protein structures identifies functional residues. *Journal of Molecular Biology*, v. 344, p. 1135–1146, 2004. DOI:10.1016/j.jmb.2004.10.055.

57 PATRA, S. M.; VISHVESHWARA, S. Backbone cluster identification in proteins by a graph theoretical method. *Biophysical Chemistry*, v. 84, n. 1, p. 13–25, 2000.

58 ARTYMIUK, P. J.; POIRRETTE, A. R.; GRINDLEY, H. M.; RICE, D. W.; WILLETT, P. A graph-theoretic approach to the identification of three-dimensional patterns of amino acid side-chains in protein structures. *Journal of Molecular Biology*, v. 243, n. 2, p. 327–344, 1994.

59 HUAN, J.; BANDYOPADHYAY, D.; WANG, W.; SNOEYINK, J.; PRINS, J.; TROPSHA, A. Comparing graph representations of protein structure for mining family-specific residue-based packing motifs. *Journal of Computational Biology*, v. 12, n. 6, p. 657–671, 2005.

60 N. KANNAN; VISHVESHWARA, S. Identification of side-chain clusters in protein structures by a graph spectral method. *Journal of Molecular Biology*, v. 292, p. 441–464, 1999.

61 BöDE, C.; KOVáCS, I. A.; SZALAYB, M. S.; PALOTAIB, R.; KORCSMáROSB, T.; CSERMELY, P. Network analysis of protein dynamics. *FEBS Letters*, v. 581, n. 15, p. 2776–2782, 2007.

62 CSERMELY, P.; SANDHU, K. S.; HAZAI, E.; HOKSZA, Z.; KISS., H. J. M.; MIOZZO, F.; VERES, D. V.; PIAZZA, F.; NUSSINOV, R. Disordered proteins and network disorder in network descriptions of protein structure, dynamics and function. hypotheses and a comprehensive review. *Current Protein and Peptide Science*, v. 13, n. 1, p. 19–33, 2012.

63 CONTE, L. L.; AILEY, B.; HUBBARD, T. J. P.; BRENNER, S. E.; MURZIN, A. G.; CHOTHIA, C. Scop: a structural classification of proteins database. *Nucleic Acids Research*, v. 28, n. 1, p. 257–259, 2000.

64 BERMAN, H.; WESTBROOK, J.; FENG, Z.; GILLILAND, G.; BHAT, T.; WEISSIG, H.; SHINDYALOV, I.; BOURNE, P. The protein data bank. *Nucleic Acids Research*, v. 28, n. 1, p. 235–242, 2000.

65 CHANDONIA, J.; HON, G.; WALKER, N.; CONTE, L. L.; KOEHL M. LEVITT, P.; BRENNER, S. E. The astral compendium in 2004. *Nucleic Acids Research*, v. 32, p. D189–D192, 2004. Supplement 1.

66 CHANDLER, D. *Introduction to modern statistical mechanics*. Oxford: Oxford University Press, 1987.

67 PYTHON programming language - official website. Disponível em: <http://www.python.org/>. Acesso em: 13 fev. 2013.

68 OLIPHANT, T. E. *Guide to numpy*. 2006. Disponível em: <http://www.tramy.us/numpybook.pdf>. Acesso em: 13 fev. 2013.

69 JONES, E.; OLIPHANT, T.; PETERSON, P. SciPy: open source scientific tools for Python, 2001. Disponível em: <http://www.scipy.org/>. Acesso em: 13 fev. 2013.

70 HUNTER, J. D. Matplotlib: a 2d graphics environment. *Computing In Science & Engineering*, v. 9, n. 3, p. 90–95, 2007.

71 MIYAZAWA, S.; JERNIGAN, R. L. Estimation of effective interresidue contact energies from protein crystal structures quasi-chemical approximation. *Macromolecules*, Washington, v. 18, n. 3, p. 534–552, 1985.

72 MUNIZ, H. Estudo computacional da difusão térmica em proteínas termoestáveis, 2013. 113 p., Dissertação (Mestrado em Física) - Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2013.

73 MAKHATADZE, G. I.; PRIVALOV, P. L. Heat capacity of proteins I. partial molar heat capacity of individual amino acid residues in aqueous solution hydration effect. *Journal of Molecular Biology*, v. 213, n. 2, p. 375–384, 1990.

74 FREUND, J. E. *Modern elementary statistics*. Upper Saddle River: Pearson, 2006.

75 VALENTE, T. W.; CORONGES, K.; LAKON, C.; COSTENBADER, E. How correlated are network centrality measures? *Connect (Tor)*, v. 28, n. 1, p. 16–26, 2008.

76 SABIDUSSI, G. The centrality index of a graph. *Psychometrika*, v. 31, n. 4, p. 581–683, 1966.

APÊNDICE *A*

Demonstrações

A.1 Efeito da Superfície na Contagem de Vizinhos

As equações (4.1.2) e (4.1.5) relacionam o número de vizinhos contado por cada partícula em uma distribuição aleatória com parâmetros da própria distribuição, e são válidas para partículas longe da superfície da esfera que engloba a mesma. Para as partículas próximas da superfície, o número de vizinhos contados necessariamente deve ser menor, pois parte do volume da esfera que cada partícula enxerga está fora da distribuição. Demonstraremos nesta seção que, no regime em que a distância de *cutoff* é pequena em relação ao raio da distribuição, as equações (4.1.2) e (4.1.5) são aproximadamente válidas para todas as partículas.

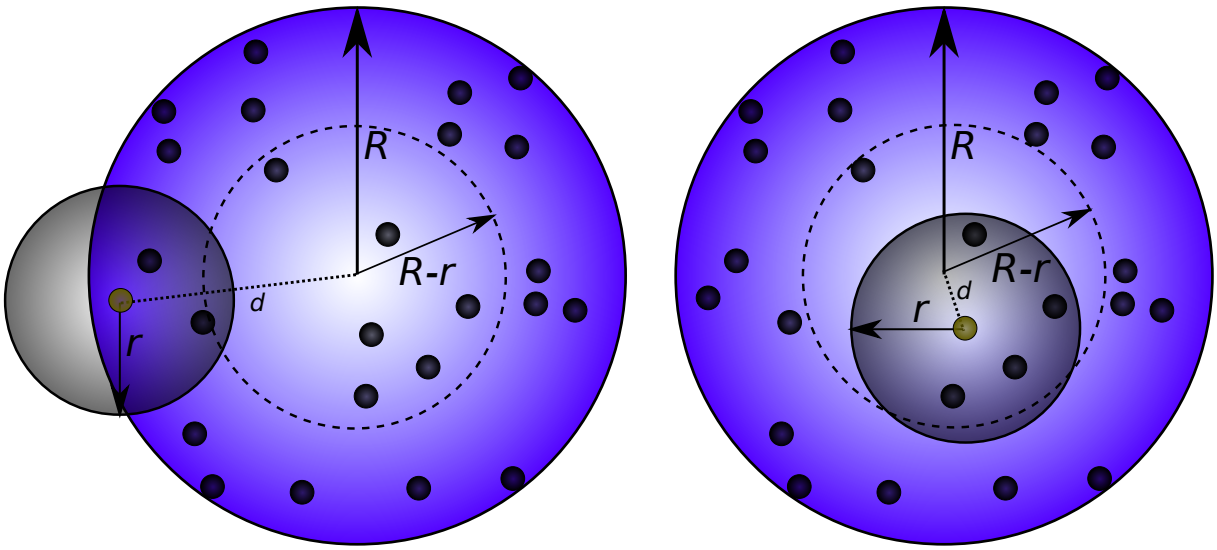

Figura A.1 – *O número de vizinhos contado por uma partícula depende do volume que ela enxerga e da sua distância em relação ao centro da distribuição. Partículas muito próximas da superfície contam menos vizinhos, pois parte dos volumes que elas enxergam não faz parte da distribuição.*

Formalmente, se o raio da distribuição for R e o *cutoff* for r, todas as partículas localizadas a uma distância $d > R - r$ do centro da esfera contarão um número de vizinhos proporcional ao volume efetivo $V_{ef}(R, r, d)$ da intersecção entre a esfera de raio r centrada na partícula e a

esfera de raio R que contém a distribuição. O volume dessa intersecção obedece a uma forma funcional complicada, mas é evidente pela simetria do problema que a *razão* entre $V_{ef}(R,r,d)$ e $V(r)$ (volume da esfera de raio r) se mantém constante quando R, r e d são multiplicados pelo mesmo fator. Usaremos este fato para reduzir o número de variáveis do problema fazendo $V_{ef}(R,r,d) = f(R,r,d)V(r)$, em que f corresponde à fração do volume da esfera que faz parte da intersecção. Por simetria, a dependência de f deve ser da forma $f(\frac{r}{R},\frac{d}{R})$. Vem:

$$V_{ef}(r,\tfrac{r}{R},\tfrac{d}{R}) = \begin{cases} f(\frac{r}{R},\frac{d}{R})V(r), & R-r \leq d \leq R \\ V(r), & 0 \leq d \leq R-r \end{cases} \tag{A.1.1}$$

Conhecido V_{ef}, o número de vizinhos efetivo m_{ef} será dado por:

$$m_{ef}(r,\tfrac{r}{R},\tfrac{d}{R}) = \int n\, dV_{ef}(r,\tfrac{r}{R},\tfrac{d}{R}) = \int_0^r n\frac{dV_{ef}(r,\frac{r}{R},\frac{d}{R})}{dr}\, dr \tag{A.1.2}$$

Onde $n = \frac{N}{V}$. O valor de V_{ef}, dado pela equação (A.1.1), pode ser estimado da seguinte maneira: Considere o número de partículas que se encontra a $0 \leq d \leq R-r$ do centro, para as quais $f = 1$. A relação entre este número e o número total de partículas pode ser interpretada como a *probabilidade* de encontrar uma partícula a uma distância menor ou igual a d do centro da distribuição. Analogamente, a relação entre o número de partículas que se encontra a uma distância $R-r \leq d \leq R$ do centro, para as quais f é uma fração desconhecida, e o volume total, é a probabilidade de encontrar uma partícula a uma distância entre $R-r$ e R do centro. Podemos utilizar essas probabilidades para ponderar o volume efetivo. Como o número de partículas em cada região é proporcional ao volume da mesma, temos, em média:

$$V_{ef}(r,\tfrac{r}{R},\tfrac{d}{R}) = \frac{\frac{4\pi}{3}(R-r)^3}{\frac{4\pi}{3}R^3}\frac{4\pi}{3}r^3 + \frac{[\frac{4\pi}{3}R^3 - \frac{4\pi}{3}(R-r)^3]}{\frac{4\pi}{3}R^3}f(\tfrac{r}{R},\tfrac{d}{R})\frac{4\pi}{3}r^3$$

$$V_{ef}(r,\tfrac{r}{R},\tfrac{d}{R}) = \frac{4\pi}{3}r^3\left\{\frac{(R-r)^3 + f(\frac{r}{R},\frac{d}{R})\,[R^3-(R-r)^3]}{R^3}\right\}$$

$$V_{ef}(r,\tfrac{r}{R},\tfrac{d}{R}) = \frac{4\pi}{3}r^3\left\{\frac{(R-r)^3}{R^3} + f(\tfrac{r}{R},\tfrac{d}{R})\left[1-\frac{(R-r)^3}{R^3}\right]\right\} \tag{A.1.3}$$

A expressão obtida independe de f no limite em que $r \ll R$, como pode ser verificado pela expansão de Taylor de V_{ef} para $\frac{r}{R}$ em torno de 0. Reescrevendo V_{ef} em termos de $\frac{r}{R}$ e derivando, vem:

$$V_{ef}(r, \tfrac{r}{R}, \tfrac{d}{R}) = \frac{4\pi}{3} r^3 \left\{ (1 - \tfrac{r}{R})^3 + f(\tfrac{r}{R}, \tfrac{d}{R}) \left[1 - (1 - \tfrac{r}{R})^3 \right] \right\}$$

$$V_{ef}(r) = \sum_{n=0}^{+\infty} \frac{1}{n!} \frac{\partial^n V_{ef}}{\partial (\frac{r}{R})^n} (\tfrac{r}{R})^n = \frac{4\pi}{3} r^3 + \mathcal{O}(r^4)$$

A primeira e a segunda derivadas de V_{ef} se anulam quando calculadas em $\frac{r}{R} = 0$, e o termo que sobrevive na terceira derivada não depende de f. O resultado é o volume da esfera de raio r, conforme esperado, pois quando r é pequeno a maior parte das partículas fica a uma distância do centro menor que $R - r$. Substituindo esse resultado na equação (A.1.2), vem:

$$m(r) = \int_0^r n \frac{dV_{ef}(r, \frac{r}{R}, \frac{d}{R})}{dr}\, dr = \int_0^r n 4\pi r^2\, dr \tag{A.1.4}$$

Para uma distribuição aleatória, $n = n_{\text{gás}}$ é constante, e a integral pode ser resolvida:

$$m_{\text{gás}}(r) = \int_0^r n_{\text{gás}} 4\pi r^2\, dr = n_{\text{gás}} 4\pi \frac{r^3}{3} \tag{A.1.5}$$

Para uma distribuição estruturada, $n = n_{\text{est}}(r) = n_{\text{gás}} g(r)$, e o resultado é dado na forma de integral:

$$m_{\text{est}}(r) = \int_0^r n_{\text{est}} 4\pi r^2\, dr = n_{\text{gás}} 4\pi \int_0^r g(r) r^2\, dr \tag{A.1.6}$$

Ficando, assim, demonstrada a validade das equações (4.1.2) e (4.1.5) no limite $r \ll R$.

A.2 Limite de $g(r)$

Mostraremos nesta seção que a equação (4.1.6) implica na equação (4.1.7). Considere o seguinte teorema:

Teorema 1 *Sejam $f(x)$ e $g(x)$ funções integráveis definidas no intervalo $[a;+\infty)$. Admita que $f(x) \neq 0$ neste intervalo*. Dadas as hipóteses:*

(I)

$$|\int_a^{+\infty} f(t)\,dt| = \infty$$

(II)

$$\lim_{x\to+\infty} g(x) \quad \textit{existe.}$$

Então:

(III)

$$\lim_{x\to+\infty} \frac{\int_a^x f(t)g(t)\,dt}{\int_a^x f(t)\,dt} = \lim_{x\to+\infty} g(x)$$

Demonstração: De (I), (II) e $f(x) \neq 0$, decorre que o limite em (III) satisfaz as condições da aplicação da regra de l'Hospital. Pela aplicação da mesma, vem:

$$\lim_{x\to+\infty} \frac{\int_a^x f(t)g(t)\,dt}{\int_a^x f(t)\,dt} = \lim_{x\to+\infty} \frac{f(x)g(x)}{f(x)} = \lim_{x\to+\infty} g(x)$$

Fazendo $f(x) = f(r) = \frac{dV(r)}{dr} = 4\pi r^2$ (ou $f(r) = \frac{dV_{ef}(r)}{dr}$ no caso mais geral) e $g(x) = g(r)$ no intervalo $[0,+\infty)$, fica demonstrada a equação (4.1.7). O fato de que $f(0) = 0$ (isto é, somente no extremo inferior do intervalo) pode ser contornado fazendo $a = \frac{1}{x}$ nos limites de integração, com poucas alterações na demonstração.

*A condição $f(x) \neq 0$ é necessária para a aplicação da regra de l'Hospital na demonstração, que requer que a derivada da função no denominador seja não-nula. Em alguns enunciados da regra ela é substituída pela condição de que as funções do numerador e denominador sejam diferenciáveis em todo o intervalo. Ela é incluída aqui por completeza.