Universidade de São Paulo
Instituto de Física

Determinação da Estrutura e do Estado Oligomérico de Proteínas Chaperonas por Espalhamento de Raios-X a Baixos Ângulos (SAXS)

Luiz Fernando de Camargo Rodrigues

Orientador: Prof. Dr. Leandro Ramos Souza Barbosa

Dissertação apresentada ao Instituto de Física da Universidade de São Paulo como requisito parcial para a obtenção do título de Mestre em Ciências.

Banca examinadora:
Prof. Dr. Leandro Ramos Souza Barbosa - Orientador (IF-USP)
Profª. Drª. Patrícia Soares Santiago (UNESP - Registro)
Prof. Dr. Luís Fernando Mercier Franco (FEQ - UNICAMP)

São Paulo
2020

FICHA CATALOGRÁFICA
Preparada pelo Serviço de Biblioteca e Informação do Instituto de Física da Universidade de São Paulo

Rodrigues, Luiz Fernando de Camargo

Determinação da estrutura e do estado oligomérico de proteínas chaperonas por espalhamento de raios-X a baixo ângulo (SAXS). São Paulo, 2020.

Dissertação (Mestrado) – Universidade de São Paulo. Instituto de Física. Depto. de Física Geral

Orientador: Prof. Dr. Leandro Ramos Souza Barbosa

Área de Concentração: Biofísica – Biofísica Molecular

Unitermos: 1. Biofísica molecular; 2. SAXS; 3. Proteínas; 4. Bioquímica.

USP/IF/SBI-026/2020

University of São Paulo
Institute of Physics

Determination of Structure and Oligomeric State of Chaperone Proteins by Small-Angle X-Ray Scattering (SAXS)

Luiz Fernando de Camargo Rodrigues

Supervisor: Prof. Dr. Leandro Ramos Souza Barbosa

Dissertation submitted to the Institute of Physics of the University of São Paulo as partial fulfillment of the requirements for the degree of Master of Science.

Examining committee:
Prof. Dr. Leandro Ramos Souza Barbosa - Orientador (IF-USP)
Prof. Dr. Patrícia Soares Santiago (UNESP - Registro)
Prof. Dr. Luís Fernando Mercier Franco (FEQ - UNICAMP)

São Paulo
2020

Em memória de José Lourenço Pires

Agradecimentos

À minha mãe Maria Angelica por tudo. Absolutamente **tudo**.

À minha irmã Maria Rita por ser a pessoa mais companheira que eu poderia pedir.

Ao meu tio José Lourenço por todo o apoio e incentivo a ser quem eu sou e me ensinar a sonhar com galáxias muito, muito distantes. Eu vou assistir todos os filmes do Hitchcock e finalmente entender a diferença entre Cary Grant e James Stewart. Boldly go, tio Joe. Descanse em paz, e que a força esteja com você.

Ao meu pai Nelson pelo carinho e me ajudar por tantas encruzilhadas na vida.

À minha família pelo amor e apoio em todos os momentos no decorrer da minha vida. Especialmente à vovó Célia pelos bolos de cenoura com cobertura de chocolate.

À minha companheira Isabela por ser minha luz em todos os momentos, quem me ensinou a falar "impedimento estérico" e "eu te amo". Muito obrigado pelo amor e paciência. Te amo, girassol.

Ao meu orientador Leandro Ramos Souza Barbosa pela amizade, por me mostrar que a pesquisa em física vai muito além do óbvio e por me guiar pelo mundo acadêmico.

Às amigatinhas Paçoca, Capitu e Leucina pela companhia e carinho em forma de cabeçadinhas, miados e ronrons.

Ao Marcos Cardoso, por me mostrar o significado de compromisso e seriedade no que se faz. Também ao Marcio Bernardino, pela amizade e por todos os momentos tão queridos no preparo físico.

Aos amigos do Instituto de Física, em especial Lukinha$ e Caíke por todos esses anos de amizade e trabalho duro aprendendo teoria de perturbação e tendo discussões acaloradas sobre Turma da Mônica e Star Trek.

Ao João por tantas discussões e recomendações na hora do almoço. Ao Dani pelo

companheirismo e amizade tão intensa. À Laís pelos lanches e conversas noite adentro sobre música. À Claudia pela paciência e compreensão.

Aos amigos de colégio: Felipe, Léo, Pedro, Janini, Xará e Mussa. Obrigado por todos os nossos momentos de descontração.

Aos amigos de treino no Esporte Clube Pinheiros, em especial Nicolas, Julio, Richard e Heitor, por toda a convivência e amizade por horas e mais horas em todos esses anos.

Aos amigos de laboratório: Marcelo, Fred, Natu, Juliana, Raphael, Bruna, Letícia, Bárbara e Mayra. Por toda a paciência nos desabafos e ajuda, meu muito obrigado.

Aos colaboradores e amigos de São Carlos: Prof. Júlio, Noeli, Silvia, Amanda e Vanessa. Aprendo e me divirto muito quando encontro com vocês e espero vê-los novamente em breve. Que eu não cheire nenhum erlenmeyer com cultura de células desta vez. Também aos colaboradores e amigos de Campinas: Prof. Carlos, Glaucia e Natália.

À Universidade de São Paulo, meu lugar favorito do planeta, por toda a estrutura que me acolheu desde os tempos de graduação e me ensinou a me sentir em casa comigo mesmo do jeito que sou.

Aos amigos da Associação Atlética Acadêmica Gleb Wataghin (AAAGW), por todo o companheirismo dentro e fora de quadra. Especialmente ao Grassetti e aos Glebtrotters.

Aos amigos do Centro de Estudos de Física e Matemática (CEFISMA) por todo o trabalho e luta em defesa dos estudantes.

Aos amigos do Instituto de Psicologia por todo o acolhimento desde 2014. Não imaginava que uma edição do BIFE me traria tantos amigos queridos.

A todos os trabalhadores que fazem a nossa sociedade funcionar, desde o professor que leciona Mecânica Estatística no IF à dona Maria que varre os ônibus em que andamos.

À CAPES pela minha bolsa, ao CNPq pela verba e aos projetos temáticos: processos nº 2015/15822-1, 2012/01953-9, 2016/05019-0, Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP). Também ao LNLS pelo uso das instalações em que coletamos os dados de SAXS. Este trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

A todos os pesquisadores que vieram antes de mim em cujas publicações e trabalhos encontrei uma área de pesquisa fascinante e envolvente.

Sem essas pessoas e entidades não teria como chegar onde cheguei. Faço minhas as palavras de Lennon e McCartney: *I get by with a little help from my friends.*

All things are a part

All things are apart

All things are a part

Peter Hammill

Resumo

Espalhamento de raios-X a baixos ângulos (SAXS) é uma técnica estrutural em estudo de materiais a baixa resolução com um amplo leque de aplicações. Nas últimas décadas, houve progresso em métodos de análise de dados de SAXS que aumentaram a utilização dessa ferramenta, especialmente em biologia estrutural para o estudo de proteínas. As chaperonas moleculares constituem um conjunto diversificado de proteínas que tem como função biológica a prevenção e correção de problemas de enovelamento proteico e agregação, responsáveis por diversas patologias. Neste trabalho, foram aplicadas metodologias de análise de SAXS para o estudo de proteínas chaperonas das famílias *70-kDa* e *90-kDa heat shock proteins* (Hsp70 e Hsp90). Foram comparados membros de Hsp70 humana de diferentes compartimentos celulares e mostrou-se que não há variação apreciável observada por SAXS entre essas proteínas. Estas se encontram em conformação levemente elongada com baixa flexibilidade e em estado monomérico em solução. A Hsp70 de retículo endoplasmático (*Binding immunoglobulin protein*, Bip) foi utilizada para modelagens *ab initio*, de *ensemble* e de corpo rígido, com obtenção da orientação entre os domínios e do envelope da proteína a baixa resolução. Foi feito o estudo comparativo de uma *Gro-P like Protein E* de mitocôndria humana (GrpE-L1) em diferentes solventes: em tampão e em agentes oxidante e redutor. Obteve-se o envelope da proteína e mostrou-se que não há compactação sensível da proteína em agente redutor, indicando que pontes dissulfeto não são fatores únicos para a manutenção da forma altamente elongada observada para a GrpE-L1, encontrada como dímero em solução. Também foi estudada a Hsp90 de *Aedes aegypti* (AaHsp90) em presença de diferentes nucleotídeos de adenosina. Foi mostrado que a AaHsp90 se encontra como um dímero em solução com conformação bastante elongada e pouca flexibilidade, sem variação conformacional apreciável ao ligar-se a diferentes nu-

cleotídeos. Foi obtida a estrutura da AaHsp90 em solução a partir de modelagens de corpo rígido e *ab initio*. A baixa plasticidade estrutural observada pode indicar que em Hsp90s de *Aedes aegypti* há maior dependência de co-chaperonas para que a função da proteína seja desempenhada neste organismo ou que interações com proteínas-cliente modulem sua ação, estando de acordo com estudos anteriores com membros de Hsp90 humana.

Palavras-chave: SAXS, enovelamento de proteínas, chaperonas moleculares, Hsp70, Hsp90

Abstract

Small-Angle X-Ray Scattering (SAXS) is a structural technique in materials' studies at low resolution with a wide range of applications. In last decades, there has been progress in SAXS data analysis methods which increased the use of this tool, especially in structural biology for protein studies. Molecular chaperones constitute a diverse set of proteins that have the biological function of preventing and correcting protein folding and agregation problems, responsible for various pathologies. In the present work, SAXS analysis tools have been applied to the study of 70-kDa and 90-kDa heat shock protein (Hsp70 and Hsp90) families of chaperones. Human Hsp70s of different cell compartments were compared and it was shown that there is no appreciable variation for these proteins seen by SAXS. They are found in a slightly elongated conformation in solution with low flexibility and in monomeric state in solution, with deviations from previous SAXS studies on Hsp70 members. Endoplasmic reticulum Hsp70 (Bip) was used for *ab initio*, ensemble and rigid-body techniques, obtaining the protein envelope and domain orientations at low resolution. A comparative study of a human mitochondrial Gro-P like protein E (GrpE-L1) was performed in different solvents: in buffer and in oxidizing and reducing agents. The protein envelope was obtained and it was shown that there is no sensible protein compactness in reducing agent, indicating that disulfide bonds are not the only factors for the maintenance of GrpE-L1's highly elongated shape, found as a dimer in solution. Hsp90 from *Aedes aegypti* (AaHsp90) was also studied in presence of different adenosine nucleotides. It was shown that AaHsp90 is found as a dimer in solution with highly elongated shape and low flexibility, with no appreciable conformational variation when bound to different nucleotides. AaHsp90 solution structure was obtained by rigid-body and *ab initio* methods. The observed low structural plasticity indicated that in Hsp90s from *Aedes*

aegypti there is more dependence on co-chaperones for protein function or that interactions with client-proteins modulate its action, in accordance to previous studies with human Hsp90 members.

Keywords: SAXS, protein folding, molecular chaperones, Hsp70, Hsp90

Lista de Figuras

Lista de Tabelas

Lista de Abreviações

AMP: Adenosina Monofosfato;

ADP: Adenosina Difosfato;

ATP: Adenosina Trifosfato;

Bip: *Binding immunoglobulin protein*;

CTD: Domínio C-terminal (*C-terminal Domain*);

CorMap: Mapa de Correlação (*Correlation Map*);

DENSS: Densidade de Espalhamento em Solução (*DENsity from Solution Scattering*);

EOM: Método de Otimização de *Ensemble* (*Ensemble Optimization Method*);

ER: Razão de Elongamento (*Elongation Ratio*);

GAJOE: Algoritmo Genético para Julgamento de Otimização de *Ensembles* (*Genetic Algorithm Judging Optimisation of Ensembles*);

GPA: Análise de Pico de Guinier (*Guinier Peak Analysis*);

GrpE: *Gro-P like Protein E*;

Hsc70: *70-kDa heat shock cognate protein*;

Hsp: Proteína de Choque Térmico (*Heat Shock Protein*);

Hsp70: Proteína de Choque Térmico de 70-kDa (*70-kDa Heat Shock Protein*);

Hsp90: Proteína de Choque Térmico de 90-kDa (*90-kDa Heat Shock Protein*);

IDP: Proteína Instrinsecamente Desordenada (*Instrinsically Disordered Protein*);

IFT: Transformada Inversa de Fourier (*Inverse Fourier Transform*);

LNLS: Laboratório Nacional de Luz Síncrotron;

MD: Domínio Médio (*Middle Domain*);

NBD: Domínio de Ligação a Nucleotídeo (*Nucleotide Binding Domain*);

NSD: Discrepância Espacial Normalizada (*Normalized Spatial Discrepancy*);

NTD: Domínio N-terminal (*N-terminal Domain*);

PDB: Banco de Dados de Proteínas (*Protein Data Bank*);

PQC: Controle de Qualidade Proteica (*Protein Quality Control*);

RANCH: Cadeia Aleatória (*RANdom CHain*);

RMN: Ressonância Magnética Nuclear;

SASBDB: Banco de Dados Biológico de Espalhamento a Baixos Ângulos (*Small Angle Scattering Biological Data Bank*);

SBD: Domínio de Ligação a Substrato (*Substrate Binding Domain*);

SAXS: Espalhamento de Raios-X a Baixos Ângulos (*Small-Angle X-Ray Scattering*).

Sumário

Capítulo 1

Introdução

1.1 Proteínas: estrutura e função

Proteínas são moléculas que ocorrem em grande variedade na natureza e com diversas funções, estando presentes em praticamente todos os processos biológicos e compartimentos celulares. Estruturalmente as proteínas são polímeros compostos por aminoácidos como subunidades monoméricas ligadas por meio de ligações covalentes chamadas *ligações peptídicas* (Nelson et al., 2008). Aminoácidos são compostos por um carbono-α ligado a um grupo amino (NH_2), um grupo carboxila (COOH), um átomo de hidrogênio e um grupo radical (R), que é particular a cada aminoácido. Como ilustrado na Figura 1.1, o carbono central é quiral, exceto para o caso da glicina, cujo grupo R é um átomo de hidrogênio. As ligações peptídicas ocorrem entre os grupos amino e carboxila, havendo liberação de uma molécula de água. Ao extremo da proteína que se encontra com o grupo amino sem ligação se dá o nome N-terminal, enquanto ao grupo carboxila no outro extremo da sequência é identificado como C-terminal. Cada aminoácido tem características determinadas a partir do seu grupo R, que podem ser ácidos, básicos, neutros, polares, apolares ou podem conter grupos aromáticos, grupos amida ou mesmo enxofre em sua composição, e esses traços servem para classificá-los, havendo 20 aminoácidos comuns encontrados na natureza, com outros adicionais que podem ser ou de difícil ocorrência ou sintetizados.

Uma vez que proteínas são polímeros compostos por subunidades muito diversas entre si e com diferentes composições químicas, cada subunidade irá interagir de maneiras próprias com o ambiente próximo, seja ele contendo moléculas do solvente ou até mesmo outras regiões da própria proteína, com interações podendo ser de caráter eletrostático, hidrofóbico ou mesmo estérico. Essa rede de interações faz com que a cadeia polipeptídica tenha um

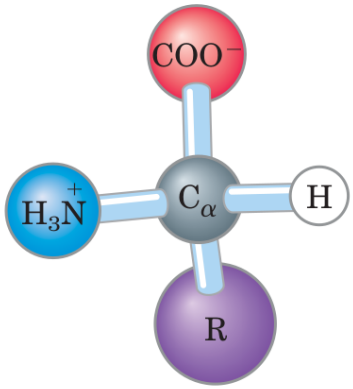

Figura 1.1: Estrutura geral de um aminoácido (exceto prolina), em que o grupo radical (R) é único para cada aminoácido (Nelson et al., 2008).

conjunto de configurações espaciais termodinamicamente favoráveis características, e a esse fenômeno de arranjo tridimensional da proteína se dá o nome de *enovelamento proteico* (*protein folding*) (Anfinsen, 1973). Também pode ocorrer de mais de uma cadeia interagir com outra de modo a formar complexos de proteínas de mesmo tipo ou de tipos diferentes, sendo este processo chamado *multimerização* (ou, quando se trata de poucas proteínas, *oligomerização*). Essas constatações levaram à classificação da estrutura de proteínas em quatro diferentes níveis, ilustrados na Figura 1.2, sendo eles:

- Estrutura primária: sequência de aminoácidos na cadeia polipeptídica, ordenadas do N ao C-terminal.

- Estrutura secundária: arranjos estáveis de regiões da cadeia que resultam em padrões estruturais que conferem uma rigidez às regiões envolvidas nessas interações. Há duas estruturas principais. A primeira é a α-hélice, que consiste em uma sequência periódica de ligações de H entre aminoácidos próximos na cadeia, se repetindo a cada três ou quatro aminoácidos, formando uma espiral. A segunda é a folha-β, formada também por ligações de H porém entre aminoácidos mais distantes, de modo que a forma característica é um zigue-zague com as conjuntos de aminoácidos envolvidos nas ligações formando dois segmentos paralelos ou anti-paralelos.

  Proteínas com alta quantidade de α-hélice e folha-*beta* são ditas bem-enoveladas. Há proteínas que apresentam menor ocorrência dessas estruturas no estado nativo e significativas regiões de ausência de estrutura secundária (chamada de *random-coil*), o que lhe confere maior flexibilidade conformacional. Estas são chamadas

proteínas intrinsecamente desordenadas (ou apenas IDPs, de *Instrinsically Disordered Proteins*) e estão envolvidas principalmente em processos de catálise (Uversky, 2014).

- Estrutura terciária: configuração espacial da cadeia de aminoácidos. Para a classificação dessas conformações são levadas em conta partes independentemente estáveis, chamadas *domínios* (geralmente consistindo de 50 a 300 aminoácidos) (Kolodny et al, 2013), e *motivos*, compostos por padrões espaciais de estrutura secundária, como *coiled-coil* (α-hélices paralelas) ou barril-β (conjunto de várias folhas-β em forma de tubo).

- Estrutura quaternária: quando trata-se de uma proteína multimérica, descreve o arranjo entre os diferentes protômeros e a formação de complexos com diferentes ligantes.

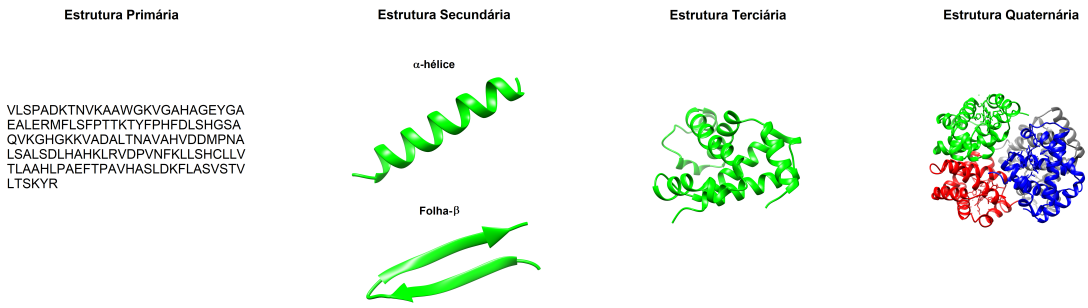

Figura 1.2: Níveis de organização estrutural de uma proteína (usando como exemplo a estrutura de desoxihemoglobina humana, entrada 1A3N do PDB) (Tame e Vallone, 2000). A estrutura primária se refere à sequência de aminoácidos, a secundária a estruturas periódicas, como α-hélice e folha-β, a terciária à conformação tridimensional da cadeia de aminoácidos e a quaternária à disposição entre as cadeias.

Recentemente também surgiu o conceito de *estrutura quinária*, que diz respeito a interações proteicas mais fracas e transientes que as encontradas na estrutura quaternária (Guin e Gruebele, 2019), como as várias interações do ribossomo com outras proteínas da célula durante a síntese de novos polipeptídeos ou mesmo o efeito de aglomeração (do inglês *crowding*) entre componentes da célula. Isso faz com que estudos *in vitro* tenham um escopo limitado sobre este grau estrutural, ilustrando a complexidade de sistemas biológicos a partir do nível celular e reforçando a necessidade de estudos *in vivo*. Assim, também, o paradigma da biologia estrutural de estrutura implicar na função apresenta mais sutilidades, especialmente frente avanços recentes em estudos de proteínas flexíveis,

com importante papel de interações da proteína com o ambiente e outras moléculas no desempenho de sua função.

Há iniciativas para a descrição da estrutura a partir de diferentes critérios e de motivos estruturais, como a *CATH Protein Structure Classification database* (Orengo et al., 1997) e a *Structural Classification of Proteins database* (SCOP) (Murzin et al., 1995). A importância dessas bases de dados se baseia na informação de que funções estão geralmente associadas à estrutura de um domínio ou à ação conjunta de mais de um, e no fato de que, em geral, similaridades de sequência primária entre diferentes proteínas leva a estruturas similares. Assim, utilizando bases de dados de classificação é possível buscar proteínas estruturalmente similares que, portanto, possam desempenhar funções similares.

O estudo da estrutura de proteínas, portanto, é de grande importância para o entendimento do funcionamento das mesmas e, portanto, de um amplo leque de processos biológicos. O estudo a fundo destes processos leva muitas vezes à constatação de mudanças estruturais nas proteínas envolvidas, de modo que fenômenos biológicos são altamente dinâmicos, podendo envolver desde mudanças conformacionais sutis a alterações no estado oligomérico proteico, e esta dinâmica está quase sempre intimamente ligada à função de uma proteína em um processo biológico. Alterações estruturais podem ocorrer por fatores que vão desde o processo de protonação e desprotonação de resíduos ou até mesmo de mutações de aminoácidos específicos à presença de ligantes ou alterações no meio. Mudanças que levam proteínas a adotar estruturas distintas das nativas podem acarretar em perda de função, o que é muitas vezes ligado a patologias que podem, em última instância, levar a fenômenos como apoptose ou morte celular. Assim, métodos experimentais que enderecem questões ligadas à estrutura e função de proteínas são ferramentas muito valiosas em estudos ligados à vida.

1.2 Controle de Qualidade Proteica: proteínas chaperonas como mecanismo de defesa da célula

As células dependem de proteínas em conformações nativas para que desempenhem as suas atividades com normalidade. Assim, para evitar patologias, a célula dispõe de mecanismos para manter suas proteínas enoveladas corretamente, em especial em situações de estresse. Estes mecanismos compõem o chamado *Controle de Qualidade Proteica* (de

Protein Quality Control, ou PQC).

Um dos principais é um conjunto de proteínas chamadas *chaperonas* que têm como função a prevenção de problemas de enovelamento não-nativo e formação de agregados. Uma vez que muitas delas são expressas especialmente sob estresse térmico, também trata-se de chamadas proteínas de choque térmico (do inglês *Heat Shock Proteins*, ou Hsps) (Hartl et al., 2011). Importante ressaltar que nem todas as chaperonas são Hsps, e nem todas as Hsps são chaperonas, porém chaperonas são nomeadas como Hsps e enumeradas de acordo com a massa molecular de cada protômero, sendo divididas em famílias de acordo com sua função e massa do protômero (como Hsp90, Hsp70 etc). Para que desempenhem suas funções diversas chaperonas dependem de outras moléculas, desde nucleotídeos como ATP a proteínas co-chaperonas.

Uma das questões discutidas em biologia estrutural é como as proteínas alcançam suas conformações nativas uma vez que buscas aleatórias não permitiriam alcançar esses estados em tempo biologicamente relevante, como aponta o *paradoxo de Levinthal* (Dill, 1999). Há a proposição de que proteínas buscam sequências de alterações conformacionais até chegar ao estado nativo, e há a de que essas se encontram em um número vasto de conformações (também chamado de *ensemble*), e que, em um perfil de energia livre em forma de funil, a proteína se desloca em mínimos locais que restringem o espaço de busca e guiam o processo de enovelamento (podendo ir a estruturas nativas ou a agregados) (Dill, 1999).

A Figura 1.3 exemplifica a visualização de funil para o enovelamento proteico: cada mínimo local de energia indica uma conformação possível para um polipeptídeo, de modo que configurações intermediárias são ilustradas por mínimos menos profundos enquanto estados nativos e agregados de difícil solubilização são mais bem definidos. Condições de estresse térmico aumentam a energia de um sistema, permitindo que mais estados sejam acessíveis, inclusive estados pouco enovelados, e chaperonas são responsáveis por guiar os polipeptídeos para estados nativos.

Famílias de chaperonas apresentam diferentes funções no PQC, podendo ter mais de uma e atuar de forma conjunta com outras famílias. De forma simplificada pode-se dividir as funções chaperona em (Tiroli-Cepeda e Ramos, 2011):

- Holdase: estabilização de cadeias em estado de desenovelamento ou agregação e transporte;

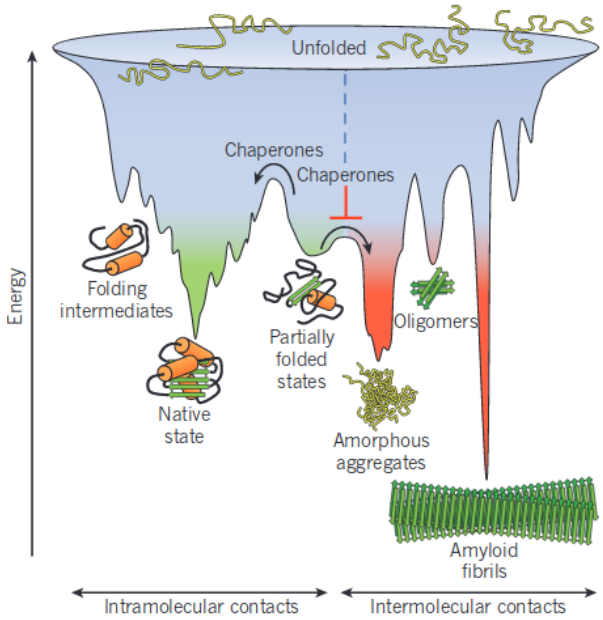

Figura 1.3: Descrição do enovelamento proteico como uma superfície de energia livre, onde a proteína busca a conformação de menor energia. À esquerda há maior proeminência das interações dentro da própria molécula que levam a conformações nativas ou parcialmente enoveladas, enquanto à direita predominam interações entre moléculas, o que pode levar a estados oligoméricos nativos ou a agregados potencialmente patogênicos (Hartl et al., 2011).

- Foldase: auxílio no reenovelamento de proteínas, sejam nascentes do ribossomo quanto livres;
- Disagregase: solubilização de agregados.

Essas funções estão ilustradas na Figura 1.4. Junto dessas funções a célula também dispõe do *complexo ubiquitina-proteassomo*, que é o elemento do PQC responsável pela degradação de polipeptídeos em conformações não-nativas de volta para aminoácidos para que possam sintetizar novas proteínas.

1.2.1 70kDa Heat Shock Proteins (Hsp70)

A família das Hsp70s é altamente versátil promíscua e conservada, estando presente em todos os compartimentos celulares assim como no citosol, sendo um pivô de todo o sistema de chaperonas das célula (Mayer, 2013). A atuação das Hsp70s depende da hidrólise de ATP e está especialmente associada à função foldase, porém há indícios também de atuação em agregados. Essas funções se traduzem por meio da associação a trechos hidrofóbicos flanqueados por resíduos positivamente carregados de proteínas-cliente em

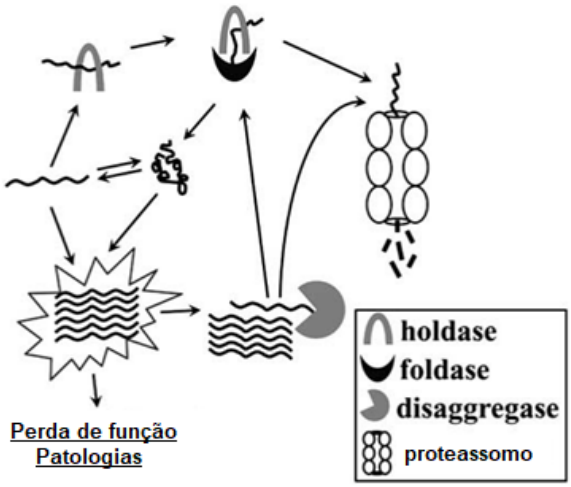

Figura 1.4: Esquematização do controle da proteostase por chaperonas. Proteínas mal enoveladas, seja por terem sido removidas de agregados por disagregases ou por outros caminhos quando livres em solução, são estabilizadas por holdases, que por sua vez as levam a foldases, que auxiliam a proteína-cliente a adotar uma conformação nativa. Polipeptídeos em conformações não-nativas também podem ser degradadas pelo complexo proteassomo (adaptado de Tiroli-Cepeda e Ramos, 2011).

estágios de tradução, translocação e desenovelamento (Mayer e Gierasch, 2019). O fato dessa região de associação ser comum em muitas proteínas e frequente no decorrer de cada sequência faz com que as Hsp70s possam interagir com um amplo leque de proteínas-cliente.

A estrutura das Hsp70s é composta por dois domínios: o domínio de ligação a nucleotídeos (de *nucleotide binding domain*, ou apenas NBD) e o domínio de ligação a substrato (de *substrate binding domain*, ou apenas SBD). Cada domínio é subdividido em dois: o NBD é mais compacto e dividido em dois lóbulos (I e II), enquanto o SBD tem forma de pinça, sendo dinâmico, e é composto pelos subdomínios SBDα (rico em α-hélices) e SBDβ (com alto conteúdo de folhas-β). Acredita-se que a região entre ambos os domínios seja composta por um *linker* (conexão) flexível (Dores-Silva et al., 2015; Mayer e Gierasch, 2019). A estrutura das Hsp70 é altamente dinâmica e duas conformações principais foram observadas para a DnaK (Hsp70 de *Escherichia coli*), sendo uma mais aberta quando ligada a ATP, e outra, mais fechada e flexível, quando na presença de ADP ligado ou sem nucleotídeos. É dito que a conformação com ADP é fechada por haver uma maior proximidade entre os subdomínios do SBD, enquanto com ATP esses se encontram mais afastados. Acredita-se que cada conformação apresenta diferentes afinidades por substrato devido a abertura e fechamento da "pinça" do SBD (Mayer e Gierasch, 2019). Uma esquematização da estrutura e da dinâmica das Hsp70 está presente na Figura 1.5.

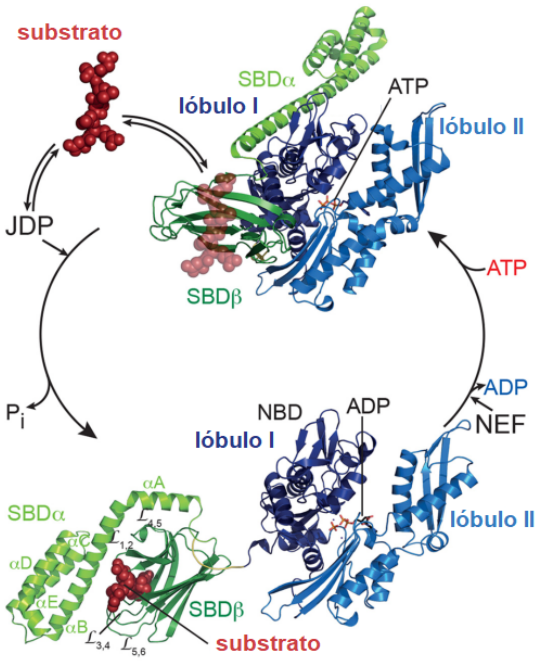

Figura 1.5: Modelo de estrutura e dinâmica de membros da família das Hsp70. As Hsp70 são compostas pelos domínios NBD (com dois lóbulos: I e II, cuja orientação determina a afinidade por ADP ou ATP) e SBD (com os subdomínios SBDα e SBDβ, que se reorientam para regular afinidade por substrato). Quando ligada a ATP o modelo sugere que a conformação é mais aberta, enquanto com ADP ligado a conformação é mais fechada e flexível, com domínios mais independentes (adaptado de Mayer e Gierasch, 2019).

Diversas funções das Hsp70s dependem de outras moléculas para que sejam realizadas, entre elas há outras chaperonas, como as Hsp90 (Luengo et al., 2018), fatores de troca de nucleotídeo (de *nucleotide exchange factor*, ou NEF), como as GrpE (Bracher e Verghese, 2015), e co-chaperonas, como a Hop (de *Hsp70/Hsp90 organizing protein*), sendo esta responsável pela mediação da interação de membros da família Hsp70 com da família Hsp90 (Johnson et al., 1998).

Há diferentes membros das Hsp70s em diferentes compartimentos celulares com sutis diferenças de sequência primária entre si. Essas diferenças estão relacionadas a distintos graus de expressão, especificidades por proteínas-cliente e localização, com alguns membros sendo muito pouco conhecidos (Daugaard et al., 2007). A Figura 1.6 mostra diferentes membros da família Hsp70 em humanos, indicando para quatro (do total de oito) membros as características relativas à estrutura primária para cada um.

Apesar disso, estudos recentes indicam que ainda há muito a ser estudado em relação às Hsp70, desde os mecanismos sugeridos para reenovelamento de proteínas-cliente (*entropic pulling*) (Sousa e Lafer, 2019; Kellner et al., 2014) ao papel das Hsp90 quando em

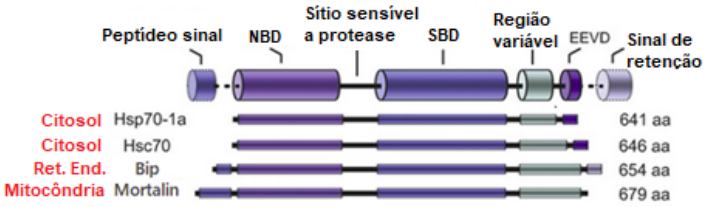

Figura 1.6: Esquematização das diferenças de sequência entre quatro membros da família das Hsp70 humanas, indicando também o principal compartimento celular de cada membro. Cada membro tem regiões específicas, como peptídeo-sinal ou peptídeo de retenção, porém o que é comum a todos é a presença de domínios de ligação a ATP e de ligação a peptídeos e uma região variável no C-terminal (adaptado de Daugaard et al., 2007).

concentrações fisiológicas (Luengo et al., 2018).

1.2.2 GrpE: peça fundamental para a família Hsp70

A liberação de ADP das Hsp70s ocorre com o auxílio de um NEF, a GrpE, sendo esta portanto uma parte importante do PQC em mitocôndria e cloroplastos em eucariotos (Harrison, 2003). Ela se encontra em estado dimérico e em humanos há dois homólogos em mitocôndria: GrpE-L1 e GrpE-L2. Pouco se sabe estruturalmente sobre elas.

As informações estruturais até o presente são duas estruturas cristalográficas, sendo uma de *Escherichia coli* em conjunto com o NBD da DnaK (Harrison et al., 1997) e uma da bactéria extremófila *Thermus thermophilus* HB8 (Nakamura et al., 2010), esta última presente na Figura 1.7. Também há estudo de SAXS que verifica a ocorrência em solução das estruturas cristalográficas propostas a baixa resolução (Borges et al., 2003).

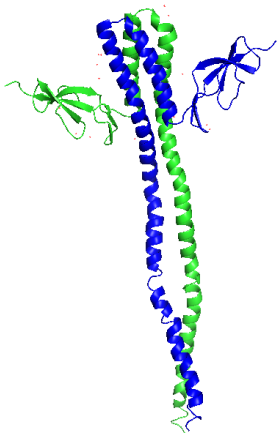

Figura 1.7: Estrutura cristalográfica da GrpE de *Thermus thermophilus* HB8, com cada protômero pintado de uma cor (entrada 3A6M do PDB, publicada por Nakamura et al., 2010).

Há questões ainda não elucidadas que vão desde as diferenças estruturais entre ambas

as GrpE humana quanto à plasticidade da região de *coiled coil* (α-hélices entrelaçadas) que podem ajudar a compreender melhor o funcionamento desse NEF e, portanto, a liberação de ADP de membros da família Hsp70.

1.2.3 90kDa Heat Shock Proteins (Hsp90)

As Hsp90 compõem uma importante família de chaperonas dependentes de ATP que atua principalmente em proteínas em estágios finais de enovelamento, compondo mais um nível de regulação para a proteostase celular. Juntamente com as Hsp70, essa família compõe um papel central de foldase, porém desempenhando funções mais específicas e para um leque menor de substratos, havendo inclusive atuação conjunta de ambas as famílias e co-chaperonas. As Hsp70 realizam um primeiro passo de reenovelamento da proteína-cliente e a entrega às Hsp90 para a etapa final do processo. Uma vez que entre as centenas de clientes de Hsp90 há elementos de processos centrais de crescimento e proliferação celular, essa família de chaperonas se apresenta como um alvo potencial para terapias de câncer (Biebl e Buchner, 2019).

A família das Hsp90 é abundante e altamente conservada, com alta similaridade mesmo entre membros de *E. coli* e de humanos. Porém, em procariotos ocorre apenas uma isoforma em citoplasma, enquanto em eucariotos pode ocorrer mais uma, sendo uma expressa em condições fisiológicas e a outra mais expressa em condições de estresse (por exemplo, as Hsp90α e Hsp90β no caso humano), podendo haver também homólogos para outras organelas (como a TRAP1, de mitocôndria) (Biebl e Buchner, 2019).

Estruturalmente, dimerização é uma característica vital para a função das Hsp90 *in vivo*, sendo elas compostas por três domínios: um N-terminal (NTD) encarregado pela ligação de ATP contendo uma tampa (do inglês *lid*), conectado por um *linker* flexível, de tamanho variável, a um domínio M (MD) intermediário, que tem como função de ligação a clientes e co-chaperonas e participação na hidrólise de ATP, e um domínio C-terminal (CTD) onde ocorre a dimerização. Também há membros da família com um motivo MEEVD no C-terminal que interage com co-chaperonas que contenham um domínio TPR (*tetratricopeptide repeat*) (Schopf et al., 2017). Uma esquematização estrutural das Hsp90 está presente na Figura 1.8.

O atual modelo para o ciclo de funcionamento das Hsp90 consiste nos seguintes passos (Figura 1.9): i) chaperona em conformação aberta quando livre em solução; ii) ligação de

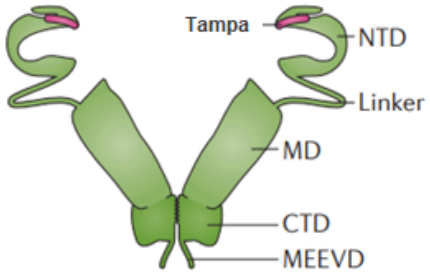

Figura 1.8: Organização dos domínios de membros da família Hsp90 de chaperonas, sendo compostos por um domínio N-terminal (do inglês *N-terminal Domain*, NTD) contendo uma tampa (*lid*), ligado por um *linker* a um domínio médio (do inglês *Middle Domain*, MD) e um domínio C-terminal (do inglês *C-terminal Domain*, CTD), com este último podendo ou não ser seguido por um motivo MEEVD. Para a realização de suas funções as Hsp90 se encontram em estado dimérico em solução (adaptado de Schopf et al., 2017).

ATP ao NTD; iii) fechamento da tampa do NTD, confinando o ATP; iv) dimerização do NTD, com fechamento parcial da estrutura; v) interação do NTD com o MD, permitindo a hidrólise de ATP e um maior fechamento da estrutura; vi) abertura da tampa, liberação de ADP e fosfato e abertura da estrutura, voltando ao estágio inicial. O processo ao todo, visualmente, assemelha-se a uma tesoura. As proteínas-cliente se associam ao domínio M com o fechamento da estrutura, e há a atuação de co-chaperonas tanto em etapas específicas (como a p23, que atua na estrutura mais fechada) quanto no processo como um todo (como ocorre com a PPIase) (Schopf et al., 2017).

Apesar de o modelo descrito ser bastante útil para descrição do fenômeno, há estudos que indicam que o processo parece ser mais complexo. Parece haver indícios de equilíbrio entre conformações abertas e fechadas a partir de constatações de flexibilidade (Seraphim et al., 2017) e estudos de microscopia eletrônica que indicam que o equilíbrio entre estados aparenta variar conforme o organismo, o que sugere maior dependência de co-chaperonas para o desempenho de função para determinadas espécies (Southworth e Agard, 2008).

1.3 Principais técnicas em biologia estrutural e a necessidade de abordagens híbridas

Em biologia estrutural, pode-se obter informações sobre uma proteína utilizando métodos de alta e baixa resolução. Em alta resolução, são obtidos mapas atômicos fazendo uso de cristalografia ou de ressonância magnética nuclear (RMN), enquanto em baixa resolução são obtidos parâmetros como determinadas distâncias, tamanho e massa de uma partícula

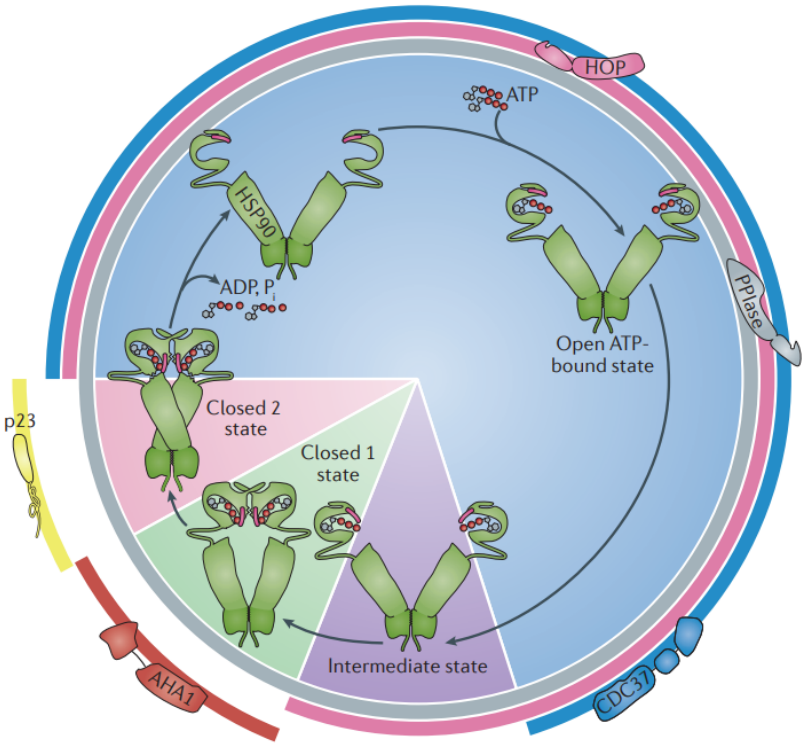

Figura 1.9: Esquematização do ciclo de funcionamento de membros das Hsp90. Uma chaperona em conformação aberta (azul) se liga a ATP, transiciona para um estado intermediário (roxo) ao fechar a tampa do NTD, que então sofre uma dimerização, levando a estrutura a um estado fechado (verde) e, ao interagir com o MD fecha ainda mais a estrutura (vermelho), com a reabertura da estrutura com a liberação de ADP e fosfato fechando o ciclo. Também está indicado em que etapas participam determinadas co-chaperonas (segmentos de circunferência externos) (Schopf et al., 2017).

ou mesmo envelopes proteicos, como é o caso do espalhamento de raios-X a baixos ângulos (SAXS).

Cada técnica experimental tem vantagens e limitações. Cristalografia permite a obtenção de mapas atômicos com resoluções da ordem de 2Å a depender da amostra, porém o processo de cristalização muitas vezes se baseia em tentativa e erro, e nem sempre é possível devido a características da própria proteína. Ao mesmo tempo, por se tratar de apenas um mapa, regiões flexíveis ou são mostradas em apenas uma conformação ou estão ausentes. Assim, o que se tem é um retrato da proteína como se ela fosse um ente rígido, o que nem sempre é o que ocorre em solução. RMN, por outro lado, tem suas medidas realizadas em solução, porém tem como contrapartida a exigência de que a massa da proteína de estudo seja relativamente baixa (Barbosa et al., 2013).

SAXS permite a obtenção de medidas em solução para diversos sistemas em diversas

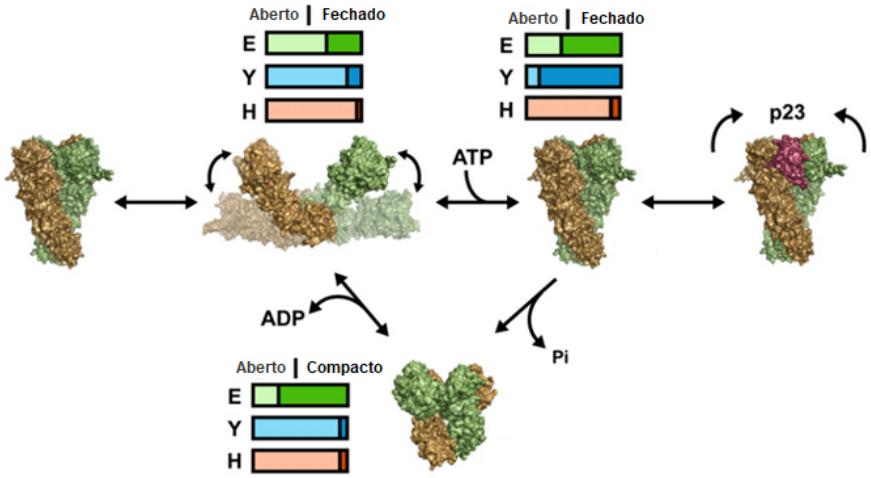

Figura 1.10: Equilíbrio entre estados de Hsp90 em diferentes espécies. A depender do ponto do ciclo, diferentes organismos contém diferentes proporções de cada conformação, indicado por uma barra, com o tom mais claro indicando proporção de estruturas abertas e o mais escuro, fechadas. (adaptado de Southworth e Agard, 2008).

condições de tampão, dando resultados que permitem a avaliação de parâmetros estruturais gerais e flexibilidade de uma proteína à obtenção de modelos a baixa resolução (Borges et al., 2016). Apesar disso, possíveis problemas devido a baixa razão sinal-ruído, por exemplo, devem ser estudados a depender da amostra.

Cada técnica experimental vem encontrando avanços fazendo uso tanto de novas tecnologias de manuseio de amostra e detecção de sinal quanto, devido a computadores cada vez mais velozes, de ferramentas computacionais que vão desde métodos de Monte Carlo para a elaboração de modelos (Svergun, 1999) quanto ferramentas como *machine-learning* (Franke et al., 2018) para o cálculo de parâmetros estruturais de proteínas (seções 3.4 e 3.5). Por conta desses avanços e da complexidade das questões a serem respondidas no âmbito biológico, o uso de técnicas complementares para a elucidação de determinado fenômeno se torna cada vez mais comum e necessário, havendo desenvolvimentos também com relação a metodologias de análise híbridas utilizando SAXS em conjunto com dados de cristalografia e/ou RMN (Petoukhov e Svergun, 2005; Bernadó et al., 2005; Bernadó et al., 2007). A disponibilidade de modelos e dados em grandes bancos de dados de fácil acesso ajudam não apenas na curadoria de conteúdo mas também no uso de resultados anteriores na elaboração de novos estudos (Berman et al., 2000; Valentini et al., 2014).

Capítulo 2

Objetivos

2.1 Justificativa

Neste trabalho foram estudadas características estruturais e dinâmicas de proteínas chaperonas de diferentes famílias e organismos utilizando a técnica de SAXS, ilustrando como a técnica pode contribuir para estudos aprofundados e abrangentes e combinar informações anteriormente obtidas para a elaboração de modelos detalhados do comportamento de proteínas em solução.

2.2 Objetivo Geral

Aplicação de metodologias de análise de SAXS, especialmente mais recentes, para estudo de estrutura e dinâmica de proteínas chaperonas em solução.

2.3 Objetivos Específicos

- Estudar possíveis particularidades entre membros da família de Hsp70 humana a baixa resolução;
- Avaliar o papel de pontes dissulfeto na manutenção da estrutura *coiled-coil* das GrpE a partir de medições em diferentes solventes;
- Obtenção da estrutura a baixa resolução da Hsp90 de *Aedes aegypti* e comparação das características de estrutura e dinâmica com Hsp90 de outros organismos.

Capítulo 3

Materiais e Métodos

3.1 Expressão e purificação de proteínas

As proteínas foram expressas pelos membros do Laboratório em Bioquímica e Biofísica de Proteínas (LBBP) do Instituto de Química de São Carlos da Universidade de São Paulo (IQSC-USP), coordenado pelo Prof. Dr. Júlio César Borges, e do Laboratório de Bioquímica de Proteínas, coordenado pelo Prof. Dr. Carlos Henrique Inácio Ramos, do Instituto de Química da Universidade Estadual de Campinas (IQ-UNICAMP).

O processo de expressão e purificação de proteínas é complexo e depende da proteína que se deseja estudar, porém ele segue algumas regras gerais. Um plasmídeo (fragmento de DNA circular bacteriano) que codifica determinada proteína é inserido em bactérias por meio de um processo chamado *transfecção*. A membrana plasmática das células é fragilizada por meio de choque térmico, expondo-as brevemente a temperaturas de 42°C com subsequente imersão em gelo. Uma vez que o plasmídeo também confere imunidade às células a certos anticorpos, estes são adicionados ao meio com as células para selecionar as que tenham capturado o plasmídeo com sucesso, e estas crescem em meio de cultura celular. Essas culturas podem ser utilizadas para *indução*, que é o processo em que as células crescem rapidamente em erlenmeyers para que seja possível obter a proteína a ser expressada em alguns mg. Após isso as células são lisadas e os componentes são separados inicialmente por centrifugação e posteriormente por processos cromatográficos, como utilizando colunas de Ni^{2+} para cromatografia de afinidade a proteínas que tenham His-tag, e cromatografia por exclusão em gel para seleção de acordo com tamanho e massa molecular.

As proteínas foram expressas em *E. coli* com pureza mínima de 95%.

3.2 Espalhamento de Raios-X a Baixos Ângulos: Teoria

O fundamento de SAXS é similar ao da difração de raios-X, sendo aplicado para estudo de partículas com tamanho de ordem maior que 10Å. Observando a lei de Bragg

$$sen\theta = \frac{\lambda}{2d}, \tag{3.1}$$

pode-se constatar que maiores valores de d implicam em menores ângulos de medição, o que indica que objetos maiores são observadas a ângulos menores (Roe, 2000). A Figura 3.1 mostra de maneira simplificada o aparato experimental para SAXS, que consiste em uma fonte de raios-X (podendo ser um filamento ou uma fonte síncrotron), monocromador para seleção de comprimento de onda, sistema de colimação de feixe, porta-amostra (podendo ser, por exemplo, um capilar ou uma armação de metal com janelas de mica) e detector, tudo isso em vácuo (Barbosa et al., 2013). Também é importante notar a existência de um *beam stopper* (inglês para), que consiste em uma pequena peça para que incidam os raios-X não espalhados pela amostra, sendo importante uma vez que sem ela o detector seria danificado nessa região.

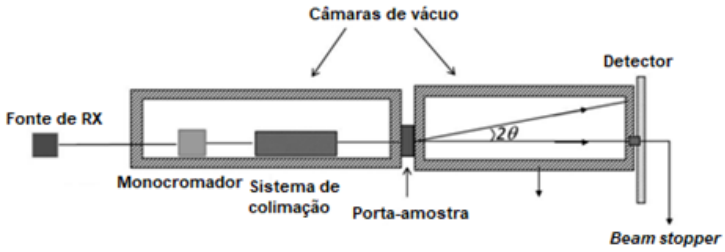

Figura 3.1: Esquematização somplificada de um experimento de SAXS. Fótons de raios-X emitidos por uma fonte (*X-ray source*), seja por filamento ou luz síncrotron, têm uma estreita faixa de comprimento de onda selecionada por um monocromador (*monochromator*), são colimados (*collimation system*) e incidem na amostra contida no porta-amostra (*sample-holder*). Ao interagir com os elétrons da amostra os fótons são espalhados a um ângulo 2θ, sendo medidos no detector. Um *beam stopper* é responsável por recever os fótons são espalhados, preservando o detector. Todo o processo ocorre em câmaras de vácuo (*vacuum chambers*) (Barbosa et al., 2013).

Uma vez que a intensidade de espalhamento é medida para ângulos baixos pode-se apenas considerar contribuições devidas a espalhamento coerente, ou seja, apenas para fótons de mesma energia que os incidentes.

3.2.1 Equação de Debye para a descrição do processo de espalhamento

A teoria de SAXS foi elaborada inicialmente por Debye no início do século XX (Debye, 1915; Debye e Fuoss, 1954). Ele partiu das expressões da interação de cargas com ondas eletromagnéticas para entender como elétrons interagiam com radiação incidente, entendendo-os como dipolos oscilantes. Uma das grandezas centrais para o SAXS é o *vetor de espalhamento* q, dado por

$$q = \frac{4\pi}{\lambda}\text{sen}(\theta), \tag{3.2}$$

em que λ refere-se ao comprimento de onda da radiação incidente e θ à *metade* do ângulo de espalhamento. A partir da expressão de Hertz para campos emitidos por este tipo de fonte, chegou-se a uma expressão que descreve a intensidade de espalhamento para vários átomos com elétrons orbitando em anéis em torno do núcleo, dada por (Debye, 1915)

$$\frac{I(q)}{I(0)} = \frac{Np}{R^2}\, r_0^2\, \frac{1+cos^2\theta}{2} \sum_{n=0}^{n=p-1} \frac{sen\left(4ka\ sen\left(\frac{n\pi}{p}\right) sen\frac{v}{2}\right)}{\left(4ka\ sen\left(\frac{n\pi}{p}\right) sen\frac{v}{2}\right)}, \tag{3.3}$$

em que o volume irradiado contém N átomos com p elétrons cada, $r_0 = e^2/mc^2$ é o raio clássico do elétron, R a distância entre o ponto de observação e a origem do sistema de coordenadas, a o raio de órbita do elétron em torno do núcleo (com elétrons distribuídos igualmente nesta órbita), k o número de onda e $senv/2$ um termo relacionado às direções das componentes do campo relacionado à onda espalhada. O índice n diz respeito ao n-ésimo elétron da somatória e θ é o ângulo de observação com relação a feixe incidente.

Pode-se chegar à expressão de espalhamento deduzida por Debye por meio de uma abordagem mais direta que consiste em pensar mais geometricamente a partir da Figura 3.2. Por se tratar de um espalhamento coerente as magnitudes de espalhamento não se alteram, mas sim as fases detectadas, com as ondas espalhadas sendo representadas pelo fator $\exp(i\phi)$, com ϕ sendo a fase. Esta é definida como $2\pi/\lambda \times a$, em que a a diferença de caminho ótico entre as ondas espalhada e incidente.

Os fótons incidem na amostra a um vetor $\vec{k_0}$ e são espalhados a um vetor $\vec{k}$, com um ângulo de 2θ entre eles, definindo assim o vetor de espalhamento $\vec{q} = \vec{k} - \vec{k_0}$. A diferença de caminho ótico é dada por $a = -\vec{r}\cdot\vec{q}$, o que implica que $\phi = -(2\pi/\lambda)(\vec{r}\cdot\vec{q})$. O vetor

$\vec{k_0}$ tem magnitude $2\pi/\lambda$ e, por se tratar de espalhamento elástico, $\vec{k_1}$ também. Assim, por meio da lei dos cossenos e das respectivas substituições trigonométricas, tem-se que

$$|\vec{q}| = |\vec{k} - \vec{k_0}| = \frac{4\pi}{\lambda} sen\theta, \tag{3.4}$$

sendo λ o comprimento de onda dos fótons incidentes e θ a *metade* do ângulo entre os vetores dos feixes incidente e espalhado.

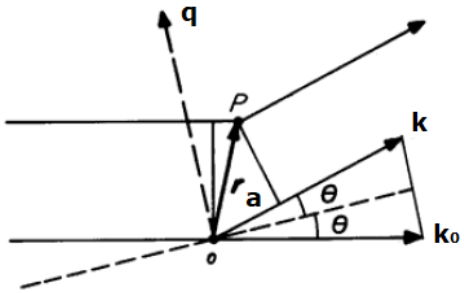

Figura 3.2: Esquematização geométria do fenômeno de espalhamento por dois pontos (em O e em P) separados por uma distância r. A diferença de caminho ótico é dada por a, e estão mostrados os vetores dos raios incidente ($\vec{k_0}$) e espalhado ($\vec{k}$) a um ângulo 2θ, além do vetor de espalhamento q (adaptado de Glatter e Kratky, 1982).

A amplitude de espalhamento referente a determinado q pode ser obtido por meio da soma de todas as ondas espalhadas, ou seja, de cada termo $\exp(-i\vec{q}\cdot\vec{r})$. Pode-se fazer uso da densidade eletrônica $\rho(\vec{r})$ para descrição de uma partícula, porém é importante levar em conta também a densidade eletrônica do solvente (ρ_s), de modo que $\Delta\rho(\vec{r}) = \rho(\vec{r}) - \rho_s$. Assim, a expressão obtida para a amplitude de espalhamento $A(\vec{q})$ é

$$A(\vec{q}) = \int_V dV \Delta\rho(\vec{r}) \exp(-i\vec{q}\cdot\vec{r}), \tag{3.5}$$

ou seja, fica explicitado que a amplitude de espalhamento corresponde à transformada de Fourier da densidade eletrônica da partícula. A intensidade é dada por

$$I(\vec{q}) = AA^* = \int_{V_1}\int_{V_2} dV_1 dV_2 \Delta\rho(\vec{r_1})\Delta\rho(\vec{r_2}) \exp[-i\vec{q}\cdot(\vec{r_1} - \vec{r_2})]. \tag{3.6}$$

Importante notar que na expressão da intensidade são envolvidos pares de elementos de densidade eletrônica. Aqui entra uma aproximação importante para SAXS: uma vez que as partículas estão orientadas aleatoriamente em solução, o sinal medido corresponde à conformação média de todas as orientações possíveis (representado pela operação $\langle \ldots \rangle_\Omega$) (Andrews, 2004). Desse modo,

$$I(\vec{q}) = \langle AA^* \rangle_\Omega = \int_{V_1} \int_{V_2} dV_1 dV_2 \Delta\rho(\vec{r_1}) \Delta\rho(\vec{r_2}) \langle \exp[-i\vec{q} \cdot (\vec{r_1} - \vec{r_2})] \rangle_\Omega. \tag{3.7}$$

Calculando a média rotacional da exponencial, tem-se que

$$\langle \exp(-i\vec{q} \cdot \vec{r}) \rangle = \frac{\text{sen}(qr)}{qr}, \tag{3.8}$$

onde $r = |\vec{r_1} - \vec{r_2}|$, o que reduz a expressão da intensidade à equação de Debye. Esse cálculo faz com que haja perda de informação em SAXS com relação às fases, havendo, na prática, perda de resolução em comparação com a difração de raios-X. Desse modo, deve-se haver cuidado ao se realizar análises de dados de SAXS para que não sejam tomadas conclusões errôneas. Em difração, erros de protocolo implicam em um impedimento da obtenção de resultados, enquanto em SAXS é possível, por exemplo, seguir todo o processo de análise mesmo com amostras agregadas, danificadas por radiação ou com problemas de subtração de fundo. Porém, com o uso de técnicas auxiliares e consciência das aproximações utilizadas no desenvolvimento teórico da técnica, SAXS torna-se uma técnica poderosa de verificação de hipótese e modelagem aprofundada (Casey et al., 2014).

Ao mesmo tempo, a intensidade de espalhamento de uma amostra é proporcional a dois fatores importantes, de modo que $I(q) \propto P(q) \times S(q)$, onde $P(q)$ é o chamado *fator de forma*, que descreve o formato da partícula, e $S(q)$ é chamado *fator de estrutura*, que depende de interações entre as partículas espalhadoras. Neste trabalho serão assumidas amostras diluídas para minimizar as interações de modo que $S(q) \to 1$ e, portanto, a curva de espalhamento contenha apenas informações sobre o formato das partículas espalhadoras.

3.2.2 Função distribuição de distâncias: p(r)

Para sistemas nos quais pode-se verificar flutuações de densidade eletrônica, como é o caso de amostras de proteína em tampão ou em água, por exemplo, é possível definir uma função que descreva essas flutuações de modo que se tenha uma descrição sobre qual a distância média na qual a densidade não varia consideravelmente (Debye e Bueche, 1949; Debye et al., 1957). Dados dois pontos, A e B, separados por uma distância r (tal que A está na posição $\vec{u}$ e B em $\vec{u} + \vec{r}$), cada um com uma respectiva densidade eletrônica $\rho_A(\vec{u})$ e $\rho_B(\vec{u} + \vec{r})$ e, portanto, $\Delta\rho(r) = \rho_A(\vec{u}) - \rho_B(\vec{u} + \vec{r})$. Pode-se calcular as diferenças de densidade eletrônica para quaisquer pontos A e B separados por r e tirar a média,

repetindo sobre todas as distâncias possíveis.

Para o caso específico de proteínas em solução será considerado um sistema com duas densidades eletrônicas, sendo uma para a proteína ($\rho(\vec{r'})$) e outra para o tampão (média, dada por ρ_s), de modo que as densidades $\rho_A(\vec{u})$ e $\rho_B(\vec{u}+\vec{r})$ podem assumir esses dois valores. Assim, considerando que em solução as partículas podem estar em qualquer orientação possível, pode-se definir uma função, denominada *função de correlação* $\gamma(r)$, que descreve a extensão média das flutuações de densidade eletrônica por meio de (Svergun e Koch, 2003)

$$\gamma(r) = \left\langle \int \Delta\rho(\vec{u})\Delta\rho(\vec{u}+\vec{r})d\vec{u} \right\rangle_{\Omega}. \tag{3.9}$$

Pela própria equação, pode-se observar dois casos extremos: quando a distância é muito baixa não são observadas variações na densidade, de modo que $\gamma(r) = 1$, enquanto para distâncias altas a diferença varia aleatoriamente, fazendo com que $\gamma(r) = 0$. Desse modo, a função de correlação, na prática, age como um histograma entre distâncias interatômicas da proteína, com contagem máxima para distâncias muito pequenas e sem contagens a partir da distância máxima medida entre um par átomos da proteína, sendo esta última a chamada D_{max}. A intensidade de espalhamento pode ser escrita, portanto, a partir da função de correlação por meio de

$$I(q) = 4\pi \int_0^{D_{max}} r^2\gamma(r)\frac{\text{sen}(qr)}{qr}dr, \tag{3.10}$$

sendo a integral até D_{max} uma vez que por argumentos físicos mostrou-se que a função de correlação é nula acima deste valor. A partir da $\gamma(r)$ é possível definir a *função de distribuição de pares*, $p(r)$, dada por $p(r) = r^2\gamma(r)$. Isso faz com que a expressão acima seja dada por

$$I(q) = 4\pi \int_0^{D_{max}} p(r)\frac{\text{sen}(qr)}{qr}dr. \tag{3.11}$$

A relação entre as funções $\gamma(r)$ e $p(r)$ está ilustrada para uma esfera na Figura 3.3. Uma vez que a $I(q)$ é dada por uma transformada de Fourier, pode-se obter a $p(r)$ a partir da transformada inversa de modo que

$$p(r) = \frac{1}{2\pi^2} \int_0^{\infty} I(q)(qr)\text{sen}(qr)dq, \tag{3.12}$$

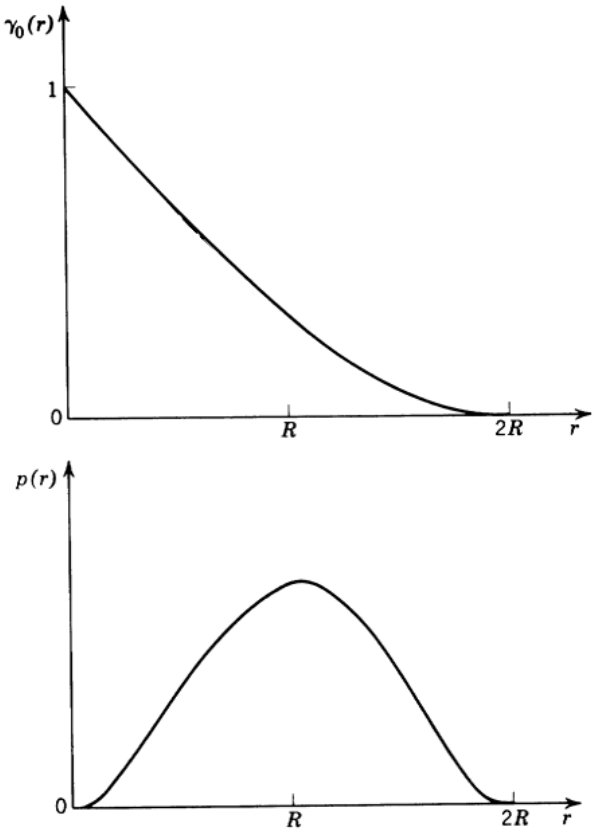

Figura 3.3: Gráficos representando as funções $\gamma(r)$ (acima) e $p(r)$ (abaixo) para uma esfera (Guinier e Fournet, 1995).

mas a $p(r)$ não pode ser obtida diretamente desta maneira uma vez que as intensidades são conhecidas apenas para um certo conjunto de valores de q. Foi elaborado um método geral (Glatter, 1977) onde a $p(r)$ é aproximada por uma série de funções no espaço real que buscam ajustar a transformada da função aos dados (Svergun et al., 2013). Dependendo da abordagem utilizada é escolhida uma base diferente, com diferenças sutis de metodologia (Glatter, 1977; Moore, 1980; Svergun, 1992). De maneira geral, portanto, a $p(r)$ aproximada ($p_{ap}(r)$) é dada por

$$p_{ap}(r) = \sum_{i=1}^{n_f} c_i \phi_i(r), \tag{3.13}$$

sendo $\phi_i(r)$ as funções que constituem a base, c_i os coeficientes de ajuste e n_f sendo o número de funções utilizadas na aproximação, com os valores de r indo de 0 a D_{max}. As funções $\phi_i(r)$ da base podem diferir de uma metodologia para outra, podendo ser utilizados, por exemplo, splines cúbicos (Glatter, 1977; Svergun et al., 2013). Assim, as intensidades aproximadas $I_{ap}(q)$ correspondentes à $p_{ap}(r)$ podem ser encontradas por meio de

$$I_{ap}(q) = 4\pi \sum_{i=1}^{n_f} c_i \int_0^{D_{max}} \phi_i(r) \frac{\text{sen}(qr)}{qr} dr, \tag{3.14}$$

com a obtenção de uma curva suave de intensidades que aproxime os dados experimentais a partir de uma estimativa de D_{max}.

Como ilustra a Figura 3.4, a função $p(r)$ apresenta perfis típicos para formas diferentes de partícula, e uma característica comum a todas é o início com $p(r) = 0$ para $r = 0$ e o decaimento suave da função conforme se aproxima de $r = D_{max}$. Assim, além de técnicas experimentais complementares, o uso deste critério para a estimativa de D_{max} é de grande auxílio no cálculo da $p(r)$ correspondente. "Ombros" no início da curva ou decaimento de modo que D_{max} seja muito alto podem indicar agregação ou problemas de subtração de fundo. Exemplos de $p(r)$ para amostras agregadas estão na Figura 3.5.

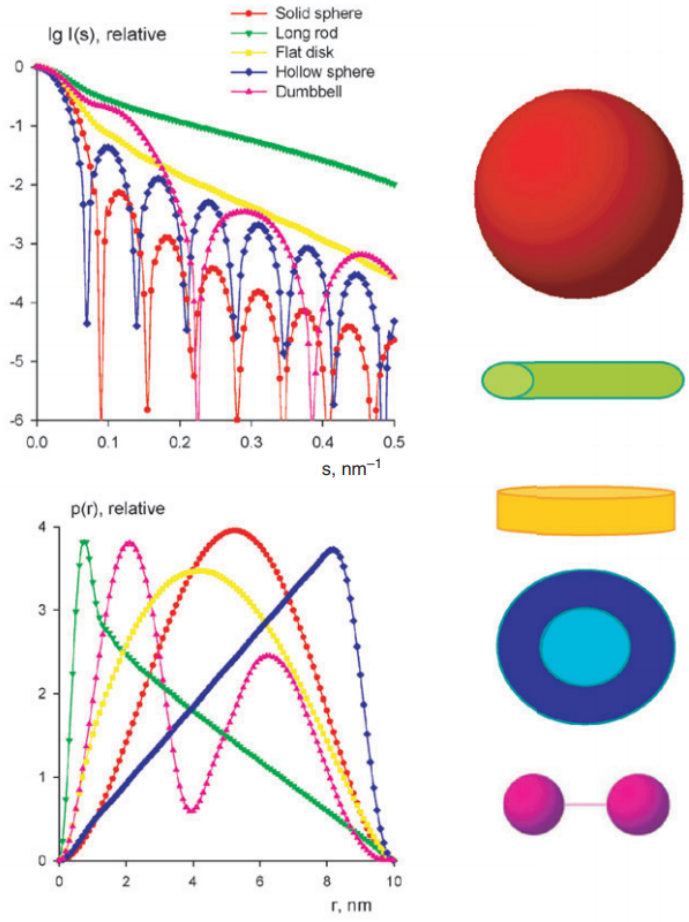

Figura 3.4: Comparação das curvas de espalhamento (acima) e $p(r)$ (abaixo) para diferentes corpos de mesma dimensão máxima: esfera (vermelho), cilindro comprido (verde), cilindro achatado (amarelo), casca esférica (azul) e haltere (magenta). A $p(r)$ ajuda a discernir quando corpos com perfis de espalhamento similares contém topologias distintas ou não (Svergun e Koch, 2003).

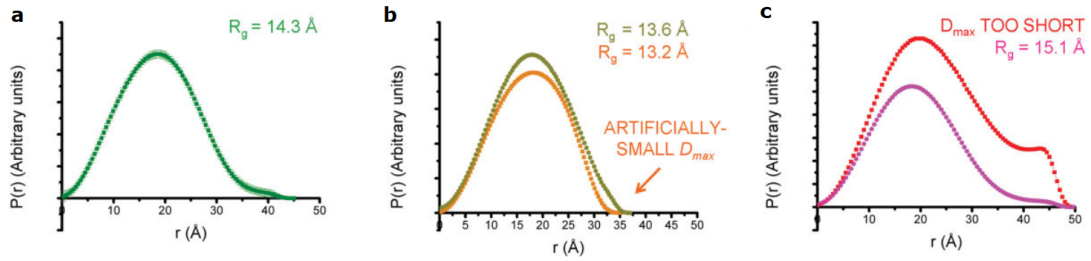

Figura 3.5: Perfis de $p(r)$ característicos quando o valor determinado de D_{max} é condizente (a), subestimado (b, laranja) ou superestimado (c, roxo). Para amostras com interações repulsivas é um pouco mais complexo (b, ocre), assim como quando há agregação pronunciada (c, vermelho) (Jacques e Trewhella, 2010).

A esse tipo de análise se dá o nome *Inverse Fourier Transform* (inglês para Transformada Inversa de Fourier, IFT) por se basear na transformada de Fourier dos dados e não em modelos pré-concebidos para a estrutura da partícula no desenvolvimento da análise. Há métodos que buscam ajustar os dados experimentais a diferentes modelos, seja de figuras geométricas ou de modelos de estruturas de polímeros, a partir de expressões analíticas de espalhamento, como os implementados no programa GENFIT (Spinozzi et al., 2014). Estes métodos dependentes de modelos não serão tratados aqui.

Pela $p(r)$ é possível definir o raio de giração R_g de uma partícula e a intensidade de espalhamento direto $I(0)$, que são grandezas muito importantes para análises de dados. O R_g indica a dispersão dos N elementos de massa da partícula em torno do seu centro de massa a partir das distâncias r (Goldstein et al., 2002), sendo definido como

$$R_g^2 = \frac{1}{N}\sum_{i=1}^{N} r_i^2, \tag{3.15}$$

e é, portanto, um indicador do *tamanho* da partícula. Ele está relacionado à $p(r)$ por meio de

$$R_g^2 = \frac{\int_0^{D_{max}} r^2 p(r) dr}{2\int_0^{D_{max}} p(r) dr}. \tag{3.16}$$

O $I(0)$, inacessível experimentalmente por conta da presença do *beam stopper* e da impossibilidade de distinguir o feixe incidente do espalhado nessa região, por meio de substituição direta na equação 3.11, é dado por

$$I(0) = 4\pi \int_0^{D_{max}} p(r)dr, \tag{3.17}$$

e está diretamente relacionado a grandezas físicas como massa molecular e volume da partícula espalhadora. No caso do volume a relação se dá por meio da expressão

$$I(0) = (\Delta\rho)^2 V^2, \tag{3.18}$$

com V sendo o volume e $\Delta\rho$ a densidade eletrônica média da partícula.

3.2.3 Análise de Guinier para a extração de parâmetros estruturais e verificação de interações entre partículas

A partir da relação entre $I(q)$ e a $p(r)$ pode-se chegar a uma aproximação muito importante para análises de SAXS. A grandeza $I(0)$ tem relação direta com a *massa molecular* (MM) da partícula espalhadora, mas não pode ser obtida com medida direta de maneira similar à intensidade quando há espalhamento devido à presença do *beam stopper*, responsável pela atenuação do feixe, que é muito mais intenso que os que contém fótons espalhados. A expressão que relaciona essas grandezas a ângulos muito baixos é dada por

$$I(q) \approx I(0)\ \exp\left[-\left(\frac{R_g^2}{3}\right)q^2\right], \tag{3.19}$$

que é a chamada *lei de Guinier*. Tomando o logaritmo natural de ambos os lados, esta fórmula assume uma forma mais prática dada por

$$\ln[I(q)] = \ln[I(0)] - \frac{R_g^2}{3}q^2, \tag{3.20}$$

permitindo encontrar R_g e $I(0)$ a partir da linearização de um gráfico de $\ln[I(q)]$ vs q^2. Porém essa linearização só pode ser feita até um ponto q_{max}, e a chamada *região de Guinier*, que corresponde ao subconjunto de dados para o qual essas aproximações são válidas, é dada por $q_{max}R_g < 1.3$ para proteínas globulares, porém podendo ser limitada por valores tais que $q_{max}R_g < 1.1$ para moléculas mais elongadas ou de cadeia desordenada (Zheng e Best, 2018).

O gráfico de Guinier permite observar interações entre partículas na solução, representadas por desvios de linearidade na região de $q \to 0$. Interações atrativas, como a formação

de agregados, aparecem como aumentos de intensidade, enquanto repulsivas, como interações eletrostáticas são observadas a partir de diminuições. Os desvios estão ilustrados na Figura 3.6. A depender do quão pronunciados os desvios de linearidade pode-se ter que realizar novas medidas ou mesmo modificar a amostra, adicionando sal, por exemplo, para a blindagem de cargas superficiais de aminoácidos carregados, ou correndo a amostra por cromatografia de exclusão para a remoção de agregados e estados oligoméricos indesejados. Quando desvios são sutis é possível utilizar os dados, porém descartando da análise de Guinier os pontos problemáticos. Também é possível estudar essas interações buscando modelá-las a partir do fator de estrutura $S(q)$ para entender mais sobre a natureza delas. Antes do experimento é importante observar as imagens bidimensionais do detector para avaliar a presença de radiação parasita, que pode levar a falsos positivos para agregação de amostra, por exemplo.

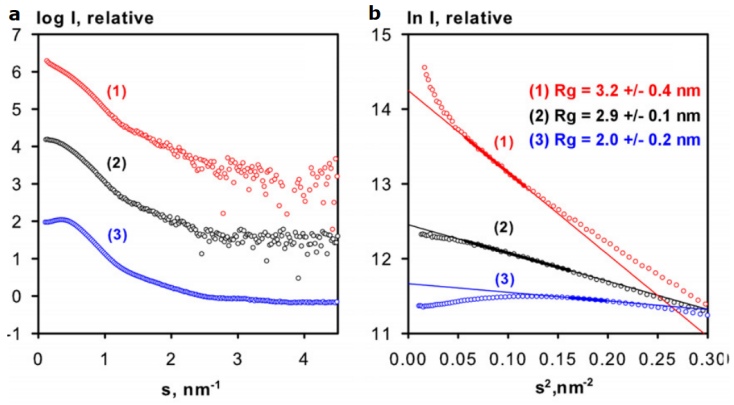

Figura 3.6: Exemplo de curvas de espalhamento (**a**) e gráficos de Guinier (***b***b) para condições sem interações (preto), com agregação (vermelho) e com interações repulsivas entre partículas (azul). Os pontos usados para a linearização estão preenchidos com as respectivas retas. Mesmo com interações ainda é possível obter linearizações com um número razoável de pontos, obtendo parâmetros sensivelmente diferentes, portanto cuidado deve ser tomado ao trabalhar com resultados de amostras com gráficos de Guinier não-ideais (Mertens e Svergun, 2010).

Comparando esta abordagem com a obtenção dos parâmetros estruturais pela $p(r)$ há prós e contras. A abordagem é bastante prática e simples, consistindo apenas em uma linearização, porém a depender do tamanho da proteína e do estado da amostra pode ser que o ajuste só seja possível com poucos pontos, levando a incertezas apreciáveis nos parâmetros obtidos, especialmente para o R_g. Por conta dessa questão no número de pontos, o uso da $p(r)$ tem a vantagem de poder utilizar todos os pontos da curva no cálculo da função, de modo que os valores apresentam uma flutuação bem menor. Porém,

o uso da $p(r)$ tem uma sutileza: a $p(r)$ deve descrever corretamente a partícula, ou seja, D_{max} deve ser determinado corretamente ao utilizar programas de cálculo da função. Por esses motivos, o cálculo dos parâmetros a partir de ambos os métodos é recomendado para que seja feita uma comparação e verificação dos resultados.

3.2.4 Lei de Porod: Análises de Kratky e Porod-Debye para estudos flexibilidade

O comportamento assintótico de uma curva de SAXS para valores mais altos de q também permite a obtenção rápida de informações estruturais de uma partícula pela chamada *lei de Porod* (Debye et al., 1957; Glatter e Kratky, 1982; Roe, 2000). A partir de um modelo de duas fases, ou seja, com dois valores fixos de densidade eletrônica pode-se mostrar que a intensidade de espalhamento pode ser descrita assintoticamente por

$$I(q \to \infty) \to \frac{2\pi S(\Delta\rho)^2}{q^4}, \tag{3.21}$$

em que S é a área superficial da partícula e $\Delta\rho$ a diferença de densidade eletrônica entre a partícula e o solvente. A demonstração completa encontra-se no Apêndice A.

O formalismo é desenvolvido assumindo que a superfície entre as fases é abrupta (do inglês *sharp*, indicando que não há transição suave entre os valores de densidade eletrônica), e, portanto, a lei só vale para proteínas bem enoveladas, e quando se trata de uma cadeia muito flexível, a superfície vista por SAXS não é tão bem delimitada por se tratar de uma estrutura média, e isso faz com que a lei não consiga descrever o espalhamento para este tipo de partícula a ângulos mais altos (para uma melhor descrição desses casos deve-se utilizar modelos de polidispersão). De maneira mais fundamental, o modelo com dois valores fixos de densidade eletrônica não é mais aplicável pois flexibilidade estrutural leva ao surgimento de flutuações nos valores de densidade eletrônica dentro da partícula, tornando a superfície menos bem definida e, portanto, desviando o comportamento do previsto pelo modelo e, portanto, pela lei de Porod.

A partir da lei de Porod pode-se elaborar ferramentas gráficas para avaliação do estado de flexibilidade de proteínas. Uma vez que a lei dita que para estruturas com volume bem delimitado a intensidade decai com q^{-4}, então fazendo uso de gráficos de $q^2I(q)$ vs q é esperado que haja um decaimento em q^{-2} para estruturas globulares na região de validade da lei, e que ocorra um platô para estruturas flexíveis. Devido à lei de Guinier o início da curva é ascendente, então os perfis para proteínas globulares mostram um formato de sino

característico (Glatter e Kratky, 1982), enquanto o perfil de proteínas flexíveis se deve à lei de Debye, que modela polímeros de cadeia aleatória, dada por

$$\frac{I(q)}{I(0)} = \frac{2(\chi - 1 + e^{-\chi})}{\chi^2}, \tag{3.22}$$

sendo $\chi = (qR_g)^2$. Para proteínas que estejam em um meio-termo, como proteínas de domínios globulares com *linkers* flexíveis, se observa um comportamento intermediário, como visto na Figura 3.7.

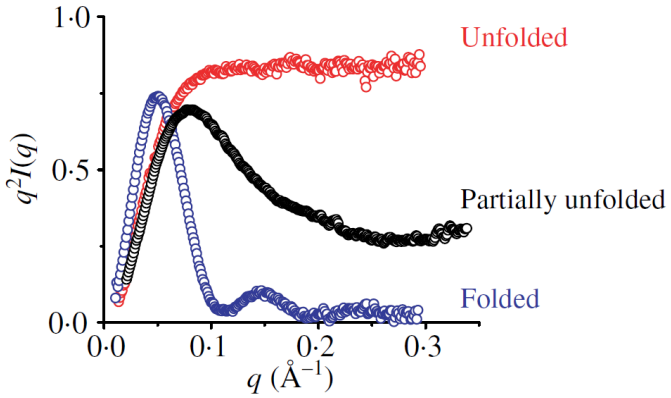

Figura 3.7: Três perfis típicos de gráficos de Kratky: para proteínas bem enoveladas (azul), parcialmente desenoveladas (preto) e desenoveladas (vermelho) (Putnam et al., 2007).

Em alguns casos, os gráficos de Kratky indicam que uma proteína é globular mesmo que esta apresente flexibilidade, de modo que esta forma de análise não se torna mais tão confiável (Bernadó, 2010). Um tipo de gráfico alternativo e interessante é o de Porod-Debye, que consiste em gráficos q vs $q^4I(q)$ e q^4 vs $q^4I(q)$. Uma vez que se baseiam também na lei de Porod, estes devem exibir um platô na região de validade da lei para proteínas que sejam mais rígidas, e ausência quando trata-se de uma mais flexível, como ilustrado na Figura 3.8 (Rambo e Tainer, 2011). O gráfico usando q^{-4} nas abscissas facilita a visualização do platô.

3.2.5 Teoria da informação de Shannon aplicada ao SAXS

Ao se calcular a $p(r)$ que melhor se ajusta aos dados, pode-se interpretar as intensidades experimentais $I(q)$ como um conjunto de amostragens discretas de uma função contínua. Utilizando teoria de informação e comunicação (Shannon e Weaver, 1949), pode-se compreender o canal de amostragem como o intervalo no qual foram tomados os dados, representados pelos pontos de q_{min} a q_{max}. Uma questão no SAXS é de que pode-se aumentar

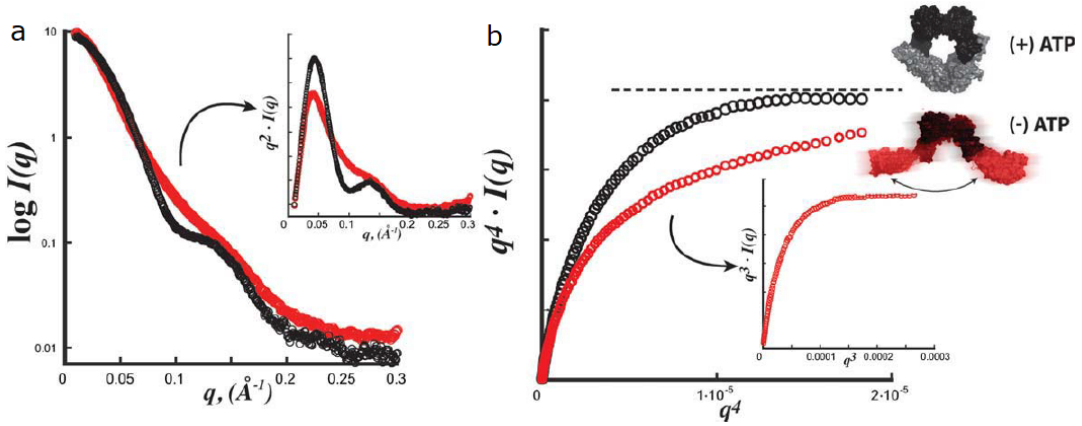

Figura 3.8: **a.** Curva experimental com gráfico de Kratky no inset. **b.** Gráfico de Porod-Debye, indicando diferentes graus de flexibilidade a partir da presença de platô nos dados, com gráfico de $q^3I(q)$ vs q^3 no inset, indicando que o gráfico de Porod-Debye tem poder de distinção para flexibilidade de proteínas (adaptado de Rambo e Tainer, 2011).

a resolução do experimento ao aumentar o tamanho do canal ao tomar dados para valores maiores de q_{max}, porém na prática o ruído é maior para valores maiores de q, e ao mesmo tempo a frequência de amostragem (intervalo de q entre cada ponto experimental), dada por $\Delta q = \pi/D_{max}$, se altera. O teorema da amostragem de Nyquist–Shannon permite a obtenção da frequência de amostragem mínima (ou seja, menor número de pontos experimentais) para a reconstituição do sinal de SAXS, definida pelo número de canais de Shannon N_s, dado por

$$N_s = D_{max}\frac{(q_{max} - q_{min})}{\pi}. \tag{3.23}$$

Um ponto curioso é que os valores típicos de N_s obtidos são da ordem de 20, ou seja, em torno de 3% do número total de pontos de um experimento típico no LNLS indo até $q_{max} = 5$ nm^{-1}. Portanto, é dito que o sinal de SAXS contém uma *sobreamostragem* (do inglês *oversampling*), ou seja, há uma amostragem excessiva com relação ao número de canais de Shannon suficientes para a descrição da intensidade ajustada pela $p(r)$. Em teoria esses pontos "excessivos" não trazem informações adicionais sobre a proteína de estudo, porém, devido a questões de ruído, o número de canais de Shannon não deve ser tomado tão literalmente uma vez que a frequência de amostragem, na prática, é influenciada por ruído. O valor de N_s pode ser interpretado, portanto, como o mínimo de parâmetros independentes que podem descrever uma curva de SAXS (Moore, 1980). De acordo com o teorema de amostragem, a curva pode ser descrita pelos N_s pontos e a sua reconstrução se

dá por meio da fórmula de interpolação de Shannon-Whittaker, dada por (Shannon, 1949)

$$I(q) = \sum_{n=1}^{\infty} I\left(\frac{n\pi}{D_{max}}\right) \frac{\text{sen}(qD_{max} - n\pi)}{qD_{max} - n\pi}. \tag{3.24}$$

Ou seja, a intensidade pode ser aproximada por meio de uma soma ponderada das intensidades espaçadas regularmente de acordo com o N_s, sendo os pesos dados pelas funções $\text{sen}(x)/x$ (também chamada de função $sinc(x)$), para $x = qD_{max} - n\pi$. Uma vez que os dados só são tomados até um certo q_{max}, a soma é truncada em um valor $n = q_{max}D_{max}/\pi$. Um ponto curioso dessa interpolação é de que quando q é múltiplo inteiro de π/D_{max}, devido à presença do fator seno, a intensidade é determinada apenas pela intensidade neste ponto (uma vez que todos os outros termos são nulos), enquanto quando q não é múltiplo a intensidade será dada de acordo com a soma ponderada de todos os pontos de Shannon. Assim, a fórmula de interpolação determina que os pontos de Shannon compõem um conjunto de observações independentes, e o resto dos pontos acabam estando correlacionados a essas observações (Rambo e Tainer, 2013). O conjunto de pontos que compõe essas observações se dá a uma frequência de π/D_{max}, porém se a amostragem for de frequência menor o sinal não pode ser recuperado apenas com base neste conjunto de informações. Como na prática experimental trata-se questões que envolvem ruído em medições, para estudar a informação contida dada uma frequência de amostragem pode-se aplicar o teorema de Shannon-Hartley, que diz respeito à taxa máxima de informação que pode ser transmitida C a partir de uma razão sinal-ruído S/N, dado por

$$R < C = \frac{2\pi}{D_{max}} \log_2\left(1 + S/N\right), \tag{3.25}$$

em que R refere-se ao espaçamento Δq entre os pontos de Shannon. Esta expressão, portanto, diz qual o intervalo máximo entre medidas para que a informação estrutural possa ser recuperada quando se estuda uma proteína de determinado tamanho (D_{max}) a determinadas condições (representadas em S/N, estando relacionado a fatores como concentração, fluxo de feixe etc).

Como questão prática, o q_{min} (ou seja, o primeiro ponto a ser utilizado em análise) não deve ser mais alto que o primeiro ponto de Shannon, de modo que $q_{min} < \pi/D_{max}$ deve ser satisfeita, senão ocorre perda de informação (Svergun e Koch, 2003). Isso se manifesta, por exemplo, no programa GNOM, onde quando esta condição não é satisfeita surge uma

mensagem de aviso, pois é como se o valor de D_{max} escolhido para a geração da $p(r)$ não fosse de fato "enxergado" uma vez que falta o primeiro ponto de Shannon para a descrição da curva experimental.

3.3 Métodos de análise: de metodologias bem estabelecidas a avanços recentes

Nos últimos anos, houve um aumento vertiginoso do uso de SAXS para estudos em biologia estrutural (Jacques e Trewhella, 2010), seja para validação de estruturas cristalográficas como também para modelagens de macromoléculas flexíveis. Pouco mais de duas décadas atrás, a técnica de SAXS fornecia pouco mais que um punhado de parâmetros estruturais sobre tamanho e massa molecular, porém especialmente desde a publicação do programa DAMMIN (Svergun, 1999) um novo leque de aplicações da técnica experimental se abriu para pesquisadores, permitindo a obtenção de envelopes de proteínas em solução com resolução de nm, com o SAXS se tornando mais do que uma técnica de baixa resolução. A presente seção busca apresentar tanto técnicas de análise já estabelecidas e amplamente utilizadas, quanto novos métodos propostos em publicações recentes que endereçam desde questões referentes a incertezas de medida quanto à obtenção de envelopes que contenham flutuações internas de densidade eletrônica.

3.3.1 Análise de Pico de Guinier (GPA)

Uma questão comum em análises de dados de SAXS é a determinação de R_g e $I(0)$ de maneira confiável. Comumente é feita a análise pela aproximação de Guinier, que respeita o limite $qR_g < 1.3$, mas para partículas mais elongadas é recomendado que esse limite seja reduzido para, por exemplo, $qR_g < 1.1$ (Zheng e Best, 2018). Ela também pode ser feita pela $p(r)$, que utiliza um valor máximo de q (q_{max}) respeitando $q_{max} > 8/R_g$ (alguns autores utilizam $q_{max} > 7/R_g$) (Petoukhov et al., 2012). Assim, utilizando mais dados o cálculo dos parâmetros pela $p(r)$ é tida como mais confiável, mas ao mesmo tempo ambiguidades na função podem levar a incertezas maiores nos valores obtidos.

Recentemente foi elaborada uma metodologia de validação do conjunto de dados válido para a aproximação de Guinier como também de R_g e $I(0)$, denominada *Análise de Pico de Guinier* (do inglês *Guinier Peak Analysis*, ou simplesmente GPA) (Putnam, 2016). O

método consiste em multiplicar ambos os lados da aproximação de Guinier por q para encontrar um pico que explicite a região de Guinier, observando o gráfico de $qI(q)$ vs q^2. Multiplicando por q ambos os lados:

$$qI(q) = I(0)\sqrt{q^2}\ \exp\left[-\left(\frac{R_g^2}{3}\right)q^2\right], \tag{3.26}$$

onde se encontra um máximo para valor de abscissa (demonstração nos apêndices)

$$qR_g \approx 1.22. \tag{3.27}$$

Portanto, quando visualizamos o gráfico assim percebemos um pico nesse ponto e depois a função decai. Esse ponto em geral pertence à região de Guinier, portanto dá informações sobre que pontos podem ser escolhidos para a linearização. Sem isso o que se pode fazer é se calcular o R_g e depois verificar se o ajuste se encontra na região de Guinier. Assim, esse gráfico permite a constatação *a priori* de que conjunto de dados permite a obtenção de R_g e $I(0)$ pela aproximação de Guinier.

A GPA também pode ser feita com um gráfico adimensional de $qR_gI(q)/I(0)$ vs $(qR_g)^2$. A multiplicação de q por R_g suprime informação sobre o tamanho da partícula, enquanto a divisão de $I(q)$ por $I(0)$ suprime a massa molecular. Essa prática de fazer gráficos adimensionais é comum em física de polímeros porém pouco usada no contexto de biomoléculas (Durand et al., 2010).

Para isso, basta multiplicar ambos os lados da equação da GPA por $R_g/I(0)$, o que nos deixa com

$$qR_g\frac{I(q)}{I(0)} = \frac{I(0)}{I(0)}(qR_g)\ \exp\left[-\left(\frac{R_g^2}{3}\right)q^2\right], \tag{3.28}$$

$$qR_g\frac{I(q)}{I(0)} = \sqrt{q^2R_g^2}\ \exp\left[-\left(\frac{R_g^2}{3}\right)q^2\right]. \tag{3.29}$$

Essa forma não altera o valor de qR_g do máximo em relação à GPA anterior, que continua se encontrando em $qR_g \approx 1.22$. Substituindo de volta na equação teremos para a ordenada o valor de aproximadamente 0.7428. Na prática, a comparação da posição do pico teórico com o obtido pelo gráfico permite uma validação dos valores de R_g e $I(0)$ encontrados. Uma vez que a GPA se baseia na aproximação de Guinier a validade do método se dá principalmente para proteínas globulares. Também é possível fazer o

caminho inverso: ajustar os dados tentando buscar para quais valores de R_g e $I(0)$ o gráfico terá o pico nos valores definidos. Apesar de ser uma forma menos precisa, ainda é uma forma de estimativa, porém esta assume uma proteína perfeitamente globular, o que não corresponde à realidade mas é uma primeira aproximação.

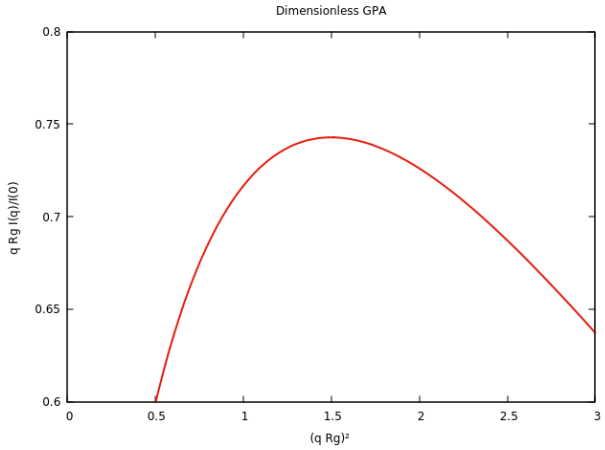

Figura 3.9: Lei de Guinier retratada em um gráfico de Guinier Peak Analysis adimensional.

3.3.2 Razão de elongamento

Ao fazer análises de SAXS que independam do ajuste de um modelo geométrico específico, informações sobre a forma de uma proteína podem ser obtidas a partir da sua $p(r)$. O perfil da função indica, em linhas gerais, qual a geometria da molécula, e o que se pode observar ao comparar os perfis característicos de uma esfera e de um cilindro ou elipsoide, é o deslocamento do máximo global da função para valores mais baixos de r em relação ao D_{max}. Portanto, a posição relativa de r nos permite ter uma ideia do quão elongada é uma proteína. Uma medida foi recentemente definida para buscar quantificar esse elongamento (Putnam, 2016). Apesar de não ser a primeira do tipo, trata-se de uma definição matemática intuitiva para o seu objetivo. Seja R o valor de r no qual a função $p(r)$ tem seu valor máximo, pode ser calculado uma *razão de elongamento* (do inglês *elongation ratio*, ou ER) a partir da equação

$$ER = \frac{\int_R^{D_{max}} p(r)dr}{\int_0^R p(r)dr}. \tag{3.30}$$

ER é, portanto, a razão entre as áreas do gráfico entre R e D_{max}, e entre 0 e R. Para uma esfera este valor é de aproximadamente 0.94, e aumenta conforme mais elongada a molécula de estudo. A interpretação dessa razão como um "fator de elongamento", porém, se restringe apenas a esferas e figuras cilíndricas e elipsoidais uma vez que, por exemplo, o perfil de uma esfera oca também apresenta um desvio similar em R apesar de ter um perfil diferente de um cilindro. Ao mesmo tempo, um disco apresenta um desvio de R para valores maiores em relação ao perfil de uma esfera. Portanto, *conceitualmente* como algo relacionado ao elongamento de uma proteína, esta razão se restringe a apenas certos formatos. Foi encontrado que há uma correlação entre a ER e o fator de anisotropia de elipsoides e altura de cilindros, porém no artigo original não foram discutidas em detalhes as relações entre ER e diferentes parâmetros estruturais (Putnam, 2016), e isso pode ser feito fazendo uso de $p(r)$ obtidas analiticamente.

3.3.3 Kratky Adimensional

Assim como a GPA tem uma versão adimensional para que se possa extrair mais informação sobre a molécula, o gráfico de Kratky também tem uma versão adimensional seguindo a mesma regra de multiplicar q por R_g e dividir $I(q)$ por $I(0)$ a fim de permitir uma análise "normalizada" de diferentes tipos de proteína. Desse modo, o gráfico de Kratky, que é originalmente de $q^2I(q)$ vs q, quando adimensional assume a forma $(qR_g)^2I(q)/I(0)$ vs qR_g (Durand et al., 2010).

Para a análise de um gráfico do tipo podemos traçar dois perfis de casos extremos idealizados e imaginar que proteínas reais surjam como casos intermediários. Para o primeiro caso seria para uma proteína perfeitamente globular, seguindo a aproximação de Guinier, e o segundo seria para um polímero de cadeia aleatória e secção transversal pontual, representando uma proteína completamente desenovelada, de acordo com a lei de Debye. Para o primeiro caso o máximo do gráfico de Kratky adimensional ocorre para $qR_g = \sqrt{3}$ (demonstração em apêndice). Para esse valor na abscissa encontramos que $(qR_g)^2I(q)/I(0) \approx 1.1$ nesse ponto de máximo. O perfil da curva é aproximadamente Gaussiano, mas levemente assimétrico.

Para o segundo caso basta aplicar a lei de Debye, que é dada por

$$\frac{I(q)}{I(0)} = \frac{2\big(\chi - 1 + \exp(-\chi)\big)}{\chi^2}, \tag{3.31}$$

na qual $\chi = (qR_g)^2$. Deixando na forma de Kratky tem-se, já substituindo χ,

$$(qR_g)^2 \frac{I(q)}{I(0)} = \frac{2\big((qR_g)^2 - 1 + \exp(-(qR_g)^2)\big)}{(qR_g)^2}. \tag{3.32}$$

Ao observar o gráfico, é nítido que há um comportamento assintótico que se aproxima de 2 para valores de qR_g cada vez maiores. Importante notar que isso é um caso extremo com a limitação de que considera a secção transversal como sendo pontual, um modelo limitado. Na prática para proteínas desenoveladas se vê um comportamento que continua crescendo mesmo após atingir o platô esperado.

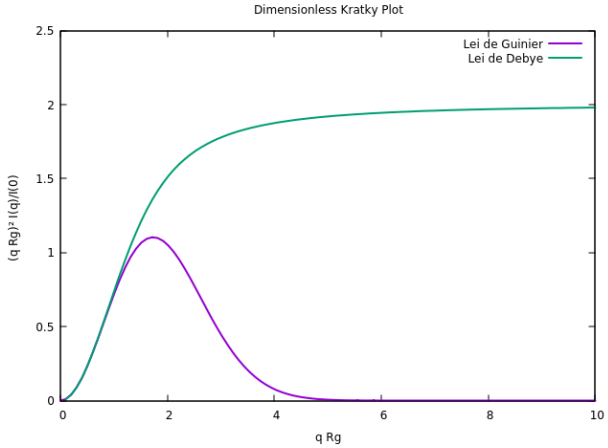

Figura 3.10: Dois perfis extremos de cadeia polipeptídica em gráficos de Kratky adimensional: proteína bem enovelada (roxo) e cadeia aleatória (turquesa)

3.3.4 Entropia da distribuição de R_g: atribuindo um número à desordem

Os gráficos de Kratky e Porod-Debye são ferramentas visuais muito importantes para a determinação do estado de flexibilidade de uma proteína. Porém, por serem métodos apenas gráficos, uma quantificação desse parâmetro seria de grande uso. Em cima disso, se elaborou um método que busca atribuir um número a esse grau de desordem que se baseia na descrição do ensemble de estados acessíveis à proteína por meio de uma distribuição de tamanhos, sendo um método independente da estrutura da proteína (Burger et al., 2016).

Pode-se descrever a intensidade de espalhamento de uma proteína em solução a partir da combinação linear das intensidades referentes a todas as conformações acessíveis a ela ponderando pela ocorrência de cada uma, e uma forma de escrever isso é utilizando funções de distribuição de probabilidade. Para uma proteína composta por N átomos mapeados pelo vetor de $3N$ coordenadas $\vec{r} = (x_1, y_1, z_1, \ldots, x_N, y_N, z_N)$, com intensidades referentes

a cada conformação dadas por $I(q,\vec{r})$ e distribuição de probabilidade conformacional dada por $P(\vec{r})$, a intensidade total é dada por

$$I(q) = \int_{\vec{r}} I(q,\vec{r})P(\vec{r})d\vec{r}, \tag{3.33}$$

sendo a integral sobre todas as conformações acessíveis. A entropia pode ser calculada a partir da distribuição de probabilidades, porém a descrição a partir da posição dos átomos é pouco prática para este fim. Deste modo, uma primeira aproximação pode ser feita: a descrição das conformações de uma proteína a partir do tamanho delas, dada pelo R_g, e assumindo-as como esferas com densidade de carga homogênea. Uma limitação dessa descrição é de que mudanças radicais de conformação que não alterem o R_g não são percebidas ao mesmo tempo que não se trata de uma boa modelagem para proteínas com graus mais elevados de anisotropia. Portanto, as intensidades parciais serão dadas por

$$I_{esf}(q,R_g) = I_{esf}(0)\frac{9}{(qaR_g)^6}\left[\text{sen}(qaR_g) - qaR_g\text{cos}(qaR_g)\right]^2, \tag{3.34}$$

sendo o fator $a = \sqrt{5/3}$ devido à conversão do raio da esfera para o R_g associado.

Utilizando o modelo proposto, denominado modelo de distribuição de raios de giração (Radius-of-gyration Distribution model, ou simplesmente RgD), a intensidade passa a ser dada por (Burger et al., 2016)

$$I_{\mu,\sigma}(q) = \int_0^\infty I_{esf}(q,R_g)\ P_{\mu,\sigma}(R_g)\ dR_g, \tag{3.35}$$

sendo $P_{\mu,\sigma}(R_g)$ a distribuição de probabilidade de R_g. Para trabalhos com raios-X, em geral adota-se esta distribuição como sendo uma log-normal, dada por

$$P_{\mu,\sigma}(R_g) = \frac{1}{\sqrt{2\pi}\sigma R_g}\ \exp\left[-\frac{(\ln[R_g]-\mu)^2}{2\sigma^2}\right], \tag{3.36}$$

sendo μ a média da distribuição e σ o desvio padrão. Fisicamente pode ser imposto que $P(0) = 0$, e o uso da log-normal é útil pois $P(R_g \to 0) \to 0$ e por ser definida apenas para $R_g > 0$. A intensidade $I_{\mu,\sigma}(q)$ é ajustada aos dados experimentais variando os parâmetros μ e σ e uma vez encontrados os parâmetros otimizados $\bar{\mu}$ e $\bar{\sigma}$ calcula-se a entropia S por meio de (Beirlant et al., 1997)

$$S = -\int_0^\infty P_{\bar{\mu},\bar{\sigma}}(R_g)\ \ln\left[P_{\bar{\mu},\bar{\sigma}}(R_g)\right]dR_g, \tag{3.37}$$

em que é importante reforçar que se trata de uma distribuição contínua, levando a um mínimo de $-\infty$ e não zero, como ocorre para a entropia de Shannon (caso discreto). No final, tem-se que

$$S = \frac{1 + \ln[\pi\bar{\sigma}^2]}{2} + \bar{\mu}. \tag{3.38}$$

A partir do cálculo de S pode-se comparar os valores de entropia calculados para perfis experimentais do SASBDB (Valentini et al., 2014) e, a partir dos gráficos de Kratky, atribuir valores de S a graus de desordem. A Figura 3.11 indica que, apesar dos valores não serem tão bem definidos para cada perfil de desordem, há valores aproximados esperados para cada tipo de gráfico de Kratky. Os valores de entropia dados pelo modelo RgD são de difícil avaliação por conta da questão das incertezas e ruído em dados de SAXS, porém o modelo nitidamente aponta para uma tendência nos valores de entropia calculados. Esta metodologia, apesar de simples, pode servir para elucidar ambiguidades entre proteínas globulares ou multiméricas com domínios globulares grandes conectados por *linkers* flexíveis, pois estas apresentam perfis de Kratky muito similares (Bernadó, 2010).

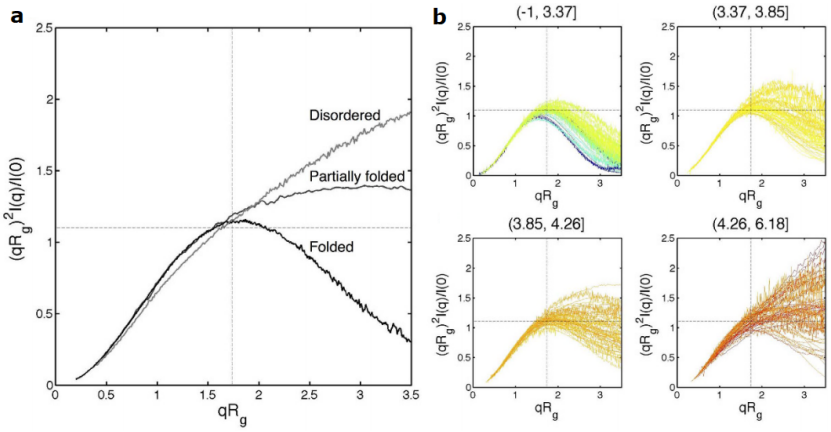

Figura 3.11: Gráficos adimensionais de Kratky: à esquerda (**a**) perfis típicos para proteínas bem enoveladas (preto), parcialmente desenoveladas (cinza escuro) e completamente desenoveladas (cinza claro), e à direita (**b**) os perfis para proteínas com valores crescentes de entropia (Burger et al., 2016).

3.3.5 Determinação de incertezas em medidas de SAXS

Uma forma de determinação de incerteza em experimentos de SAXS consiste em realizar múltiplas tomadas de quadros de curtos espaços de tempo e o cálculo da média e desvio padrão entre os diferentes quadros. Essa prática é comum especialmente para a avaliação

de dados devido à exposição da amostra à radiação, porém a depender da concentração proteica na amostra tornam-se necessárias tomadas cada vez mais longas de dados, o que dificulta a utilização da técnica para proteínas instáveis em solução. Outra forma, porém, consiste no uso de um modelo para determinação das incertezas (Sedlak et al., 2017) a partir da estatística de contagem de fótons pelo detector.

O modelo se baseia em três hipóteses: a contagem de fótons como sendo um processo de Poisson, o espalhamento pelo tampão sendo considerado aproximadamente constante para o intervalo de q medido e processos de medição para as medições de amostra e de tampão como sendo estatisticamente independentes com relação às incertezas obtidas.

Em experimentos de SAXS, como os feitos no LNLS, os fótons incidem em detectores de placa que realizam contagens em pixels, e um dos primeiros passos da redução de dados consiste no cálculo de uma média entre as contagens feitas por pixels referentes a um mesmo vetor de espalhamento q, como ilustrado na Figura 3.12. O número de pixels depende do q considerado, portanto a intensidade é (Sedlak et al., 2017)

$$I(q) = \frac{1}{N(q)} \sum_{i=1}^{N(q)} n_i, \tag{3.39}$$

o que resulta, portanto, considerando pixels como estatisticamente independentes, em uma variância dada por (Sedlak et al., 2017)

$$\sigma^2(q) = \frac{1}{N(q)^2} \sum_{i=1}^{N(q)} \sigma_i^2 = \frac{1}{N(q)^2} \sum_{i=1}^{N(q)} [n_i - I(q)]^2. \tag{3.40}$$

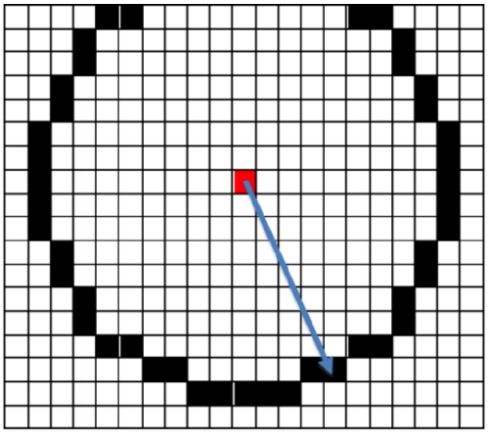

Figura 3.12: Processo de integração de um perfil bidimensional de detector de SAXS para uma curva unidimensional. Em preto os pixels a serem levados em conta na integração e em vermelho a região do feixe incidente (ou seja, a região do *beam stopper*) (Franke et al., 2015).

Naturalmente, uma maior presença de pixels não-funcionais leva a um aumento da variância, portanto um aumento é esperado nas regiões de fronteira entre placas para detectores de mais de uma placa. Nesse ponto entra mais uma das hipóteses do modelo. Assumindo que se trata de um processo de Poisson tem-se que a média e a variância têm mesmo valor e, portanto, que $\sigma_i^2 = I(q)$, o que resulta em

$$\sigma^2(q) = \frac{I(q)}{N(q)}. \tag{3.41}$$

Em SAXS, as curvas analisadas subtraem a intensidade do tampão da intensidade da amostra, ou seja, fazem que a $I_p(q) = I_a(q) - I_t(q)$, em que os índices p, a e t se referem, respectivamente, às medidas para proteína, amostra e tampão. Ao mesmo tempo, assumindo que cada medida é estatisticamente independente, a variância, por propagação, é dada por $\sigma_p^2(q) = \sigma_a^2(q) + \sigma_t^2(q)$. A partir da terceira hipótese, que diz que o espalhamento do tampão é aproximadamente constante, pode-se considerar que $I_t(q) = cI_a(q_{arb})$, ou seja, que o espalhamento do tampão corresponde simplesmente à intensidade espalhada pela amostra a um valor de q arbitrário q_{arb} multiplicada por uma constante c, de modo que esta constante depende do valor de q_{arb} escolhido. Uma vez que as intensidades de espalhamento do solvente divergem de uma constante para valores de q muito baixos, deve-se atentar para a escolha de um q_{arb} que corresponda à região de q em que não haja essa divergência.

Usando o que foi apresentado, tem-se que, para o q_{arb} escolhido,

$$I_a(q_{arb}) = I_p(q_{arb}) + I_t(q_{arb}) = I_p(q_{arb}) + cI_a(q_{arb}) \tag{3.42}$$

$$I_p(q_{arb}) = (1 - c)I_t(q_{arb}), \tag{3.43}$$

o que implica que

$$I_t(q_{arb}) = \frac{I_p(q_{arb})}{(1 - c)}. \tag{3.44}$$

Substituindo os valores de $\sigma_a^2(q)$ e $\sigma_t^2(q)$ e usando a equação acima, obtêm-se

$$\sigma_p^2(q) = \frac{I_p(q)}{N(q)} + \frac{I_t(q)}{N(q)} + \frac{I_t(q)}{N(q)}, \tag{3.45}$$

$$\sigma_p^2(q) = \frac{I_p(q)}{N(q)} + \frac{2cI_a(q_{arb})}{N(q)}, \tag{3.46}$$

o que resulta em uma expressão mais diretamente aplicável, dada por

$$\sigma_p^2(q) = \frac{1}{N(q)}\left(I_p(q) + \frac{2cI_p(q_{arb})}{(1-c)}\right). \tag{3.47}$$

Ou seja, a partir da relação constante entre o espalhamento da amostra e o do tampão a certo valor q_{arb} (fator dependente da amostra e da escolha de q_{arb}) e do número de pixels usados na obtenção das $I(q)$ (fator dependente do aparato experimental). A questão do número de pixels não é exata *a priori* por depender do tipo de detector utilizado, porém uma aproximação pode ser feita para uma geometria comum de experimento tal como a presente no LNLS. Para pixels de mesmo tamanho e detectores de placa dispostos perpendicularmente ao feixe incidente de raios-X há uma relação aproximadamente linear com r, a distância entre o pixel e o feixe direto, de modo que $N(q)$ é proporcional a $2\pi r$. Porém é possível escrever r em função de θ , de modo que

$$\tan 2\theta = \frac{r}{L_{ad}}, \tag{3.48}$$

em que L_{ad} é a distância entre a amostra e o detector e r é a distância entre o feixe direto e o pixel. Ao mesmo tempo, diretamente da definição do vetor de espalhamento segue que

$$\theta = \text{sen}^{-1}\left(\frac{\lambda q}{4\pi}\right), \tag{3.49}$$

e substituindo 3.49 em 3.48 tem-se que

$$N(q) = 2\pi L_{ad}\tan\left(2\text{sen}^{-1}\left(\frac{\lambda q}{4\pi}\right)\right), \tag{3.50}$$

que pode ser aproximada como

$$N(q) \approx L_{ad}\lambda q. \tag{3.51}$$

Deste modo a variância pode ser escrita como

$$\sigma_p^2(q) = \frac{1}{kq}\left(I_p(q) + \frac{2cI_p(q_{arb})}{(1-c)}\right), \tag{3.52}$$

em que k é uma constante dada por $k \approx L_{ad}\lambda$. No artigo também são sugeridos valores para k e c para o caso de aparelhos de bancada e para laboratórios de luz síncrotron. Importante reforçar que o modelo leva em conta apenas a incerteza relacionada à medida em si, e não a respeito de variações no manuseio e preparo de amostras, intrínseco a muitos trabalhos do ramo biológico, que podem ser fonte para contribuições adicionais à incerteza do experimento.

3.4 Programas de computador

3.4.1 Determinação automatizada dos parâmetros do gráfico de Guinier: AUTORG

A linearização da região de Guinier é um dos primeiros passos da análise de SAXS, dando informações sobre o estado da amostra e também parâmetros sobre o tamanho (R_g) e massa ($I(0)$) da partícula espalhadora, sendo eles muito utilizados em análises posteriores. Portanto, uma boa determinação desses parâmetros é fundamental para a qualidade da análise.

Um critério muito utilizado até hoje para a análise de Guinier é o visual a fim de avaliar a qualidade da linearização e os pontos envolvidos no ajuste. Porém, para evitar subjetividade de análise e para acompanhar o uso cada vez maior de linhas de análise automatizadas o programa AUTORG foi criado (Petoukhov et al., 2007). Esse programa realiza iterativamente linearizações do gráfico de $\ln\left[I(q)\right]$ vs q^2, buscando minimizar o χ^2 e atribuindo à incerteza não apenas o dos parâmetros de ajuste por propagação, mas também atribuindo uma margem de erro à determinação dos pontos a serem ajustados.

O AUTORG primeiramente descarta pontos que tenham variação muito grande com relação ao resto no início do conjunto de dados a fim de descartar problemas referentes ao *beam stopper*, e então seleciona como primeiro intervalo a região em que a intensidade decai por uma ordem de magnitude. Se busca ajustar funções cúbicas para determinar se há indícios de agregação ou repulsão entre partículas. Após isso o programa faz ajustes utilizando um número mínimo de três pontos e impondo que $q_{min}R_g < 1$ e $q_{max}R_g < 1.3$, calculando R_g e $I(0)$ para cada intervalo, sendo descartados os casos em que hajam desvios sistemáticos. Caso não haja intervalos consistentes, as condições referentes a qR_g são flexibilizadas automaticamente pelo programa.

Os intervalos são avaliados de acordo com o número de pontos adotado para a linea-

rização e o χ^2 calculado, sendo escolhidos os parâmetros do melhor ajuste, sendo atribuída uma incerteza tal que seja a soma da obtida por propagação e dos outros R_g encontrados para os outros intervalos. O programa também dá um fator de qualidade que leva em conta cinco fatores: número de intervalos encontrados, incerteza do R_g, número de pontos descartados no início da curva, desvios que indiquem agregação ou repulsão e se houve ou não flexibilização das condições referentes a qR_g. O fator de qualidade varia de 0 a 1, sendo 0 referente a não haver linearização possível. As configurações do programa, como mínimo de pontos para ajuste e limites de $q_{min}R_g$ e $q_{max}R_g$ podem ser alterados pelo usuário.

3.4.2 Obtenção da p(r): GNOM

Como já apontado na seção 3.2.2, a partir da $p_{ap}(r)$ pode-se calcular uma curva aproximada $I_{ap}(q)$ que descreva os dados experimentais para a proteína de estudo. O ajuste da curva calculada aos dados experimentais se dá minimizando o χ^2, dado por

$$\chi^2 = \frac{1}{N_{gl}-1}\sum_{j=1}^{N_p}\left[\frac{I(q_j)-I_{ap}(q_j)}{\sigma_j}\right]^2, \tag{3.53}$$

sendo N_{gl} o número de graus de liberdade do sistema, e que pode ser reescrito em termo dos coeficientes a serem minimizados por meio de (Svergun et al., 2013)

$$\chi^2 = \sum_{j=1}^{N_p}\left[\frac{I(q_j)-4\pi\sum_{i=1}^{n_f}c_i\Phi_i(q_j)}{\sigma_j}\right]^2, \tag{3.54}$$

sendo σ_j a incerteza associada ao j-ésimo ponto experimental, N_p o número de pontos experimentais e

$$\Phi_i(q_j) = \int_0^{D_max}\phi_i(r)\frac{\text{sen}(qr)}{qr}dr. \tag{3.55}$$

Porém, é possível encontrar $p(r)$ com comportamento oscilatório, o que é fisicamente problemático. Portanto, busca-se encontrar a função mais suave que possa descrever os dados fazendo uso de um multiplicador de Lagrange α que adicione uma restrição $N_{c'}$ dada por

$$N_{c'} = \sum_{i=1}^{n_f-1}(c_{i+1}-c_i)^2, \tag{3.56}$$

de modo que minimize a expressão $\chi + \alpha N_{c'}$. Ou seja, introduz-se um termo que busca um α que faz com que a solução em que o menor aumento de χ com a maior redução de oscilações seja escolhida. Assim, um maior α reduz o número de coeficientes presentes na aproximação.

O programa GNOM (Svergun, 1992) busca utilizar critérios automatizados de percepção acerca de diferentes elementos de ajuste de maneira quantitativa para encontrar o α. Os critérios têm pesos e distribuições próprias comentadas posteriormente, sendo dados eles (Svergun, 1992):

- DISCRP (peso 1, $\sigma = 0.30$): trata-se do χ^2 no espaço recíproco, que aumenta conforme α aumenta. Valor esperado < 1, sendo 0.7 estipulado como ideal;

- OSCILL (peso 3, $\sigma = 0.60$): reflete a suavidade da $p(r)$, sendo dado por

$$OSCILL = \frac{|dp(r)/dr|}{|p(r)|}, \tag{3.57}$$

  e diminui conforme α aumenta. Valor ideal de 1.1, equivalente ao critério calculado para o caso de uma esfera;

- STABIL (peso 3, $\sigma = 0.12$): retrata a estabilidade da função para variações no α, sendo dada por

$$STABIL = \frac{d(\ln|p(r)|)}{d(\ln[\alpha])}, \tag{3.58}$$

  tendo zero como valor ideal.

- SYSDEV (peso 3, $\sigma = 0.12$): indica possíveis desvios sistemáticos de $I_{ap}(q)$ com relação aos dados experimentais $I(q)$. Seja N_p o número de pontos experimentais e N_- o *número de vezes que a diferença* $I(q) - I_{ap}(q)$ *muda de sinal*, o critério é dado por

$$SYSDEV = \frac{N_-}{N/2}, \tag{3.59}$$

  sendo 1 o seu valor ideal, ou seja, o esperado é o mesmo número de pontos experimentais acima e abaixo da curva aproximada obtida. Aumenta conforme α também

aumenta. Importante ressaltar que desvios em diferentes partes da curva podem ter maior ou menor impacto na estrutura correspondente à curva calculada uma vez que cada região de q equivale a diferentes tamanhos no espaço real.

- POSITV (peso 1, $\sigma = 0.12$): corresponde à norma relativa da parte positiva da $p(r)$, tendo 1 como valor ideal. Ou seja, o critério reflete que idealmente a $p(r)$ não deve apresentar valores negativos.

- VALCEN (peso 1, $\sigma = 0.12$): diz respeito à validade do intervalo escolhido no espaço real, sendo a norma relativa da parte central da $p(r)$. Idealmente tende a 1, sendo 0.95 para uma esfera, e <<1 caso seja grande demais.

Ou seja, trata-se de critérios que são empregados para análise humana de ajustes, porém quantificados de modo que um computador possa realizar essa análise para determinar o α tal que haja o melhor ajuste possível com o mínimo de oscilações. A partir dos valores de cada critério o GNOM também calcula um valor de estimativa de qualidade Q dado por

$$Q = \frac{\sum_i P(i)W(i)}{\sum_i W(i)}, \tag{3.60}$$

sendo W(i) os pesos dados a cada critério e $P(i)$ uma gaussiana associada a cada critério, dados por

$$P(i) = \exp\left[- \left(\frac{A(i) - B(i)}{\sigma(i)} \right)^2 \right], \tag{3.61}$$

sendo $A(i)$ o valor ideal estipulado para o critério, $B(i)$ o valor calculado tal que minimiza α e $\sigma(i)$ o desvio encontrado a partir da largura a meia altura da distribuição. Assim sendo, os critérios de maior peso são referentes à oscilação da $p(r)$, à estabilidade da solução relativa a variações em α e aos desvios sistemáticos com relação aos dados experimentais.

A estimativa de qualidade Q vai de 0 a 1 conforme a $p(r)$ satisfaça mais os critérios. Ao lado do valor de Q o GNOM também apresenta a avaliação da solução em palavras, indo de ruim ou suspeita a excelente.

3.4.3 Fatores de forma a partir de arquivos PDB: CRYSOL

A técnica mais popular para a obtenção de estruturas de proteínas em alta resolução é a cristalografia com difração de raios-X, porém devido à possibilidade do processo de

cristalização poder introduzir artefatos nos resultados, como contatos adicionais ou mesmo estados oligoméricos artificiais, os resultados nem sempre retratam o que ocorre *in vivo* ou em solução. Desse modo, essas estruturas de alta resolução podem não corresponder ao que é medido em solução, que é mais próximo do que ocorre *in vivo*. Ao mesmo tempo é difícil fazer observações sobre regiões flexíveis uma vez que estas partes da proteína dificilmente são resolvidas, ou, quando observadas, podem à primeira vista parecer estáticas sem que sejam tomadas informações complementares. Uma vez que experimentos de SAXS de proteínas ocorrem em solução, comparar os dados obtidos com estruturas conhecidas é de grande uso para verificar a validade de estruturas depositadas e bancos de dados como o *Protein Data Bank* (inglês para banco de dados de proteínas, ou PDB) (Berman et al., 2000). Um método de comparação é o cálculo de curvas teóricas de SAXS para estruturas em formato PDB e o ajuste destas aos dados obtidos para avaliar a compatibilidade entre ambos, e o principal programa que faz esse cálculo é o CRYSOL (Svergun et al., 1995), que, ao contrário dos programas antecessores, modela a camada primeira camada de solvente em contato com a macromolécula, evitando erros sistemáticos presentes anteriormente.

Um modelo para descrição de uma macromolécula em solução consiste em uma partícula com densidade eletrônica $\rho_p(\vec{r})$ cercada de solvente com densidade ρ_0 e uma camada de hidratação aproximada por uma borda de espessura Δ e densidade ρ_b, que pode ser diferente de ρ_0 (de modo que $\delta\rho = \rho_b - \rho_0$). No modelo as densidades eletrônicas são médias, não fazendo distinção entre flutuações internas de cada região, mas faz distinção entre solvente ligado à proteína e solvente livre em solução (também chamado *bulk*), o que é observado, por exemplo, para água. Para o cálculo do espalhamento teórico de uma proteína é necessário levar em conta a contribuição de cada região, de modo que as intensidades são dadas por (Svergun et al., 1995)

$$I(q) = \langle |A_p(\vec{q}) - \rho_0 A_e(\vec{q}) + \delta\rho A_b(\vec{q})|^2 \rangle_\Omega, \tag{3.62}$$

em que $A_p(\vec{q})$, $A_e(\vec{q})$ e $A_b(\vec{q})$ se referem às amplitudes da partícula isolada, do volume excluído dela e da camada de hidratação. A amplitude de espalhamento para a partícula com átomos localizados em coordenadas $\vec{r}_j = (r_j, \omega_j)$ é dada pela expressão comum à cristalografia

$$A_a(\vec{q}) = \sum_{j=1}^{N} f_j(q) e^{i\vec{q}\vec{r}_j}, \tag{3.63}$$

sendo N o número de átomos e $f_j(q)$ os fatores de espalhamento atômicos, e a expressão pode ser expandida em multipolos a partir de amplitudes parciais $A_{lm}(q)$ e harmônicos esféricos $Y_{lm}(\omega)$ (Stuhrmann, 1970) por meio de

$$A_a(\vec{q}) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} A_{lm}(q) Y_{lm}(\omega), \tag{3.64}$$

com as amplitudes parciais dadas por

$$A_{lm}(q) = 4\pi i^l \sum_{j=1}^{N} f_j(q) j_l(qr_j) Y^*_{lm}(\omega_j), \tag{3.65}$$

sendo j_l referente às funções esféricas de Bessel. Para as amplitudes referentes ao volume excluído, $A_e(\vec{q})$, o processo é análogo ao apresentado acima, com as amplitudes parciais dadas por $E_{lm}(q)$ (análogas a $A_{lm}(q)$), sendo dadas em função dos fatores de espalhamento $g_j(q)$, de modo que

$$E_{lm}(q) = 4\pi i^l \sum_{j=1}^{N} g_j(q) j_l(qr_j) Y^*_{lm}(\omega_j). \tag{3.66}$$

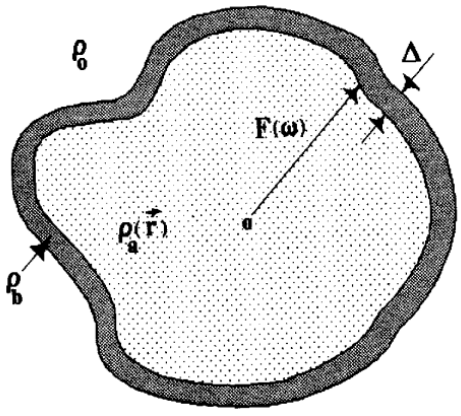

Figura 3.13: Modelo de uma proteína em solução, com densidades eletrônicas $\rho_p(Vecr)$ no interior da molécula, ρ_0 no solvente e ρ_b na camada de hidratação de espessura Δ. O formato da proteína é descrito pela função envelope $F(\omega)$ (Svergun et al., 1995).

Para descrever a densidade eletrônica da borda é necessária a chamada *função envelope* $F(\omega)$, sendo definida formalmente por

$$\rho_b(\vec{r}) = \begin{cases} 1, & \text{para } F(\omega) \leqslant r \leqslant F(\omega) + \Delta \\ 0, & \text{para } 0 < r < F(\omega) \text{ ou } r > F(\omega) + \Delta \end{cases}, \tag{3.67}$$

lembrando que ω se refere aos ângulos de orientação (θ e ϕ). Em outras palavras, a densidade da borda é simplesmente definida como unitária na borda e nula fora dela, e na expressão da intensidade a amplitude é multiplicada pela diferença entre essa densidade e a do solvente livre. A Figura 3.13 resume o modelo usado pelo CRYSOL para descrição da partícula. A partir de transformadas de Hankel, que expandem uma dada função em funções de Bessel (Glatter e Kratky, 1982), as amplitudes parciais $B_{lm}(q)$ são dadas por

$$B_{lm}(q) = i^l \sqrt{\frac{2}{\pi}} \int_0^\infty \rho_{lm}(r) j_l(qr) r^2 dr, \tag{3.68}$$

com

$$\rho_{lm}(r) = \int_\omega \rho_b(\vec{r}) Y_{lm}^*(\omega) d\omega. \tag{3.69}$$

Também pode-se obter $B_{lm}(q)$ por meio de

$$B_{lm}(q) = i^l \sqrt{\frac{2}{\pi}} \int_\omega Y_{lm}^*(\omega) d\omega \int_{F(\omega)}^{F(\omega)+\Delta} j_l(qr) r^2 dr. \tag{3.70}$$

Portanto, após simplificações devido à média para todas as orientações, as intensidades podem ser decompostas em coeficientes até um valor máximo $l = L$ que determina a resolução da curva calculada de acordo com

$$I(q) = \sum_{l=0}^{L} \sum_{m=-l}^{l} |A_{lm}(q) - \rho_0 E_{lm}(q) + \delta\rho B_{lm}(q)|^2. \tag{3.71}$$

Os fatores de forma usados para o cálculo de $A_{lm}(q)$ e $C_{lm}(q)$, uma vez que em geral arquivos PDB não mostram explicitamente os átomos de H, podem se referir tanto a átomos como a grupos atômicos, como NH ou CH, e são calculados de acordo com

$$g_j(q) = G(q) V_j \, \exp\big(-\pi q^2 V_j^{2/3}\big), \tag{3.72}$$

com $V_j = 4\pi/3 r_{wj}^3$ o volume de solvente (com moléculas de raio r_{wj}) deslocado pelo j-ésimo átomo ou grupo atômico (de raio r_{gj}) e $G(q)$ um fator de expansão tal que

$$G(q) = \left(\frac{r_0}{r_m}\right)^3 \exp\left[-\left(\frac{4\pi}{3}\right)^{3/2} \pi q^2 \left(r_0^2 - r_m^2\right)\right], \tag{3.73}$$

sendo $r_m = 1/N \sum_{j=1}^{N} r_{gj}$ o raio médio do átomo ou grupo atômico, e r_0 um parâmetro variável, chamado raio atômico efetivo, que ajusta o volume total excluído. Este é o primeiro parâmetro de ajuste do programa, enquanto o segundo é o $\delta\rho$, ou seja, a diferença de densidade eletrônica entre a borda e o solvente livre. Com isso obtém-se os coeficientes $A_{lm}(q)$ e $E_{lm}(q)$.

Para o cálculo de $B_{lm}(q)$ é necessário encontrar $F(\omega)$, e a ideia geral do processo consiste em atualizar a função utilizando cada átomo ou grupo atômico para cada direção ω_i caso este seja o elemento mais externo. Formalmente $F(\omega_i)$, a cada passo angular ω_i contendo um certo número de átomos (indicados por j), é dada por

$$F(\omega_i) = \max\left\{F'(\omega_i), (r_{ji} + 0.5 r_{gj})\right\},$$

sendo $F'(\omega_i)$ o valor atual da função envelope, r_{ji} a projeção na direção i da distância do centro geométrico da macromolécula ao k-ésimo átomo contido no passo angular ω_i, e r_{gj} é o raio atômico. O processo é repetido para todos os átomos em ω_i e, depois, para cada ω_i, de modo que ao final do processo $F(\omega)$ contenha todas as distâncias entre o centro geométrico da proteína e a superfície da mesma. A partir disso os termos $B_{lm}(q)$ podem ser encontrados e, portanto, também a curva de espalhamento teórica para a proteína de estudo.

Uma vez que o modelo utilizado para a borda é simples, estruturas mais complexas, como por exemplo estruturas ocas (que apresentariam mais de uma solução para $F(\omega)$ o CRYSOL modela apenas a camada externa, ignorando características estruturais possivelmente importantes da proteína.

Existem outros programas com abordagens diferentes que podem ser utilizados para o cálculo de curvas teóricas de SAXS, como o FoXS (Schneidman-Duhovny et al., 2013), de desempenho similar ao do CRYSOL porém com tempo de processamento menor, sendo interessante para a análise de conjuntos muito grandes de dados ou para implementação em pipelines de análise.

3.4.4 *Modelagem ab initio: de geração de modelos com DAMMIF a filtragem e questões de ambiguidade*

Uma das aplicações mais populares de SAXS é a obtenção do envelope tridimensional da proteína de estudo por meio de metodologias *ab initio*, como é o caso do programa DAMMIF (Franke e Svergun, 2009), implementação mais rápida do programa DAMMIN (Svergun, 1999). O programa não exige nenhum conhecimento *a priori* sobre a proteína e usa como entrada os parâmetros obtidos pela análise de Guinier e dados da $p(r)$ e parte da aproximação de que a densidade eletrônica da proteína é homogênea em todo o seu volume. O programa aceita restrições impostas pelo usuário ao modelo gerado, como imposições relativas à simetria e anisotropia (corpo oblato vs corpo prolato) da partícula gerada. Ao final do programa é possível utilizar o modelo gerado como entrada para uma execução do programa DAMMIN, de cálculo mais lento e custoso que o DAMMIF, para o refino da estrutura obtida.

O algoritmo do DAMMIF funciona da seguinte maneira: é gerado um espaço de busca esférico de raio $R = D_{max}/2$ densamente ocupado por esferas, denominadas *dummy atoms* (inglês para *átomos-modelo*), que representam elementos de densidade eletrônica fixa e raio $r_0 << R$, sendo atribuídos dois valores de densidade a cada esfera: um para esferas pertencentes à proteína e outro para pertencentes ao solvente. Cada valor é atribuído a um vetor de configuração X com M entradas binárias tais que $M \approx (R/r_0)^3$ representando cada esfera no espaço. O programa calcula a intensidade de espalhamento do modelo de esferas descrito por esse vetor a partir da equação

$$I(q) = 2\pi^2 \sum_{l=0}^{\infty} \sum_{m=-l}^{l} |A_{lm}(q)|^2, \tag{3.74}$$

sendo $A_{lm}(q)$ as amplitudes parciais. Dado um vetor posição em coordenadas polares $\vec{r_j} = r_j\hat{r} + \omega_j\hat{\theta}$, essas amplitudes podem ser expandidas em funções de Bessel $j_l(qr)$ e harmônicos esféricos $Y_{lm}(\omega)$ de acordo com

$$A_{lm}(q) = i^l\sqrt{2\pi}\nu_a \sum_{j=1}^{M} j_l(qr_j)Y^*_{lm}(\omega_j), \tag{3.75}$$

na qual $\nu_a = \left(\frac{4\pi}{3}r_0\right)/0.74$ é o volume deslocado por cada esfera. Iterativamente o DAMMIF varia a disposição das esferas até que a função gerada pelo conjunto delas aproxime-se

da curva experimental até um certo limite.

O programa busca minimizar a função $f(X) = R^2(I, X) + \sum_k \alpha_k P_k(X)$, denominada função-alvo (*target function*), em que X é um vetor que denomina uma conformação gerada pelo programa, sendo a primeira parcela a discrepância entre os dados experimentais e o modelo gerado, e a segunda a soma das penalidades aplicadas à geração do modelo, sendo elas relativas à compactação da estrutura, à interconectividade entre as esferas e à distância do centro de massa do modelo em relação à origem (auxiliada pelo R_g calculado previamente), além da possível restrição imposta pelo usuário a respeito da anisotropia da partícula. O fator $\alpha_k > 0$ se refere ao peso dado a cada uma das penalidades e deve permitir que as penalidades tenham contribuição de $10\% - 50\%$ à função $f(X)$ no final da minimização a fim de que as penalidades possam filtrar modelos com "peso de decisão" comparável ao do termo de discrepância.

O processo de minimização segue o método de *simulated annealing* (expressão em inglês para "recozimento simulado") (Kirkpatrick e Vecchi, 1983), cuja ideia principal consiste em realizar alterações aleatórias no sistema, no caso o nosso vetor de configuração X, e armazenando os que vão diminuindo a energia, representada pela função $f(X)$, mas também permitindo alguns aumentos a fim de se evitar mínimos locais sutis utilizando um parâmetro de "temperatura" T que diminui conforme mais iterações são feitas. O algoritmo segue a ideia de que a cada passo uma esfera seja alterada aleatoriamente de posição no espaço de busca, indo de uma configuração X para uma X' e é calculada a diferença $\Delta = f(X') - f(X)$. Caso $\Delta < 0$ o estado X' é aceito, e caso $\Delta > 0$ o estado X' é aceito com probabilidade $P = \exp(-\Delta/T)$. Este processo é feito um certo número de vezes ($100M$ iterações ou $10M$ substituições feitas com sucesso) e após isso T é diminuído (por exemplo, em 10%), de modo que a probabilidade de se aceitar um modelo X' tal que $\Delta > 0$ diminui exponencialmente com a diminuição da temperatura. O algoritmo se inicia a uma temperatura T_0 relativamente alta, como $T_0 = f(X_0)$ e segue até que $f(X) = f(X_f)$ seja suficientemente estável, selecionando a configuração X_f.

Não há um critério amplamente aceito sobre qual o intervalo de valores de q para o qual os métodos de obtenção de envelope devem ser aplicados a fim de que seja utilizado um q_{max} que permita a descrição de elementos essenciais da estrutura, como formato geral e buracos, mas que não chegue à região de ângulos mais altos, que não só contém menor razão sinal/ruído mas também se referem a detalhes da estrutura que estão fora da resolução

permitida pela técnica de SAXS (Petoukhov et al., 2012). Trabalhos de modelagem em geral sugerem a utilização de um intervalo de dados do primeiro ponto (desde que não haja agregação) até $q_{max} = 8/R_g$, com alguns trabalhos fazendo uso de $q_{max} = 7/R_g$. Ou seja, o intervalo utilizado passa a depender do tamanho da proteína de estudo.

Uma vez que se trata de uma reconstrução tridimensional a partir de uma curva de intensidades unidimensional, em que perdemos a informação de fase, há o problema de mais de um envelope corresponder à curva de espalhamento. Para contornar este problema o programa pode gerar múltiplas reconstruções (recomenda-se um mínimo de dez estruturas), alinhá-las usando o programa SUPCOMB e calcular um "modelo médio" representativo do conjunto, seja utilizando o conjunto de modelos gerados fazendo o descarte dos muito discrepantes dos outros (pelo DAMAVER) quanto buscando subconjuntos de formas similares entre si ao executar o programa mais vezes, as separando em *clusters* (ou aglomerados, pelo DATCLUST) e calculando a média de cada subconjunto.

Também há outros métodos de reconstrução do envelope, como o implementado no programa GASBOR (Svergun et al., 2001), que faz uso de elementos de volume, denominados *dummy residues*, com conteúdo físico mais significativo do que as esferas utilizadas no DAMMIF, atribuindo espaçamentos de acordo com a estrutura de aminoácidos e inserindo uma camada de hidratação. Também há outros programas, como o recente SHAPES (Badger, 2019), que atribui um potencial entre as esferas para que haja um elemento de interação (inserindo, portanto, um elemento de minimização de estrutura) e utiliza a $p(r)$ como alvo para o ajuste, e não os dados experimentais, obtendo buracos bem definidos e estruturas tipo halter com espaçamentos mais bem definidos, além de maior convergência estrutural com relação aos programas anteriores.

O programa SUPCOMB é responsável pelo alinhamento de arquivos no formato *.pdb*, seja em alta ou baixa resolução, pela minimização de uma medida denominada *Normalized Spacial Discrepancy* (NSD), dada por $\rho(S_1, S_2)$. A ideia é de que cada estrutura a ser alinhada é descrita por um conjunto de dados S_k contendo N_k pontos, chamados s_{ki}. Ou seja, tem-se respectivamente $S_1 = \{s_{1i}, i = 1, 2, ..., N_1\}$ e $S_2 = \{s_{2i}, i = 1, 2, ..., N_2\}$, sendo o valor mínimo entre todas as distâncias calculadas entre um ponto específico do conjunto S_1 e cada um dos pontos de S_2 chamada $\rho(s_{1i}, S_2)$. Analogamente se define $\rho(s_{2i}, S_1)$. A NSD é dada por uma média ponderada entre as distâncias entre os pontos por meio de

$$\rho(S_1, S_2) = \left\{ \frac{1}{2} \left[\frac{1}{N_1 d_2^2} \sum_{i=1}^{N_1} \rho^2(s_{1i}, S_2) + \frac{1}{N_2 d_1^2} \sum_{i=1}^{N_2} \rho^2(s_{2i}, S_1) \right] \right\}, \tag{3.76}$$

sendo d_k uma quantidade chamada $fineness$, definida pela distância média entre pontos vizinhos no conjunto S_k. Deste modo, para duas estruturas sobrepostas o NSD tende a zero. Na prática, entende-se que pares de estruturas com $NSD < 0.8$ indicam estabilidade entre as estruturas geradas, enquanto valores bem maiores que 1 indicam maior grau de variabilidade nos modelos obtidos (Petoukhov & Svergun, 2015). O programa busca minimizar a NSD por meio do cálculo dos tensores de inércia de cada conjunto e aplicando diferentes rotações e deslocamento buscando o alinhamento dos eixos principais de ambos os conjuntos. Na prática, ao usar o programa o usuário escolhe uma estrutura de *template* e outra a ser transladada e rotacionada para que ambas sejam alinhadas e tenham a NSD calculada.

O DAMAVER funciona da seguinte maneira (Volkov e Svergun, 2003; Kozin e Svergun, 2001): faz-se o alinhamento entre si para cada par de estruturas geradas, calculando a NSD para cada par, e se calcula uma NSD média ($\langle NSD \rangle$) com a respectiva medida de dispersão ($\Delta(NSD)$). Calcula-se então a NSD_k, que é a média das NSD calculados para cada reconstrução ao alinhá-la com todas as outras e se seleciona a reconstrução com menor valor de NSD_k. Este processo é seguido pela filtragem de estruturas discrepantes ao remover reconstruções tais que $NSD_k > \langle NSD \rangle + 2\Delta(NSD)$. Os modelos restantes são sobrepostos e a cada ponto da grade é atribuído um "fator de ocupação" definido por quantas esferas se encontram na vizinhança daquele ponto. Os pontos com mais esferas na vizinhança são selecionados de modo que o volume excluído pela estrutura final seja igual à média dos volumes excluídos por cada reconstrução, e toma essa seleção de pontos como sendo estrutura "média" do conjunto. Ao conjunto de pontos com ocupação não-nula se dá o nome de "região de *spread* total", e ao conjunto com maiores valores de ocupação selecionados se denomina "volume mais populado". Na Figura 3.14 estão ilustradas várias estruturas de esferas, com suas respectivas sobreposições e mapa de população.

Importante notar que o espalhamento calculado por essa estrutura média não necessariamente ajusta os dados da forma como cada estrutura gerada o faz individualmente, mas visualmente o modelo preserva as características comuns às reconstruções e remove elementos particulares de cada uma. Para que se tenha uma estrutura que ajuste os dados é comum gerar um modelo utilizando o DAMMIN (programa de reconstrução tridimensional

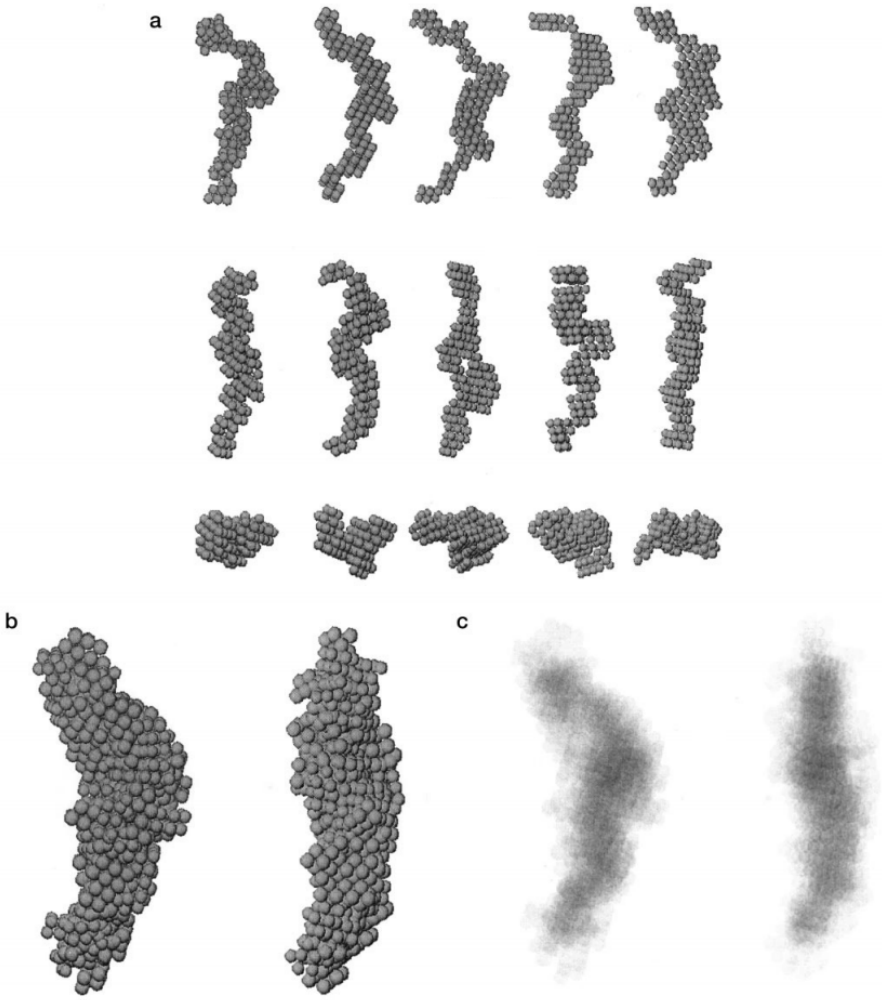

Figura 3.14: Reconstruções tridimensionais *ab initio* para uma proteína em solução (**a**), região de *spread* total (**b**) e mapa de população da região de *spread* total (Funari et al., 2000).

anterior ao DAMMIF, porém mais lento e com algumas sutis diferenças), que permite definir o volume inicial de busca com esferas, de modo que seja feito um "refino" do resultado do DAMAVER, obtendo um modelo final.

Ao se fazer reconstituições tridimensionais para certos tipos de partícula pode-se observar problemas nas modelagens *ab initio* refletidos na dificuldade do método em gerar reconstruções para curvas simuladas de certos tipos de sólidos satisfatoriamente, mostrando uma disparidade muito grande entre a forma teórica calculada pela fórmula de Debye e a obtida pelos programas, podendo ser fonte de falsos positivos na obtenção de modelos. Essa ambiguidade de determinação de estruturas pode ser avaliada pelo programa AMBIMETER.

O programa contém uma biblioteca de curvas de espalhamento de um conjunto de estruturas de diferentes topologias geradas pelo cálculo de curva teórica de estruturas interco-

nectadas contendo até 7 esferas dispostas tridimensionalmente em uma grade hexagonal de maneiras diferentes (gerando mais de 14 mil combinações). Exemplos de topologias estão presentes na Figura 3.15. O que se observa é que, representando as curvas em gráficos de qR_g vs $I(q)/I(0)$ todas elas coincidem na região de Guinier, o que é esperado, e que a partir deste ponto algumas regiões do gráfico são mais populadas que outras, havendo, portanto, topologias com maior grau de ambiguidade. Numa tentativa de quantificação dessa ambiguidade é introduzida uma "medida de dissimilaridade" d, que é basicamente a diferença entre os gráficos de Kratky adimensionais para o conjunto de dados e uma curva da biblioteca. Essa medida d é definida por (Petoukhov e Svergun, 2015)

$$d^2 = \frac{1}{N}\sum_{i=1}^{N}\left\{(q_iR_g)^2\left[\frac{I_{\exp}(q_iRg)}{I_0} - I_{top}(q_iRg)\right]\right\}^2, \tag{3.77}$$

sendo N o número de pontos experimentais, $I_{\exp}(q_iR_g)$ a intensidade da curva experimental e $I_{top}(q_iR_g)$ a intensidade da curva teórica da respectiva topologia da biblioteca gerada previamente. É escolhido um limite de $d^2 < 0.0016$ (Petoukhov e Svergun, 2015) para identificar uma curva como sendo condizente com a respectiva topologia da biblioteca, e isso é repetido para cada uma das 14 mil estruturas. O número de topologias m condizentes com os dados é usado para o cálculo de um *a-score* definido por $a - score = log(m)$, como medida de ambiguidade. Portanto, o uso dessa quantidade permite a predição da unicidade de modelos gerados por métodos *ab initio* a partir do conjunto de dados, sendo atribuído $a - score < 1.5$ como forte indicativo de unicidade nas estruturas obtidas, enquanto valores mais altos, em especial maiores que 2.5, indicam uso da divisão em clusters em casos mais brandos a imposição de simetrias e anisotropias em casos mais extremos. O trabalho original deste programa (Petoukhov e Svergun, 2015) mostra que partículas anisotrópicas são as que apresentam mais ambiguidade (com exceção de partículas muito elongadas), e a categoria de partículas de maior ambiguidade na obtenção de envelopes de SAXS é a de anisotropia oblata.

3.4.5 Modelagem híbrida de corpo rígido com métodos ab initio: BUNCH

Uma abordagem possível além da metodologia *ab initio* é a modelagem de corpo rígido. Para proteínas que têm estruturas parciais disponíveis em alta resolução é possível calcular o espalhamento teórico de subunidades e, a partir dessa informação e do número de subuni-

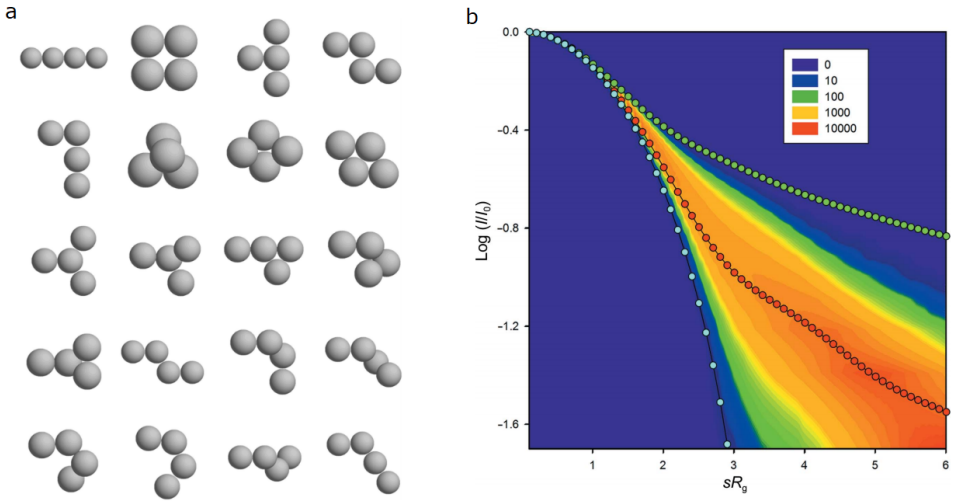

Figura 3.15: Exemplos de modelos de esferas utilizados para gerar curvas de SAXS (**a**) e gráfico de temperatura para indicar regiões de maior ou menor incidência de curvas simuladas de SAXS com modelos de esferas (Petoukhov e Svergun, 2015).

dades e da sequência de aminoácidos, realizar uma modelagem considerando as estruturas PDB como entes rígidos, buscando reorientá-las de modo que o espalhamento teórico da estrutura final ajuste os dados.

Há um leque de programas publicados em um mesmo trabalho (Petoukhov e Svergun, 2005) que realizam esta tarefa de diferentes maneiras e com diferentes escopos, podendo ser específico para oligômeros simétricos, por exemplo, como o GLOBSYMM, ou para dímeros, como o DIMFOM. Também há outros, de aplicação mais abrangente, como o SASREF. Porém é comum encontrar estruturas que não foram completamente resolvidas, o que pode levar inviabilizar a análise por corpo rígido. Por conta disso há programas que realizam uma modelagem híbrida, aplicando método de corpo rígido para os arquivos PDB de entrada e também modelagem *ab initio* para inserir regiões ausentes na estrutura gerada, desde *loops* (inglês para laços) a domínios inteiros. As intensidades, portanto, ao serem expandidas em amplitudes parciais, podem ser escritas em termos das amplitudes relativas por meio de (Petoukhov e Svergun, 2005)

$$I(q) = 2\pi^2 \sum_{l=0}^{L} \sum_{m=-l}^{l} \left| \sum_i A_{lm}^i(q) + \sum_j D_{lm}^j(q) \right|^2 , \tag{3.78}$$

sendo $A_{lm}(q)$ relativas a cada uma das i estruturas PDB, e $D_{lm}(q)$ relativas a cada uma das j esferas do modelo *ab initio*. O programa busca minimizar a função-alvo E dada por

$$E = \Sigma(\chi^2)_k + \alpha_{sob}P_{sob} + \alpha_{lig}P_{lig} + \alpha_{die}P_{die} + \alpha_{est}P_{est}, \tag{3.79}$$

sendo α_l referentes aos pesos dados a cada penalidade, especificada nos respectivos termos P_l. Os termos de índice *sob* se referem a impedimento estérico entre os PDBs e as esferas modeladas (impedindo que ambas se sobreponham), *lig* e *die* se referem a ângulos de ligação e diedrais para as esferas, e *est* penaliza conformações muito estendidas para os *loops*. A penalização para *loops* de conformação muito estendida se dá por meio de restrições ao R_g da região por meio de

$$P_{est} = \frac{\sum_k \left(max\left(0, R_g^k - 3\sqrt[3]{M_k}\right)\right)^2}{9\sum_k \sqrt[3]{M_k^2}}, \tag{3.80}$$

sendo R_g^k o raio de giração do k-ésimo *loop* modelado contendo M_k esferas. O valor $3\sqrt[3]{M_k^2}$ é uma aproximação para o R_g correspondente a uma proteína globular com M_k resíduos (Petoukhov e Svergun, 2005).

Outro método que realiza modelagem similar é o CORAL (Petoukhov et al., 2012), que para a modelagem utiliza uma biblioteca pré-construída de *loops* com diferentes tamanhos e conformações para reduzir o tempo computacional da modelagem com esferas, sendo portanto uma modelagem de corpo rígido, e não um método híbrido.

3.4.6 Modelagem híbrida considerando flexibilidade estrutural: EOM

Em solução, proteínas podem apresentar diferentes conformações possíveis, e o espalhamento medido corresponde a uma média de todas as N diferentes configurações possíveis, incluindo diferentes estados oligoméricos, com intensidade dada por

$$I(q) = \sum_{i=1}^{N} \nu_i I_i(q), \tag{3.81}$$

sendo $I_i(q)$ a intensidade de espalhamento devida à i-ésima conformação e ν_i a fração de volume ocupada por ela (de modo que $\sum_{i=1}^{N} \nu_i = 1$). Um problema com relação aos métodos de modelagem apresentados até aqui é que nenhum deles consegue endereçar a questão da flexibilidade indo além da obtenção de uma estrutura média. O fato do SAXS permitir a obtenção de estruturas que correspondam à uma média entre todas as conformações presentes em solução já é de grande auxílio quando a cristalização de proteínas não é possível, porém uma vez que proteínas intrinsecamente desordenadas (IDPs) são

altamente abundantes em qualquer proteoma (Uversky, 2013), a estrutura média obtida (tanto por envelope quanto por modelagem de corpo rígido) muitas vezes não apresenta informação relevante. Inclusive, para proteínas com domínios bem enovelados ligados por *linkers* flexíveis pode-se concluir erroneamente que a estrutura apresenta geometria elongada, quando na verdade esta apresenta flexibilidade conformacional. A depender do tamanho dos domínios em relação aos *linkers* o gráfico de Kratky pode apontar para uma estrutura altamente globular quando não se trata deste caso (Bernadó, 2010).

Considerando apenas polidispersão conformacional, sem variação de estado oligomérico, é possível fazer uso de métodos de Monte Carlo para entender melhor como se dá o equilíbrio dessas diferentes estruturas em solução. O programa *Ensemble Optimization Method* (inglês para *Método de Otimização de Ensemble*, EOM) endereça o problema da polidispersão gerando um conjunto contendo um número suficientemente alto de diferentes conformações (*pool*) (Levitt, 1976; Bernadó et al., 2005) a fim de descrever o espaço conformacional acessível à proteína, calculando as intensidades para cada elemento da *pool*, e por fim usando um algoritmo genético para filtrar as estruturas cuja intensidade média minimize o χ^2 com relação aos dados experimentais (*ensemble*). O EOM utiliza, portanto, uma abordagem de "força bruta" (Bernadó et al., 2007). A rotina de geração de *pool* é chamada *RANdom CHain* (inglês para *Cadeia Aleatória*, RANCH), e o algoritmo genético de seleção de *ensemble* é o *Genetic Algorithm Judging Optimisation of Ensembles* (inglês para *Algoritmo Genético para Julgamento de Otimização de Ensembles*, GAJOE), com o EOM consistindo na implementação de ambos, sendo também possível utilizar cada programa separadamente (Tria et al., 2015).

No EOM tanto é possível buscar modelar proteínas inteiras como desordenadas como também realizar a modelagem de algumas regiões, utilizando arquivos PDB para regiões rígidas, como domínios globulares, modelando apenas linkers ou His-tags, por exemplo. Aqui é importante reforçar as diferenças entre o EOM e o BUNCH neste último caso para além do número de estruturas geradas: enquanto o BUNCH gera e altera uma única estrutura de modo que a intensidade gerada por ela se ajusta aos dados, no EOM as estruturas são geradas por força bruta independente de ajuste aos dados e, ao final do processo, se busca quais combinações de estruturas resultam em uma intensidade que se ajusta aos dados experimentais.

Na modelagem de regiões desordenadas o programa pode gerar estruturas mais ou

menos aleatórias de acordo com a especificação do usuário, e isso é feito em cima da relação de Flory (Flory, 1953), que relaciona o raio de giração de um polímero (R_g) a parâmetros característicos, dada por

$$R_g = R_0 N^{\nu}, \tag{3.82}$$

sendo R_0 uma constante relacionada ao comprimento de persistência da cadeia, N o número de meros da cadeia e ν um fator empírico característico do tipo de cadeia (como proteínas intrinsecamente desordenadas, proteínas desnaturadas etc), sendo atribuídos diferentes valores para R_0 e ν de acordo com o tipo de cadeia a ser modelada especificada pelo usuário. É dito que para proteínas globulares $\nu \approx 0.33$, enquanto para altamente desordenadas $\nu \approx 0.60$ (Riback et al., 2017).

O programa leva em conta diagramas análogos aos de Ramachandran (Nelson et al., 2008; Ramakrishnan e Ramachandran, 1965), mas que fazem uso apenas da disposição espacial de carbonos-α (em um gráfico de Ramachandran adaptado para ângulos entre C_α de consecutivos na cadeia) para que não sejam gerados *loops* com conformações de pouco significado físico (Kleywegt, 1997).

Para permitir parâmetros de comparação entre as estruturas da *pool* e as selecionadas pelo algoritmo genético que vão para além da inspeção visual, o EOM calcula os R_g e D_{max} de cada estrutura e produz histogramas para cada parâmetro (Figura 3.16), comparando cada um dos conjuntos. Em geral distribuições da *pool* são mais amplas em torno de um valor mais proeminente (podendo estar na mediana ou não a depender da proteína modelada), enquanto o ensemble selecionado pode ter distribuições mais ou menos amplas. Para proteínas com conformações mais bem determinadas e rígidas são observados picos bem determinados, com distribuições mais estreitas, enquanto para proteínas com um espaço conformacional muito amplo as distribuições se assemelham à da *pool*, como ilustrado na Figura 3.17. Recomenda-se executar o algoritmo genético múltiplas vezes para a elaboração do histograma do *ensemble* de estruturas selecionadas.

O EOM também busca traduzir o grau de flexibilidade da proteína de estudo por meio de parâmetros calculados a partir das distribuições de R_g e D_{max}: R_{flex} e R_σ. Para calculá-los primeiro deve-se representar os histogramas em funções de densidade de probabilidade $D = (X, P)$ com M entradas, sendo $P = (p_1, p_2, ..., p_M)$ as probabilidades associada aos intervalos $X = \{x_1, x_2, ..., x_M\}$, de modo que $\sum_{j=1}^{M} p(x_j) = 1$. Neste contexto o conceito

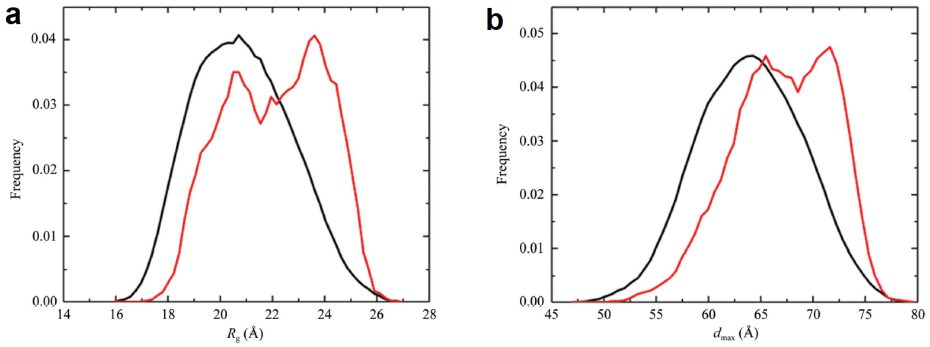

Figura 3.16: Distribuições de R_g (**a**) e D_{max} (**b**) geradas pelo programa EOM. Em preto são representadas as distribuições relativas ao *pool* de conformações totais e em vermelho as distribuições de *ensembles* selecionados (Trewhella et al., 2017).

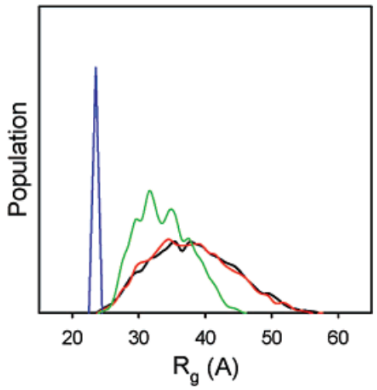

Figura 3.17: Exemplo de distribuições de R_g para *ensembles* de uma proteína bem enovelada (azul), parcialmente enovelada (verde) e totalmente desenovelada (vermelho) sobrepostos ao *pool* de conformações geradas pelo programa (preto) (Bernadó et al., 2007).

de entropia de Shannon ($H_b(S)$) torna-se bastante útil, sendo dado por (Tria et al., 2015)

$$H_b(S) = -\sum_{j=1}^{M} p(x_j)\ \log_b\left[p(x_j)\right], \tag{3.83}$$

fazendo que $\log_b(0) = 0$ para $p(x_j) = 0$. O EOM é implementado de modo que $b = \left|X\right|^j$ é representante da "cardinalidade do alfabeto" (Shannon e Weaver, 1949; Tria et al., 2015), com $H_b(S) \in [-1, 0]$, de modo que $H_b(S) \to -1$ para distribuições com maior largura (com $H_b(S) = -1$ para uma distribuição uniforme), e $H_b(S) \to 0$ para baixa incerteza (sendo $H_b(S) = 0$ para distribuições de um único valor).

Assim, para proteínas com distribuições de *ensemble* similares às de *pool*, ou seja, proteínas extremamente desordenadas, espera-se que $H_b(S)$ com valores mais próximos a -1, enquanto para proteínas de conformação muito bem definida, ou seja, com distribuições bem mais estreitas de *ensemble* com relação às de *pool* espera-se que $H_b(S)$ seja mais

próximo a 0. A entropia de Shannon é adaptada em um parâmetro muito similar, o R_{flex}, dado por

$$R_{flex} = -H_b(S), \tag{3.84}$$

podendo ser também multiplicado por 100 para indicar, em uma escala de 0 a 100% o grau de desordem de uma cadeia polipeptídica, sendo 0% referente a uma cadeia perfeitamente rígida e 100% a uma totalmente desordenada.

Pode-se introduzir também um parâmetro que faça referência às variâncias das distribuições, chamado R_σ, dado por

$$R_\sigma = \frac{\sigma_E}{\sigma_P}, \tag{3.85}$$

sendo σ_E a variância do *ensemble* e σ_P, da *pool*. Espera-se $R_\sigma \to 1$ quando o *ensemble* descreve uma proteína muito flexível, apresentando variâncias similares para os conjuntos. Para proteínas mais ordenadas espera-se $R_\sigma < 1$. O estudo de ambos os parâmetros é possível descrever mais seguramente o estado de desordem de uma proteína de interesse, uma vez que dados problemáticos (por exemplo, contendo agregação) ou análises descuidadas podem dar resultados como $R_{flex} < 1$ significativamente e $R_\sigma > 1$. Também são introduzidos parâmetros auxiliares, como desvio padrão, de forma a descrever as distribuições da forma mais quantitativa possível.

O uso das distribuições de R_g e D_{max} geradas pelo EOM podem servir como ferramenta de distinção quando em dúvida se uma estrutura é elongada ou flexível, tornando esta metodologia uma ferramenta poderosa de análise (Bernadó, 2010), assim como RgD (seção 3.3.4). Inclusive o uso da entropia obtida pelo uso do RgD pode ser útil como informação adicional para a geração de modelos pelo EOM. Porém é importante reforçar que os modelos filtrados pelo algoritmo genético não representam estruturas rígidas, como as N conformações rígidas que a proteína pode ter em solução, mas sim que devem ser entendidos como representações aproximadas das conformações predominantes na amostra.

Há programas que buscam endereçar questões relativas a polidispersão utilizando SAXS, utilizando abordagens um pouco diferentes, como é o caso do *Minimal Ensemble Search* (MES), que faz uso de dinâmica molecular (Pelikan et al., 2009). Para o caso de polidispersão de estados oligoméricos é possível utilizar um software de ajuste de curvas obtidas

pelo CRYSOL de estruturas em formato PDB, chamado OLIGOMER, mas recentes implementações do EOM também incluem esta funcionalidade (Tria et al., 2015). Métodos de refinamento de *ensemble* não apenas em SAXS mas também em FRET (Förster Resonance Energy Transfer) são especialmente úteis para o estudo de IDPs. FRET permite a obtenção de R_g a partir da distância entre dois fluoróforos, porém até recentemente era apenas mediante hipóteses de modelo para a cadeia, e metodologias de *ensemble* permitem contornar esta questão, permitindo resultados mais condizentes entre ambas as técnicas (Fuertes et al., 2017; Best et al., 2018; Fuertes et al., 2018), apesar de ainda haver discussões sobre a possibilidade de eventuais diferenças entre resultados fazendo uso de cada técnica serem devido a metodologias de análise ou de colapso estrutural devido à presença de fluoróforos (Riback et al., 2017; Riback et al., 2018; Riback et al., 2019; Best, 2020).

Análises a partir de modelos poliméricos também são possíveis porém menos utilizados, como ajustes com a lei de Debye, válida para proteínas com $\nu \approx 0.5$. Recentemente foi elaborada uma metodologia de análise, especialmente útil para IDPs, que consiste no trabalho com um termo a mais na lei de Guinier, proporcional a q^4, e em uma modificação na lei de Flory, permitindo a extensão da região de análise para $qR_g < 2.0$ e a obtenção do parâmetro ν (Zheng e Best, 2018). Porém o trabalho com métodos computacionais com *ensembles* está se tornando mais popular devido a computadores com poder de processamento cada vez maior e amplo leque de aplicabilidade.

3.4.7 Definindo o intervalo de dados útil para análise usando canais de Shannon: SHANUM

Uma questão importante em experimentos de SAXS é a escolha de um intervalo de pontos para modelagem uma vez que há as questões relacionadas à razão sinal-ruído, especialmente para valores mais altos de q, e ao excesso de pontos experimentais medidos. Portanto, perguntas relacionadas a até qual valor de q há informações úteis na curva de SAXS são bastante relevantes. Há regras qualitativas de se buscar valores como $q \geqslant 3.5\text{nm}^{-1}$, $q \approx 8/R_g$, ou regiões cuja razão sinal-ruído seja (subjetivamente) satisfatória porém não há critérios amplamente aceitos e rigorosamente elaborados quanto a isso. Em cima disso, pode-se elaborar um uma metodologia baseada nas ideias de amostragem de Shannon para determinar até que valor de q há informações úteis para modelagem (Konarev e Svergun, 2015).

A função $p(r)$ pode ser escrita em termos de uma somatória de acordo com

$$p(r) = \frac{1}{2\pi^2} \sum_{n=1}^{\infty} a_n sen\left(\frac{\pi n r}{D_{max}}\right), \tag{3.86}$$

de modo que a inserção dessa $p(r)$ resulta na fórmula de interpolação de Shannon, dada por, utilizando uma função auxiliar $U(q) = qI(q)$,

$$U(q) = \sum_{n=1}^{\infty} q_n a_n \left[\frac{\text{sen}(D_{max}(q - q_n))}{D_{max}(q - q_n)} - \frac{\text{sen}(D_{max}(q + q_n))}{D_{max}(q + q_n)}\right]. \tag{3.87}$$

Para medidas de SAXS tem-se um conjunto de dados tomados para valores finitos de q. Assim, chamemos $p_M(r)$ e $U_M(r)$ as funções truncadas para um valor $n = M$, e utilizando esses M canais de Shannon pode-se aproximar o conjunto de dados ao minimizar o χ^2. Chamando este de $\chi^2(M)$, tem-se que, para um conjunto contendo N pontos experimentais e M canais de Shannon,

$$\chi^2(M) = \sum_{i=1}^{N} \frac{1}{2q_i^2\sigma_i^2}[q_i I(q_i) - U_M(q_i)]^2, \tag{3.88}$$

de modo que haja um equilíbrio entre o número de canais M usados no ajuste para que sejam suficientes para a descrição dos dados porém que não haja *over-fitting* em que o uso de mais valores de a_n não melhorem o χ^2. Assim, busca-se um M' tal que haja esse equilíbrio tal que a informação da curva seja preservada, e este valor ótimo está relacionado com o q_{max} que ainda contém informação útil para modelagem por meio de $q_{max} = \pi M'/D_{max}$, como apontado na seção 3.2.5, e este valor não necessariamente é igual ao nominal N. Ou seja, o número de canais contendo informação útil não necessariamente equivale ao do teorema de Nyquist-Shannon, que contém a informação mínima para a reconstrução completa do sinal de SAXS, podendo ser maior que o número de pontos obtido pelo experimento para baixo ruído, e menor quando há pouca precisão na determinação das intensidades.

O programa SHANUM foi implementado para buscar este número ótimo M' a partir dos dados experimentais a partir de dois critérios realizando uma busca para um conjunto de valores tal que $M_i = \max(3, 0.2N_S)$ e $M_f = 1.25N_S$ são os limites do intervalo de busca para determinação de M'. O primeiro é o já mencionado $\chi^2(M)$, que é aplicado sobre o espaço recíproco, porém apenas isso é insuficiente pois não é bem definido o quão significativos devem ser os valores de a_n para que seja usado um valor para M' e não outro. Em cima

disso, o segundo critério diz respeito ao espaço real, avaliando a significância dos a_n a partir da expansão da $p(r)$ em uma somatória, de modo que, de acordo com a fórmula de interpolação de Shannon, valores maiores de truncamento inserem oscilações de frequências mais altos, e portanto a superestimação de M' leva a oscilações sem contribuição ao ajuste (processo ilustrado na Figura 3.18). Desse modo, o segundo critério é dado por uma medida $\Omega(p)$ definida por

$$\Omega(p) = \int_0^{D_{max}} \left[\frac{dp_M(r)}{dr}\right]^2. \tag{3.89}$$

O uso dos dois critérios combinados é dado pela métrica conjunta $f(M) = \chi^2(M) + K\Omega(p_M)$, com K sendo uma constante de proporcionalidade dada pela razão entre $\chi^2(M)$ e $\Omega(p_{M_i})$, e o programa busca minimizar $f(M)$ para encontrar M'. O programa, portanto, dá um critério robusto para a escolha do conjunto de dados a ser usado em ajustes e modelagens, indo além de regras empíricas.

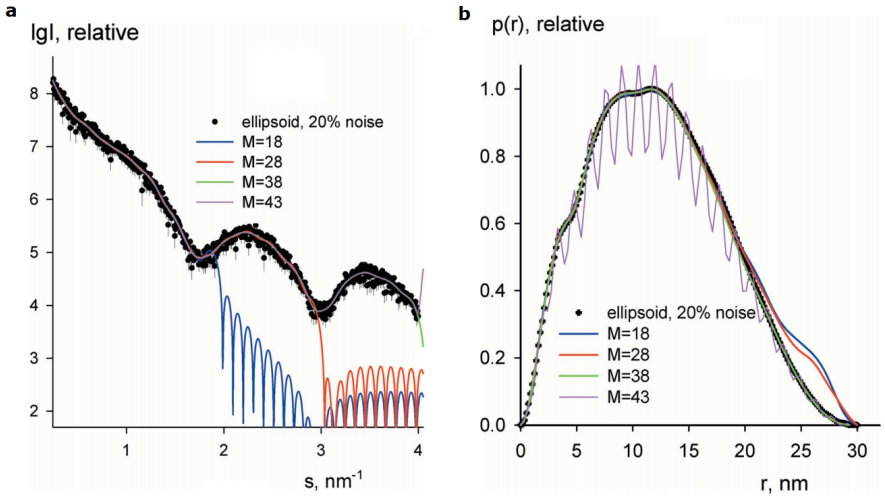

Figura 3.18: Ajustes a dados de SAXS (**a**) e funções $p(r)$ (**b**) por meio da fórmula de interpolação truncada para diferentes números de canais de Shannon, M (Konarev e Svergun, 2015).

3.4.8 Usando computação para encontrar flutuações internas de densidade eletrônica: DENSS

Uma das questões limitantes principais para a modelagem tridimensional em SAXS é a ausência de flutuações na densidade eletrônica das estruturas, de modo que é tomada uma densidade média para a proteína inteira. Isso se deve ao problema da fase não ser

resolvido, somado ao fato de que cada valor $I(q)$ corresponde à média tomada sobre um número de pixels. Porém, é possível a realização de métodos computacionais de busca exaustiva para o ajuste configurações tridimensionais de densidade eletrônica aos dados experimentais, como ocorre para o programa DENSS (*DENsity from Solution Scattering*, inglês para *Densidade de Espalhamento em Solução*) (Grant, 2018).

O algoritmo do DENSS consiste na atribuição de valores aleatórios de densidade eletrônica a cada elemento de volume (chamado *voxel*) em uma grade cúbica. Em cima dessa distribuição é calculada a transformada de Fourier e depois a média esférica desse resultado para a obtenção do perfil de espalhamento unidimensional de maneira análoga ao que ocorre para detectores bidimensionais. Após este passo busca-se ajustar as intensidades calculadas aos dados experimentais por meio da multiplicação das intensidades na grade tridimensional por constantes. Desse modo, obtêm-se um perfil que ajusta os dados e ainda mantêm as fases. Porém, uma vez que esse perfil pode não ter significado físico devido a várias distribuições poderem gerar o mesmo perfil de intensidades unidimensionais, são aplicados algoritmos que conservam regiões com densidades positivas e mantêm regiões externas a elas com um perfil constante (Marchesini et al., 2003; Fienup, 1978). O algoritmo é então repetido até que haja uma convergência da solução, sendo o processo ilustrado na Figura 3.19. O processo pode ser repetido e tomada a média de maneira análoga ao DAMAVER, e alinhamentos podem ser feitos com arquivos como feito no SUPCOMB. Exemplos de reconstruções estão presentes na Figura 3.20.

3.4.9 Um banco de dados para experimentos e modelos de SAXS: SASBDB

Conforme uma técnica se torna popular, uma quantidade cada vez maior de dados e resultados são publicados na literatura, e a criação de bancos de dados se tornam necessários não apenas para a curadoria desses resultados, como também para que sirvam de material de referência para outros pesquisadores que busquem explorar sistemas semelhantes. Também há pesquisadores que utilizam essas bases de dados para a elaboração de novas metodologias a partir de tendências observadas nos resultados. O maior exemplo de base de dados para biologia estrutural é o *Protein Data Bank* (PDB), que contém modelos de proteínas obtidos por difração de raios-X e ressonância magnética nuclear (Berman et al., 2000).

Por conta do aumento em número de publicações utilizando SAXS, a concepção de

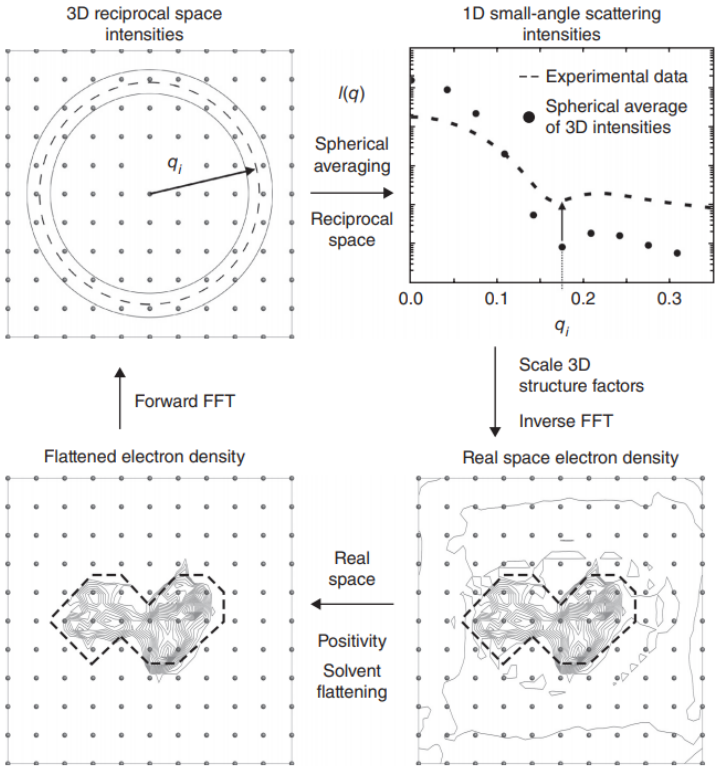

Figura 3.19: Processo de obtenção da estrutura tridimensional pelo DENSS: tentativa de ajuste de um perfil tridimensional de densidade eletrônica aleatoriamente gerada no início do processo (perfil calculado por meio de transformada de Fourier e integração radial das intensidades) a um conjunto de dados experimentais por meio da multiplicação de uma constante e correção das densidades eletrônicas a partir do ajuste obtido. O ciclo é repetido por um certo número de iterações até que haja convergência do valor de χ^2 do ajuste da distribuição de densidades eletrônicas aos dados (Grant, 2018).

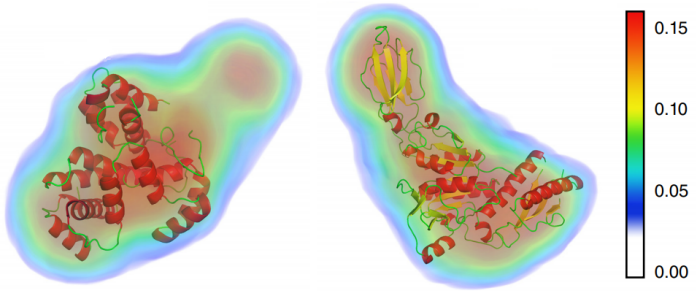

Figura 3.20: Estruturas tridimensionais obtidas pelo programa DENSS alinhadas a estruturas de alta resolução. À esquerda, a estrutura cristalográfica não está completa, e o DENSS mostra não só a região que falta, mas também que nela há alta densidade eletrônica, o que indica, possivelmente, presença de estrutura secundária (adaptado de Grant, 2018).

um banco de dados de dados experimentais e modelos obtidos pela técnica tornou-se uma realidade com o *Small Angle Scattering Biological Data Bank* (inglês para *Banco de Dados Biológico de Espalhamento a Baixos Ângulos*, SASBDB) (Valentini et al., 2014), que permite a consulta e utilização para estudos estatísticos de conjuntos de dados a mode-

los tridimensionais de proteínas, ácidos nucleicos e complexos de ambos. As entradas do SASBDB contém as curvas experimentais em escala logarítmica, podendo também conter os gráficos de Guinier e Kratky adimensional, a $p(r)$ e mesmo modelos desde corpo rígido e *ab initio* a modelos de ensemble com os respectivos ajustes. Cada entrada contém também informações sobre a amostra, onde foram tomados os dados, as condições experimentais e onde foram publicados os resultados, além de scores atribuídos aos conjuntos de dados e ajustes na aba de validação de dados (Kikhney et al., 2019). Esses scores baseiam-se em parâmetros como χ^2, p do CorMap e número de canais de Shannon, sendo exibidos em *sliders* de maneira similar aos parâmetros do PDB. Informações suplementares, como arquivos FASTA com as sequências de aminoácidos, organismo e referência da proteína no UniProt (Apweiler et al., 2004) também são apresentados. Pode-se fazer o *download* de todas as informações para curadoria e realização de outras análises. A interface gráfica do website (**https://www.sasbdb.org/**) está presente na Figura 3.21.

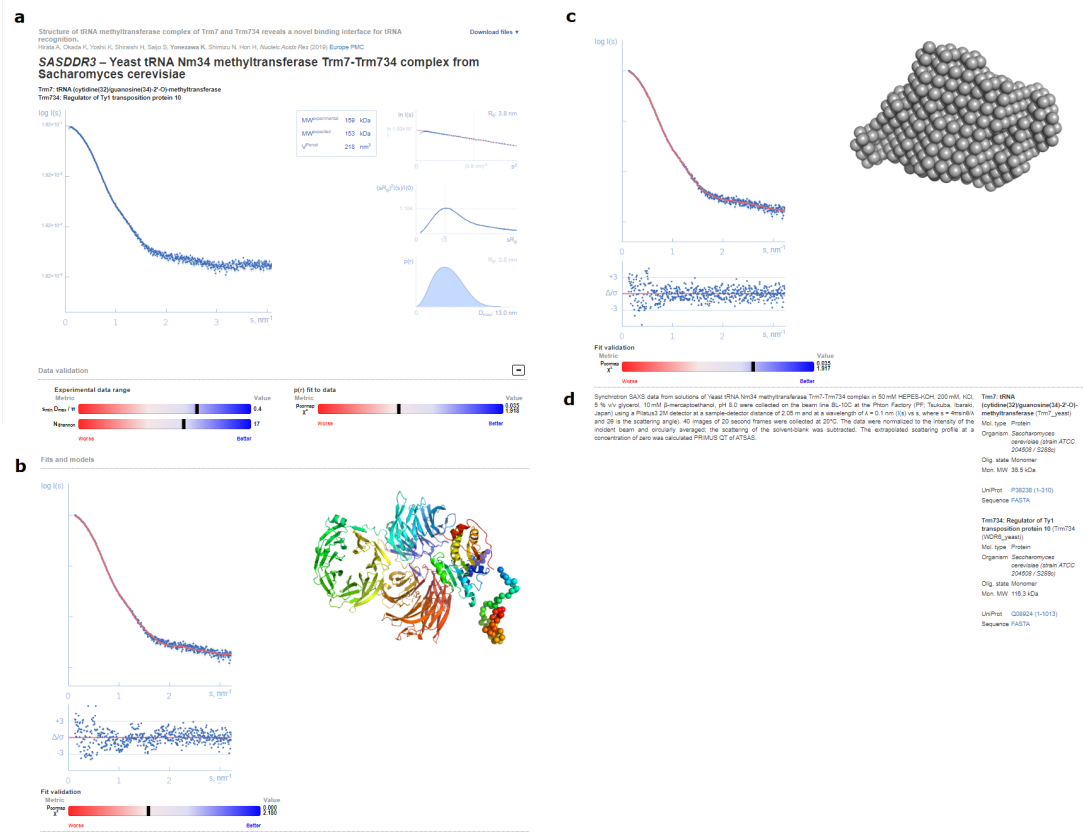

Figura 3.21: Interface gráfica do SASBDB: **a**. Nome da entrada com referência de publicação e autores, dados experimentais, gráficos de Guinier e Kratky e $p(r)$. **b**. Modelagem utilizando estruturas cristalográficas parciais de referência. **c**. Reconstituição *ab initio*. **d**. Informações sobre a tomada de dados e a proteína de estudo, como arquivo FASTA de sequência, massa molecular do protômero, organismo de origem, entrada UniProt e estado oligomérico. Scores são exibidos como sliders.

O SASBDB conta com mais de mil entradas e dois mil modelos com vários outros em processo de deposição e validação. Assim como curvas calculadas do PDB utilizando programas como CRYSOL ou FoXS, os dados experimentais do SASBDB já foram utilizados para estudos de base de dados na busca de métodos de obtenção de parâmetros como massa molecular, por exemplo (Franke et al., 2018).

3.5 Métodos de determinação da massa molecular

3.5.1 Métodos dependentes da concentração

3.5.1.1 Calibração com proteína padrão

Uma forma direta de cálculo da massa molecular consiste em utilizar uma proteína padrão, como lisozima, BSA e glicose isomerase, para calibração do $I(0)$ da proteína medida por meio de (Mylonas e Svergun, 2007)

$$MM_p = \frac{I(0)_p}{c_p}\frac{MM_{pad}}{I(0)_{pad}/c_{pad}}, \tag{3.90}$$

sendo MM_i a massa molecular e $I(0)_i$ a intensidade do feixe direto (extrapolada pela análise de Guinier ou pela $p(r)$) e c_i os valores de concentração da espécie i, sendo $i = p$ referente à proteína de estudo e $i = pad$ à proteína padrão. Assim, deve haver o cuidado extra uma vez que o resultado depende também da condição de outra proteína, e ao mesmo tempo ambas as concentrações devem ser determinadas cuidadosamente para que não hajam erros sistemáticos no cálculo de MM_p. Portanto, uma vez que o trabalho experimental com proteínas é delicado, apesar de ser um método simples em teoria, podem surgir entraves práticos não esperados, como agregação da proteína padrão, erros de concentração etc.

3.5.1.2 Calibração em escala absoluta usando água

Outro caminho para a obtenção da massa molecular é a calibração das intensidades medidas para escala absoluta a partir de padrões conhecidos, sendo o mais popular a água, cujo espalhamento depende apenas da compressibilidade isotérmica χ_T por meio de (Orthaber et al., 1999)

$$I_{água} = \rho^2 kT\chi_T, \tag{3.91}$$

ou seja, é entendido como constante, com ρ sendo uma densidade de comprimento de espalhamento. Uma vez que a compressibilidade isotérmica depende da temperatura, é necessário usar tabelas para obter os valores para χ_T. O mais usual é o uso de $\chi_T = 4.58 \times 10^{-10}$ Pa^{-1} a 293 K, implicando que $I_{água} = 1.632 \times 10^{-2}$ cm^{-1} a 20 °C. Assim, sabe-se o espalhamento *teórico* da água quando no "regime horizontal", e para que seja feita a conversão para escala absoluta o caminho é o seguinte:

- I) Medição do espalhamento do porta-amostra, da água e da amostra de estudo com proteína de interesse;
- II) Divisão do valor *constante* do espalhamento da água medido (já subtraído o espalhamento do porta-amostra, a partir de aproximadamente $q = 2nm^{-1}$) pelo valor teórico, encontrando uma constante C;
- III) Divisão das intensidades da curva experimental para a proteína pela mesma constante C.

Após este processo a curva estará normalizada, assim como o $I(0)$, obtido pela análise de Guinier. A relação entre $I(0)$ e MM é dada por (Orthaber et al., 1999)

$$MM = \frac{N_A I(0)}{c\Delta\rho_M^2}, \tag{3.92}$$

em que $N_A = 6.023 \times 10^{23}$ mol^{-1} é o número de Avogadro, c é a concentração de proteína e $\Delta\rho_M = r_0(\rho_{M,prot} - \rho_s\bar{\nu})$, sendo $r_0 = 2.8179 \times 10^{-13}$ cm o comprimento de espalhamento (ou raio clássico) de um elétron, $\rho_{M,prot} = 3.22 \times 10^{23}$ e g^{-1} a densidade numérica de elétrons por massa de proteína, $\rho_s = 3.34 \times 10^{23}$ e cm^{-3} a densidade volumétrica de elétrons do solvente e $\bar{\nu}$ o volume parcial específico da proteína (Mylonas e Svergun, 2007). Sugere-se o uso de um valor "efetivo" para o volume parcial específico proteico de $\bar{\nu} = 0.7425$ cm^3 g^{-1} (Mylonas e Svergun, 2007), apesar de que os valores, de acordo com a literatura, flutuem em 10% em torno deste valor (Perkins, 1986). O valor "efetivo" faz com que não seja necessário o uso de programas como o SEDNTERP para o cálculo teórico de $\bar{\nu}$ nem a realização de experimentos para obtenção da grandeza uma vez que os valores dela para proteínas são próximos o suficiente para que não tenham um impacto tão grande no cálculo da MM.

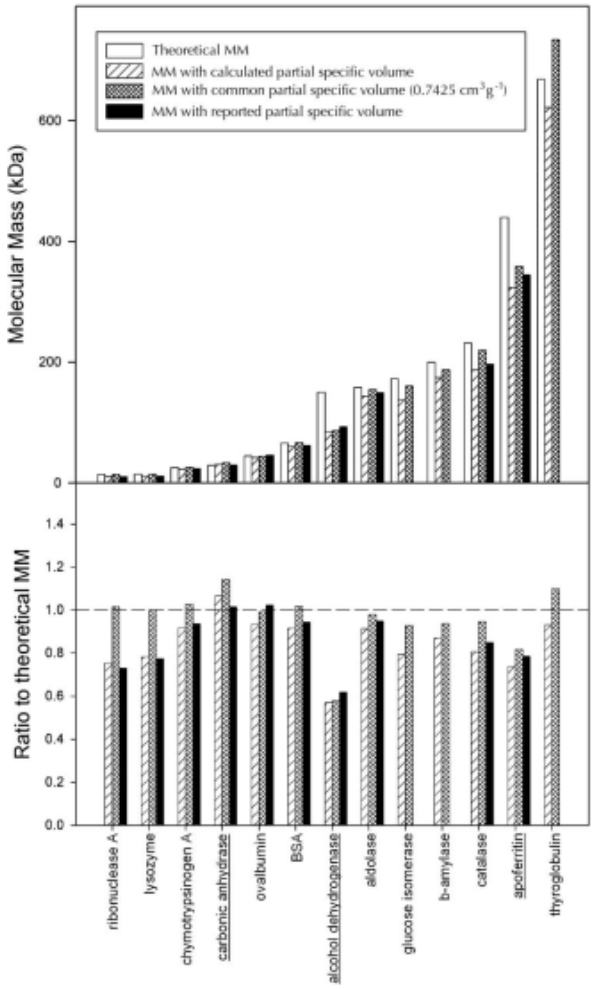

Figura 3.22: Gráfico indicando as MM calculadas e a razão entre as massas calculada e teórica para cada proteína estudada no artigo, comparando o uso de diferentes $\bar{\nu}$ nos parâmetros obtidos. (Mylonas e Svergun, 2007)

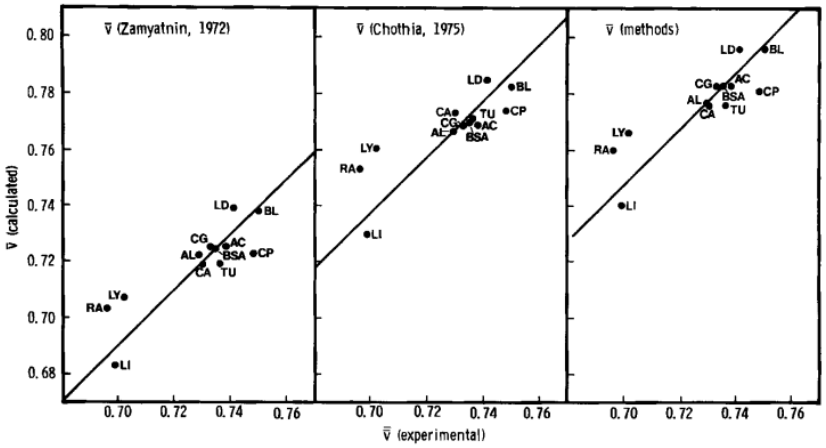

Figura 3.23: Comparações dos valores de $\bar{\nu}$ obtidos em diferentes estudos para as mesmas doze proteínas, com regressões lineares dos resultados obtidos (Perkins, 1986).

3.5.2 Métodos independentes da concentração

3.5.2.1 SAXSMoW: integrando o gráfico de Kratky

Dado que o espalhamento para $q \to 0$ é dado por (Fischer et al., 2010)

$$I(0) = N(\Delta\rho)^2 V^2 \tag{3.93}$$

e o invariante de Porod Q é dado pela integral do gráfico de Kratky, que, por sua vez, é dado por (Glatter e Kratky, 1982)

$$Q = \int_0^\infty I(q) q^2 dq = 2\pi^2 (\Delta\rho)^2 NV, \tag{3.94}$$

de modo que temos que o volume da partícula espalhadora é dado por

$$V = 2\pi^2 \frac{I(0)}{Q}. \tag{3.95}$$

Issos nos permite, portanto, obter a massa molecular MM de uma proteína por meio do produto de V com a densidade proteica média $\rho_m \approx 1.37\text{g cm}^{-3}$ a partir de

$$MM = V\rho_m = 2\pi^2 \frac{I(0)}{Q} \rho_m. \tag{3.96}$$

Ou seja, é teoricamente possível obter a massa a partir do gráfico de Kratky, sendo o único empecilho a integração em si. Porém, como observado nos últimos anos (Fischer et al., 2010), é possível elaborar um método de cálculo de massa truncando a integral até um valor q_{max}, o que definiria um invariante "aparente" Q' dado por

$$Q' = \int_0^{q_{max}} I(q) q^2 dq, \tag{3.97}$$

que, por sua vez, permite a obtenção de um "volume aparente" V', obtido de maneira análoga ao volume "exato" teórico já mencionado. A chave da metodologia elaborada por Fischer et al (Fischer et al., 2010) consiste em um estudo sobre a relação entre V e V' em função do q_{max} utilizado, incluindo um estudo para mais de mil estruturas depositadas no *Protein Data Bank*. A Figura 3.24a mostra o gráfico relacionando V e V' a partir da variação de q_{max} para uma esfera, e a Figura 3.24b mostra o mapeamento de volumes para as estruturas do PDB para três valores de q_{max} diferentes, com o espalhamentos teóricos desses arquivos sendo calculados pelo CRYSOL, de onde eles também extraíram V e V'.

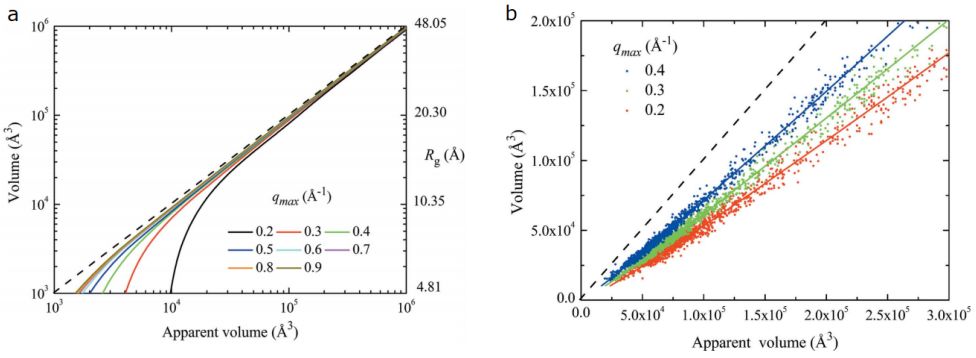

Figura 3.24: Relação entre volume real e aparente para diferentes valores de truncamento q_{max} na integração para obtenção de Q' para curvas calculadas de esferas com diferentes volumes (**a**) e de um conjunto de estruturas cristalográficas (**b**). As linhas tracejadas correspondem ao caso ideal em que o volume real e o aparente são iguais, e cores diferentes indicam diferentes valores de q_{max} adotados. (Fischer et al., 2010).

Os autores então correlacionaram os volumes "real" e aparente por meio de ajustes lineares, V = A + BV', dados por coeficientes A e B que variam de acordo com o q_{max} utilizado. Essa metodologia é empregada no programa *SAXSMoW 2.0* (disponível *online* no endereço **http://saxs.ifsc.usp.br**) (Piiadov et al., 2019). O programa recebe um conjunto de dados do usuário e realiza a análise de Guinier de maneira automatizada, imprime os gráficos com os dados, o de Guinier, de Kratky e de Porod. Após isso o usuário escolhe uma de três opções para q_{max}, sendo elas

- $q_{max} = 8/R_g$, recomendado como padrão;
- $\dfrac{I(0)}{I(q_{max})} = 10^{2.25}$, recomendado para proteínas mais elongadas (Piiadov et al., 2019);
- q_{max} determinado manualmente.

Após a escolha o programa exibe o resultado do cálculo da MM e, caso seja colocado pelo usuário, também mostra a discrepância D entre a MM calculada e a da sequência, obtida meio de

$$D = \left| \frac{m}{m_0} - 1 \right| \times 100\%, \tag{3.98}$$

incluindo também, a partir da razão entre as massas, uma estimativa do estado oligomerico da proteína. A Figura 3.25 mostra a interface do programa com um conjunto de dados de exemplo. Piiadov et al estudaram os valores de D e encontraram que para proteínas cada vez mais elongadas D aumenta progressivamente.

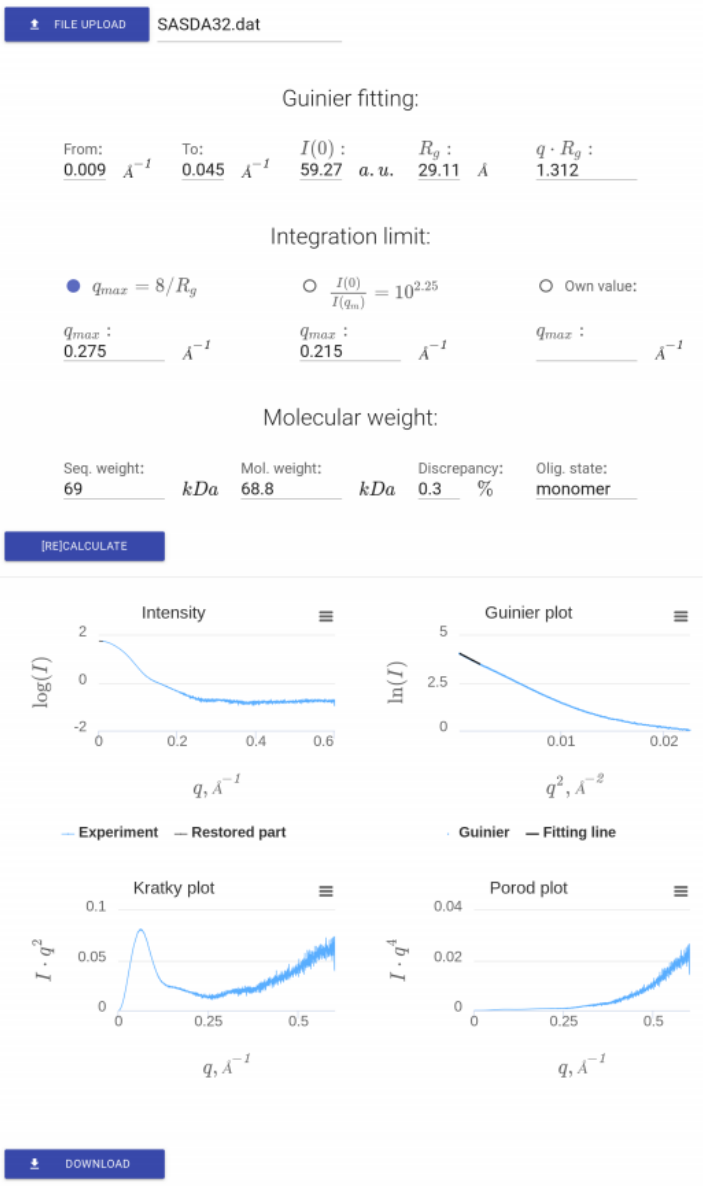

Figura 3.25: Interface do programa SAXSMoW acessado por *browser* (Piiadov et al., 2019).

Há uma outra forma implementada no ATSAS, chamada Q_p, para obtenção da MM também calcula o volume V_p por meio do invariante Q encontrado para um diferente intervalo de integração e divide o volume por 1.37, sem o uso de fatores de correção como os apresentados por Fischer et al (Fischer et al., 2010).

3.5.2.2 Volume de correlação, V_c: introduzindo um novo invariante

Apesar da praticidade, a determinação de MM por meio de Q é impraticável para proteínas parcial ou totalmente desenoveladas devido à não-convergência do gráfico de Kratky. Esta questão leva paradoxalmente a uma não definição de parâmetros físicos

como o volume V_p. Para endereçar este problema da convergência, é possível trabalhar com gráficos de $qI(q)$ vs q, cujo perfil sempre converge, representativo da intensidade total espalhada pela partícula (Glatter e Kratky, 1982). Essa intensidade total está relacionada ao comprimento de auto-correlação l_c por meio de (Rambo e Tainer, 2013)

$$\int_0^\infty qI(q)dq = cV_p(\Delta\rho)^2 \int_0^\infty \gamma(r)dr = cV_p(\Delta\rho)^2 2\pi l_c. \tag{3.99}$$

Assim, a função de correlação por si só já $\gamma(r)$ define l_c. Usando a razão entre $I(0)$ e essa intensidade total espalhada é possível definir uma grandeza invariante dada por (Rambo e Tainer, 2013)

$$V_c = \frac{I(0)}{\int qI(q)dq} = \frac{cV_p^2(\Delta\rho)^2}{cV_p(\Delta\rho)^2 2\pi l_c}, \tag{3.100}$$

e, portanto,

$$V_c = \frac{V_p}{2\pi l_c}. \tag{3.101}$$

Por se tratar da razão entre o volume de Porod V_p e o comprimento de correlação l_c a grandeza V_c é chamada *volume por comprimento de correlação*, dada em unidade de Å^2. Importante notar que, assim como Q, esta grandeza independe da concentração de amostra, e desta vez é dada por uma integral que na prática converge tanto para proteínas bem enoveladas quanto para intrinsecamente desordenadas (Rambo e Tainer, 2013).

O cálculo de massa fazendo uso do V_c se dá sem a necessidade de conversão para escala absoluta nem de hipóteses sobre a forma da partícula. Ao se produzir, para um parâmetro Q_R definido por $Q_R = V_c^2/R_g$ (com, portanto, dimensão de Å^3), um gráfico de $\ln[Q_R]$ vs $\ln[MM]$ percebe-se uma lei de potência entre ambos, dada por

$$MM = \left(\frac{Q_R}{e^d}\right)^{1/k}, \tag{3.102}$$

com os parâmetros d e k sendo empíricos específicos para tipo de molécula, podendo ser para proteínas, ácidos nucleicos ou complexos de ambos, como presente na Figura 3.26. Por estar diretamente relacionado à MM, Q_R traz informações sobre a presença de contaminantes ou estados oligomericos que não são de fácil observação a partir de R_g ou de V_c.

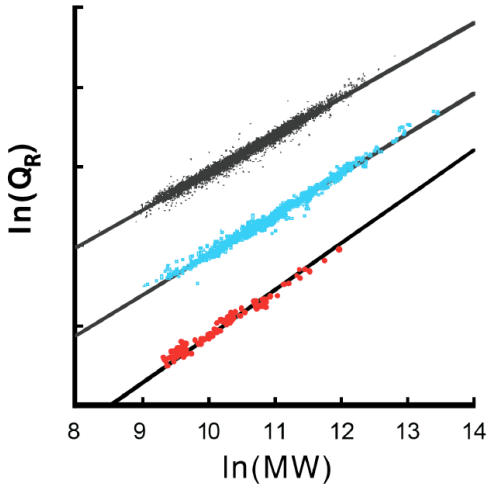

Figura 3.26: Gráficos relacionando Q_R a MM de proteínas (preto), complexos de proteína com ácido nucleico (azul) e RNA (vermelho), indicando que os parâmetros a serem usados para relacionar ambas as grandezas dependem do tipo de amostra a ser estudada (Rambo e Tainer, 2013).

3.5.2.3 *Size & Shape: machine learning a partir de gráficos de Kratky para predições de MM e D_{max}*

Outra forma, conhecida como *Size & Shape*, de endereçar a questão da massa em SAXS faz uso de bancos de dados e busca encontrar parâmetros estruturais a partir de partículas similares por meio de machine learning. Para facilitar essa busca, é necessário reduzir o número de parâmetros de busca para o mínimo possível contendo o máximo de informação estrutural sobre o objeto de estudo a fim de se formar um vetor com alguns valores que permitam entender variações estruturais a partir da distância entre vetores. Para a determinação desses vetores foram escolhidos parâmetros encontrados fazendo uso do gráfico de Kratky normalizado: os volumes aparentes normalizados V', obtidos a partir dos invariantes de Porod normalizados Q', calculados para diferentes valores máximos de qR_g para integração ($qR_g^{max} = 3$, 4 e 5). Também foi escolhido o R_g, totalizando um total de quatro componentes no vetor descritivo, e às estruturas são atribuídos pontos em um *espaço* V', e a partir deste mapeamento se estimam os parâmetros MM e D_{max} e se classifica o tipo de estrutura como compacto, estendido, achatado entre outros.

Para a determinação de MM e D_{max} se tira a média ponderada entre esses parâmetros para os k primeiros vizinhos no espaço V' (implementado originalmente considerando $k = 5$ de modo a minimizar os erros de estimativa), sendo os fatores de peso calculados de modo que quanto mais próximo o vizinho maior o peso dado no cálculo da média.

Para atribuição correta dos valores do vetor são necessários R_g e V' (e, portanto,

$I(0)$) bem determinados, de modo que o método GPA (apresentado na seção 3.3.1) é útil para validação desse tipo de análise. Ao mesmo tempo, caso MM e D_{max} sejam sabidos (quando, por exemplo, o estado oligomerico já é bem determinado) a Size & Shape pode também servir para validar os valores de R_g e $I(0)$ encontrados. Também é possível usar esta metodologia para, ao estimar o tipo de partícula, determinar restrições à modelagem *ab initio* ou até mesmo sugerir o uso de outros tipos de análise, como a de ensemble. Outro uso útil deste tipo de análise é obter uma estimativa inicial de D_{max} a ser utilizada como entrada para o GNOM. O programa que faz a classificação dos dados atribuindo uma geometria é o DATCLASS (Franke et al., 2018).

Portanto, assim como o SAXSMoW, o trabalho também faz uso dos gráficos de Kratky. Porém, neste método é usado o Kratky *normalizado* e como forma de classificação estrutural proteica.

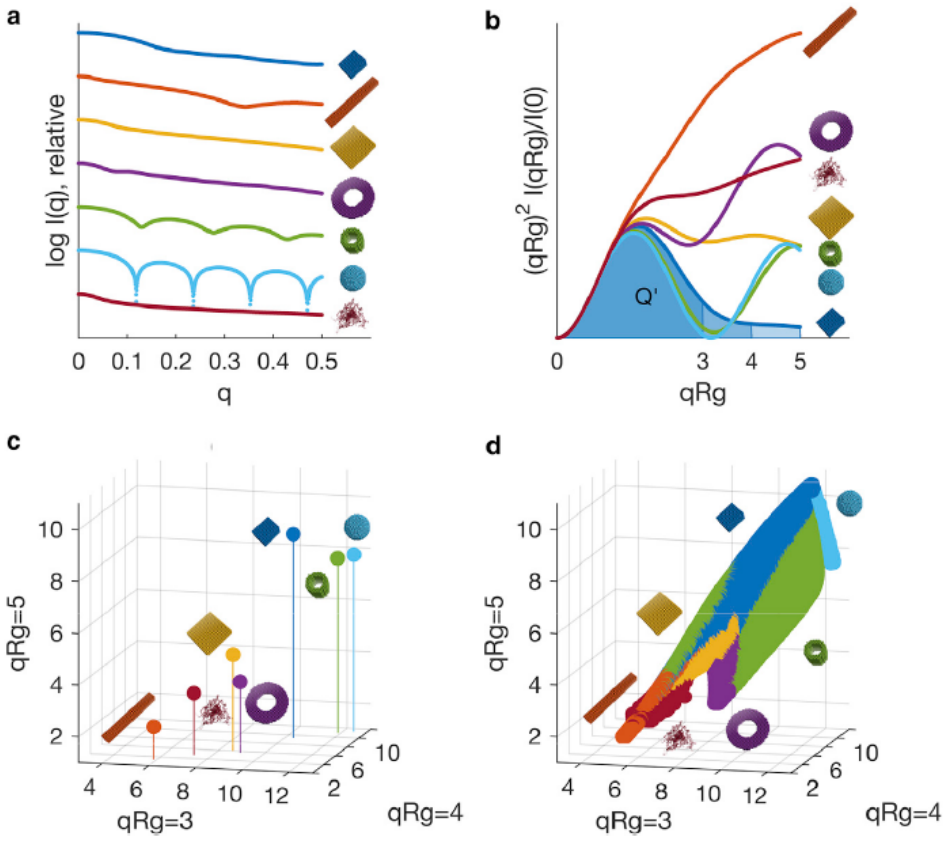

Figura 3.27: Procedimento de elaboração do espaço V': cálculo de curvas teóricas para diferentes topologias a partir de expressões analíticas (**a**), obtenção de gráficos de Kratky adimensionais para cálculo do invariante Q' para diferentes valores de truncamento da integral, $qR_g = 3$, 4 e 5 (**b**), e mapeamento desses valores no espaço tridimensional V' (**c** e **d**). As regiões referentes a cada topologia são representadas por diferentes cores, de modo que é possível relacionar uma curva experimental a uma forma (Franke et al., 2018).

3.5.2.4 Estimativas empíricas de MM por meio de V_p

Há uma regra empírica de relação entre MM e V_p que não leva em conta a densidade média proteica, dada por $MM \approx V_p/1.7$ quando aplicando a metodologia Q_p. Também é possível obter aproximadamente o volume da partícula espalhadora por um caminho que não envolve o cálculo direto do invariante Q, mas sim seria pelo cálculo do volume do envelope obtido por modelagem *ab initio*, que leva a uma regra similar: $MM \approx V_{envelope}/2$ (Petoukhov et al., 2012). Nenhuma das regras apresenta comportamento sistemático com relação à natureza da partícula, como anisometria e tamanho, como visto na Figura 3.28. São métodos menos confiáveis por terem uma acurácia mais baixa porém úteis por serem de simples aplicação em processos automatizados de análise (Petoukhov et al., 2012).

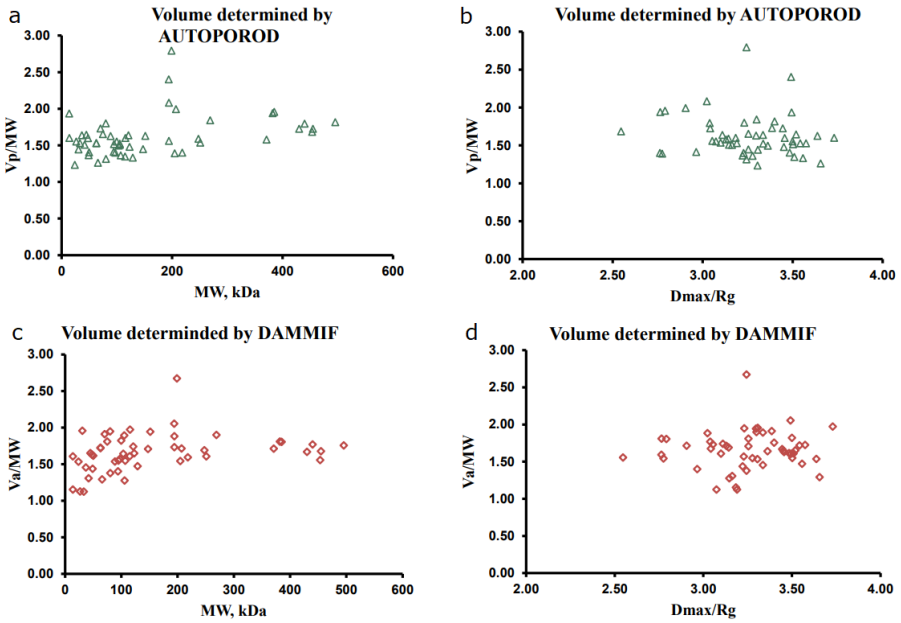

Figura 3.28: Relação entre a razão V/MM e MM (**a** e **c**) e D_{max}/R_g (**b** e **d**) usando volumes calculados pela lei de Porod (verde) e modelos *ab initio* (vermelho) (Petoukhov et al., 2012).

3.5.3 Conciliando métodos por meio de inferência Bayesiana: DatBayes

Uma vez que cada método independente de concentração tem limitações e, portanto, maior ou menor acurácia na obtenção da MM da proteína de estudo, é importante utilizar mais de uma abordagem. Uma abordagem probabilística é possível para buscar um cálculo de massa levando em conta os métodos independentes de concentração apresentados. Para isso, primeiramente obtém-se as MM de curvas calculadas para um grande conjunto de estruturas do PDB aplicando cada método e, a partir de um gráfico de $MM_{método}$ vs

$MM_{sequência}$ pode-se atribuir distribuições de probabilidade para cada valor de MM calculado (Hajizadeh et al., 2018).

É possível utilizar o teorema de Bayes como forma de inferir a MM (desconhecida) a partir dos valores calculados por cada método ao entendê-los como pedaços de informação, ou *evidências*. Seja $P(H \mid E)$ a probabilidade da *hipótese* H (no caso, a massa molecular em kDa) ter certo valor a partir de uma evidência E o teorema de Bayes diz que

$$\begin{aligned} P(H = MM(kDa) \mid E_{Q_p}, E_{V_c}, E_{MoW}, E_{S\&S}) = \\ = \frac{P(E_{Q_p} \mid H)\, P(E_{V_c} \mid H)\, P(E_{MoW} \mid H)\, P(E_{S\&S} \mid H)\, P(H)}{P(E)} \end{aligned} \tag{3.103}$$

onde $P(H = MM(kDa) \mid E_{Q_p}, E_{V_c}, E_{MoW}, E_{S\&S})$ é a probabilidade da MM ser igual à da hipótese dado que cada método n calculou um valor E_n, $P(E_n \mid H)$ é a distribuição de probabilidade para certa MM para o método n, $P(E)$ é um fator de normalização dado pela soma das probabilidades para cada método e $P(H)$ é a distribuição uniforme (assume-se que todos os valores de MM sejam equiprováveis). A metodologia de obtenção das probabilidades está sintetizada na Figura 3.29. O método Q_p é similar ao SAXSMoW (3.5.2.1) porém utilizando um valor fixo para q_{max} e não fazendo uso dos fatores empíricos de correção, V_c é o que faz uso do volume de correlação (3.5.2.2), MoW é o próprio método do SAXSMoW e $S\&S$ se refere ao método *Size & Shape* (3.5.2.3).

O método funciona da seguinte maneira: uma base de dados foi construída a partir de um conjunto de curvas teóricas calculadas de estruturas do PDB, e cada forma de cálculo de MM foi aplicada a essas curvas e os valores obtidos foram comparados aos valores reais de massa. A partir disso, foram elaborados mapas $MM_{método}$ vs $MM_{sequência}$, e para cada valor de $MM_{método}$ foram elaboradas distribuições de probabilidade para que sejam usadas na equação de Bayes e, a partir da junção dessas distribuições, se obtém uma distribuição conjunta utilizada para encontrar um valor final "por consenso" para a MM.

3.5.4 Densidade proteica como ferramenta para estimativa de estado oligomérico

Um possível entrave em análises de SAXS é não ter uma MM teórica de referência confiável para a comparação dos resultados obtidos por outros métodos, sendo uma das informações mais importantes comprometida o estado oligomérico da proteína de estudo. Para contornar este problema uma possível abordagem é o uso dos valores das MM cal-

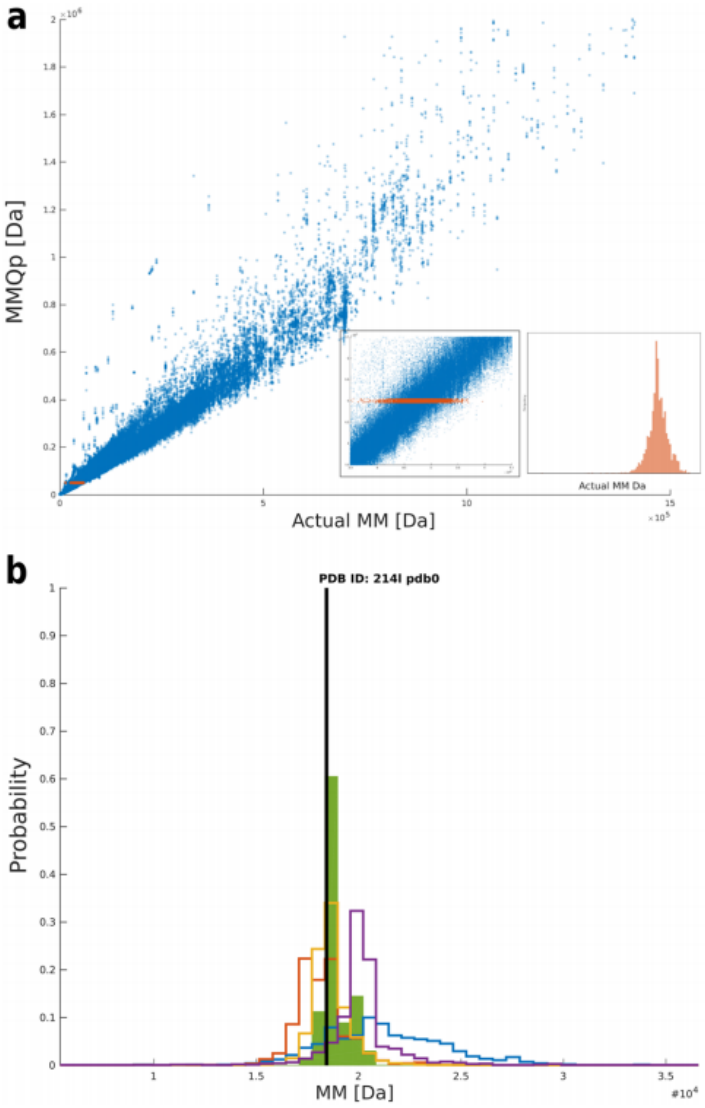

Figura 3.29: Procedimento de cálculo de MM pelo método Bayesiano com respectivas probabilidades: cálculo de MM por cada método para várias estruturas do PDB e comparação à MM da sequência, gerando distribuições de probabilidade para cada método (**a**), e, a partir das distribuições (cada método sendo representado por uma cor) e da equação de Bayes, determinação da MM com o respectivo intervalo de confiança (**b**) (Hajizadeh et al., 2018).

culadas juntamente com os volumes para a obtenção da densidade (chamada *densidade de teste*, ou ν_{teste}) dada por (Rambo e Tainer, 2011)

$$\nu_{teste} = N\frac{MM}{V_P}1.66, \tag{3.104}$$

sendo N o número de subunidades da proteína de estudo e 1.66 um fator de conversão de Da/Å^3 para g/ml (ou g/cm^3). Foram estipulados, a partir de dados de SAXS, os valores mais comuns para ν_{teste} para proteínas conhecidas e se encontrou valores na faixa

de $0.9 - 1.6\text{g/cm}^3$. Os valores abaixo do nominal para proteínas de 1.37g/cm^3 pode se dever a regiões como His-tags ou a dados de proteínas com grandes porções desordenadas, o que aumenta artificialmente o volume calculado (que faz uso de Q) uma vez que o SAXS obtém uma média de todas as orientações e conformações das partículas em solução. Utilizando esses valores de referência, pode-se variar N de modo que se encontre um estado oligomérico com uma densidade dentro de valores esperados para proteínas (Rambo e Tainer, 2011). Cada valor pode ser utilizado como referência a depender se a proteína de estudo é intrinsecamente desordenada ou bem enovelada. Importante notar que o valor de 1.37g/cm^3 pode ter aumentos sistemáticos para proteínas com baixos valores de massa molecular (Fischer et al., 2004). Portanto, comparando os valores mais comuns para ν_{teste} é possível filtrar estados oligoméricos acessíveis à proteína sem a necessidade de uma MM de referência.

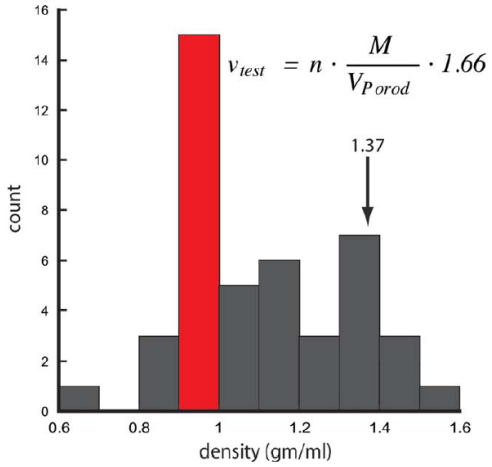

Figura 3.30: Histograma de densidades proteicas calculadas fazendo uso de SAXS, mostrando uma distribuição ampla para valores de referência, com alta ocorrência par valores abaixo do nominal de 1.37g/cm^3, possivelmente devido a regiões flexíveis que aumentam o valor de Q obtido e, portanto, também o volume V_P, levando a uma redução da densidade calculada (Rambo e Tainer, 2011).

3.5.5 Comparação entre curvas e ajustes

Uma vez que a determinação de incertezas não é ainda bem estabelecida se buscou formas de determinação de qualidade de ajustes para além da popular medida do χ^2 reduzido (χ^2_{red}), dada por

$$\chi^2_{red} = \frac{1}{N_{gl}} \sum_{i=1}^{n} \frac{[I_{\exp}(q_i) - I_{fit}(q_i)]^2}{\sigma_i^2}, \tag{3.105}$$

em que $I_{\exp}(q_i)$ é a intensidade medida no i-ésimo ponto experimental, $I_{fit}(q_i)$ é a intensidade do ajuste, σ_i é a incerteza associada ao ponto experimental e N_{gl} é o número de graus de liberdade. Para bons ajustes espera-se $\chi^2_{red} \approx 1$, porém uma vez que as incertezas podem ser mal estimadas (por um arranjo experimental diferente, por exemplo) este critério torna-se problemático, e por este motivo o uso de métodos alternativos de avaliação de qualidade de ajustes é recomendável.

3.5.5.1 Novas métricas para avaliação de ajustes e resolução utilizando canais de Shannon: χ^2_{free}, R_{SAS} e V_R

A utilização do χ^2 tal como foi concebido para o conjunto de dados de SAXS pode levar a ajustes fisicamente não representativos (sobreajuste, ou *overfitting*) quando os dados apresentam alto ruído relativo ou um número superestimado de graus de liberdade. Portanto, o problema não está no procedimento do χ^2 em si, mas sim nos conjuntos de dados e na taxa de amostragem de um experimento de SAXS. Foi elaborada uma métrica análoga ao χ^2 que faz uso apenas do conjunto de pontos descritos pelos canais de Shannon: o χ^2_{free} (Rambo e Tainer, 2013), cuja concepção é similar ao R_{free} da cristalografia (Brünger, 1992). A partir dos resultados que fazem uso da teoria de informação (seção 3.2.5), é possível recuperar um sinal de experimento de SAXS sem ruído a partir dos canais de Shannon e da estimativa de D_{max}.

O algoritmo de determinação de χ^2_{free} funciona da seguinte maneira: o conjunto de dados é dividido em intervalos regulares ao longo de q, sendo o número de intervalos igual ao número de canais de Shannon, e de cada intervalo é tirado um ponto aleatório como representante daquele conjunto, e então o χ^2 é então calculado a partir deste conjunto de pontos. Para evitar a escolha de um conjunto de pontos muito díspares do resto o procedimento de amostragem aleatória de pontos é repetida várias vezes (recomenda-se em torno de 1000 vezes), e então se escolhe a mediana dos χ^2 encontrados como o valor de χ^2_{free}. Esta nova métrica apresenta maior estabilidade com relação à relação sinal-ruído e valores menores que os do menor χ^2 para dados mais ruidosos, e o fato de o procedimento ser repetido várias vezes para alguns pontos (geralmente da ordem de ao menos 10 a 40) faz com que método seja menos sensível a pontos muito díspares e ao over-fitting (Rambo e Tainer, 2013).

O método descrito acima consiste em uma filtragem dos pontos experimentais relevan-

tes, porém não há um peso atribuído para cada região da curva de modo que para dois valores iguais de χ^2_{free} não se sabe se a concordância maior entre dados e ajuste ocorre para menores ou maiores valores de q (que leva a maiores ou menores disparidades no modelo gerado) sem avaliar graficamente, o que insere um fator subjetivo na análise. Portanto, o χ^2_{free} sozinho ainda não revela o suficiente sobre quanto um ajuste concorda com os dados experimentais ao pensar em modelo gerado. Esta questão pode ser endereçada pensando em função dos parâmetros de tamanho, dados por R_g e V_c (seção 3.5.2.2), de modo que uma menor discrepância entre estes parâmetros juntamente com baixo χ^2_{free} implicam em um bom ajuste tanto em relação ao modelo gerado como em relação aos pontos experimentais. A métrica R_{SAS} mede a disparidade entre parâmetros de tamanho, sendo definida como

$$R_{SAS} = \frac{(R_g^{\text{exp}} - R_g^{ajuste})^2}{(R_g^{\text{exp}})^2} + \frac{(V_c^{\text{exp}} - V_c^{ajuste})^2}{(V_c^{\text{exp}})^2}. \tag{3.106}$$

Utilizando o valor de R_{SAS} como referência pode-se variar o número de intervalos a serem tomados nos cálculos de discrepância entre dados e modelo, de modo que truncar o conjunto de dados até que se encontre um R_{SAS} que tenda a zero juntamente com um $\chi^2_{free} \leqslant 1.5$, por exemplo, terá no seu valor máximo de q utilizado uma medida de limite de resolução aceitável entre dados e modelo. Uma definição alternativa de resolução, portanto, pode ser definida como sendo o valor de q relativo ao canal de Shannon mais alto que apresente discrepância suficientemente aceitável com os dados. Importante apontar que, diferentemente do caso do programa SASRES (Tuukkanen et al., 2016), que é uma medida de resolução definida como variabilidade de modelos *ab initio* tridimensionais gerados, aqui trata-se de uma medida de até que resolução o ajuste descreve os dados, podendo representar modelos *ab initio*, modelos de corpo rígido ou mesmo ajustes de ensembles de conformações, tendo aplicações mais amplas. Portanto, tratam-se de "resoluções" de diferentes naturezas.

Também é interessante a introdução de um critério mais quantitativo para comparação de curvas experimentais. Em cima disso foi definido o critério razão de volatilidade (do original *volatility ratio*), V_R, que faz uso da razão entre duas curvas experimentais e de pontos amostrados por canais de Shannon, e não por todos os pontos (Hura et al., 2013). Para o cálculo de V_R primeiramente se divide uma curva por outra e se normaliza a razão entre ambas a média seja igual a 1. Para o uso de canais de Shannon, o trabalho original estabeleceu como padrão o uso de 25 canais, o que restringe as dimensões da proteína para

$D_{max} < 40\text{nm}$ e $q_{max} < 2\text{nm}^{-1}$. V_R então é definido como sendo (Hura et al., 2013)

$$V_R = \sum_{i=1}^{N=25} \left| \frac{R(q_i) - R(q_{i+1})}{\big[R(q_i) + R(q_{i+1})\big]/2} \right|, \tag{3.107}$$

ou seja, a razão de volatilidade é dada como a soma sobre os canais de Shannon relativa ao módulo da diferença de razão entre pontos de canais consecutivos dividida pela a média entre elas.

Metodologias quantitativas podem ser utilizadas para um mapeamento de diferentes conformações de uma proteína. Ao se estudar, por exemplo, uma proteína na presença de diferentes ligantes, pode-se organizar os conjuntos de dados em uma matriz em que cada elemento se traz o valor de determinado critério seja χ^2, coeficiente de correlação de Pearson (Dos Reis et al., 2011) ou V_R, como presente na Figura 3.31. Essa abordagem traz tanto um elemento quantitativo quanto visual para a comparação entre conjuntos de dados e foi cunhada mapa de comparação estrutural (ou SCR, de *Structural Comparison Map*) (Hura et al., 2013). A elaboração dessa matriz pode ser feita facilmente no *website* da linha SYBILS, do Lawrence Berkeley National Laboratory, pelo endereço **https://sibyls.als.lbl.gov/saxs_similarity**.

3.5.5.2 Contornando a questão das incertezas: CorMap

É possível abordar o problema com uma metodologia mais visual que faça uso de correlações entre variâncias e covariâncias junto com análises probabilísticas para determinar qual grau de divergência entre curvas ou entre curva e ajuste é relevante estatisticamente, e é o caso da metodologia de análise *Correlation Map* (inglês para *Mapa de Correlação*, CorMap). Para altos números de contagem de fótons, especialmente em fontes síncrotron, a distribuição de Poisson tende a uma distribuição gaussiana, e pode-se considerar as medições $I_{\text{exp}}(q)$ como uma variável aleatória que segue uma distribuição gaussiana de valor esperado $I(q_k)$ e desvio padrão $\sigma(I(q_k))$. Desse modo é possível descrever a curva experimental como uma coleção de amostras gaussianas de modo que o vetor de intensidades esperadas J e a matriz de covariância Σ:

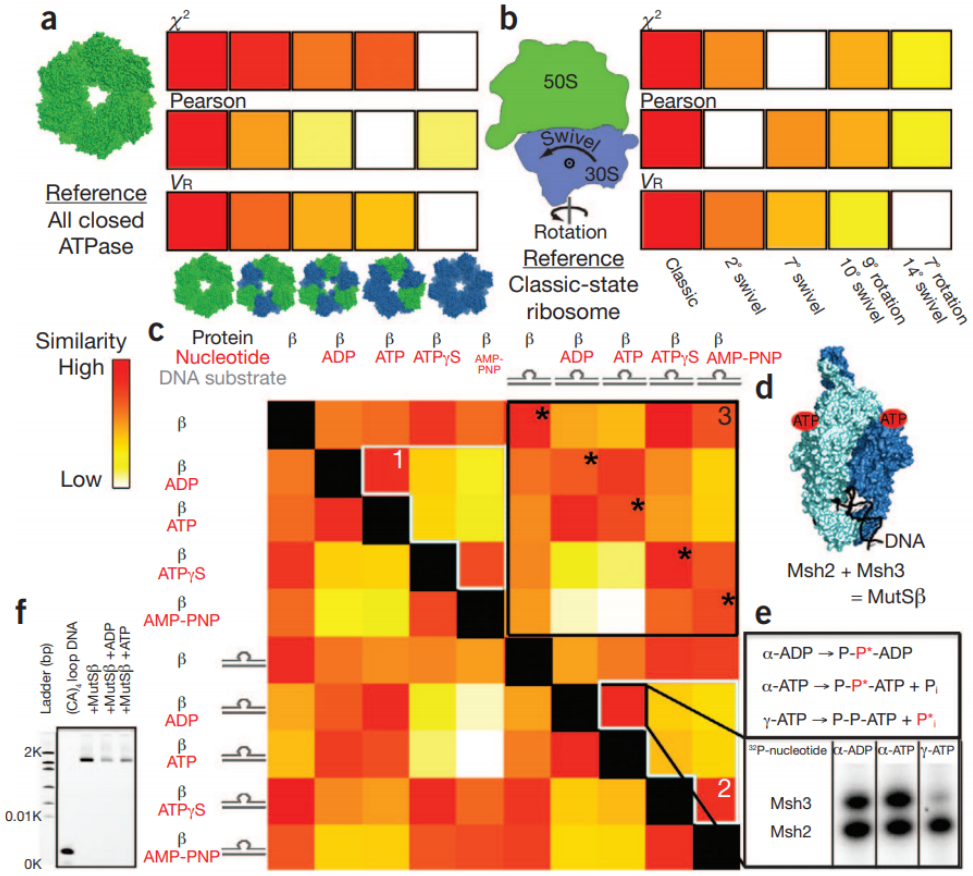

Figura 3.31: Exemplo de SCR, comparando diferentes estados da proteína ao ligar nucleotídeos de adenosina e mapear a similaridade entre os dados obtidos (Hura et al., 2013).

$$J = \begin{pmatrix} I(q_1) \\ \vdots \\ I(q_k) \\ \vdots \\ I(q_N) \end{pmatrix}, \tag{3.108}$$

sendo N o número de pontos experimentais, e

$$\Sigma = \begin{pmatrix} \ddots & & & \\ & \sigma(I(q_k))^2 & \dots & \sigma(I(q_k), I(q_l)) \\ & \vdots & \ddots & \vdots \\ & \sigma(I(q_l), I(q_k)) & \dots & \sigma(I(q_l))^2 \\ & & & \ddots \end{pmatrix}. \tag{3.109}$$

Caso sejam tomados m quadros experimentais, como é usual para verificar, por exemplo, dano por radiação, a intensidade média dos quadros será dada por

$$\bar{I}_{\exp}(q_k) = \frac{1}{m}\sum_{i=1}^{m} I_{\exp}(q_k)_i \tag{3.110}$$

enquanto os termos diagonais e não diagonais, respectivamente, serão dados por

$$\sigma(I_{\exp}(q_k))^2 = \frac{1}{m-1}\sum_{i=1}^{m}[I_{\exp}(q_k)_i - \bar{I}_{\exp}(q_k)]^2, \tag{3.111}$$

$$\sigma(I_{\exp}(q_k), I_{\exp}(q_l)) = \frac{1}{m-1}\sum_{i=1}^{m}[I_{\exp}(q_k)_i - \bar{I}_{\exp}(q_k)][I_{\exp}(q_l)_i - \bar{I}_{\exp}(q_l)]. \tag{3.112}$$

Lembrando que os índices k se refere aos k-ésimo e l-ésimo pontos experimentais, enquanto i diz respeito ao i-ésimo quadro. É possível definir uma correlação r_{kl} $(k \neq l)$ como a razão

$$r_{kl} = \frac{\sigma(I_{\exp}(q_k), I_{\exp}(q_l))}{\sigma(I_{\exp}(q_k))\sigma(I_{\exp}(q_l))}, \tag{3.113}$$

de modo que $-1 < r_{kl} < 1$. Para melhor visualização, os valores podem ser mapeados em uma matriz, porém usando uma escala de cinza indo de preto ($r_{kl} = -1$) a branco ($r_{kl} = 1$). Regiões brancas ou pretas indicam desvios sistemáticos de um quadro em relação ao outro, e mapas sem padrões claros e muita alternância entre tons claros e escuros indicam ausência de erros sistemáticos. Pode-se fazer o mesmo raciocínio que o feito acima para comparar um conjunto de dados experimentais com um ajuste. A Figura 3.32 mostra um exemplo de como o CorMap pode ser usado para comparar um conjunto de dados com ajustes provenientes de modelos.

A partir do número máximo de pontos experimentais consecutivos que apresentem correlação de +1 ou -1 (chamado de "comprimento máximo") é possível discutir a probabilidade de similaridade entre os elementos comparados. Ao se observar que o comprimento máximo segue a mesma distribuição de probabilidade da máxima sequência de caras ou coroas consecutivas em experimentos com lançamentos de moeda como feitos por Schilling (Schilling, 1990).

O trabalho de Schilling diz o seguinte: seja n o número de lançamentos de moeda, R_n a maior sequência de caras obtida e $A_n(C)$ o número de sequências com n lançamentos

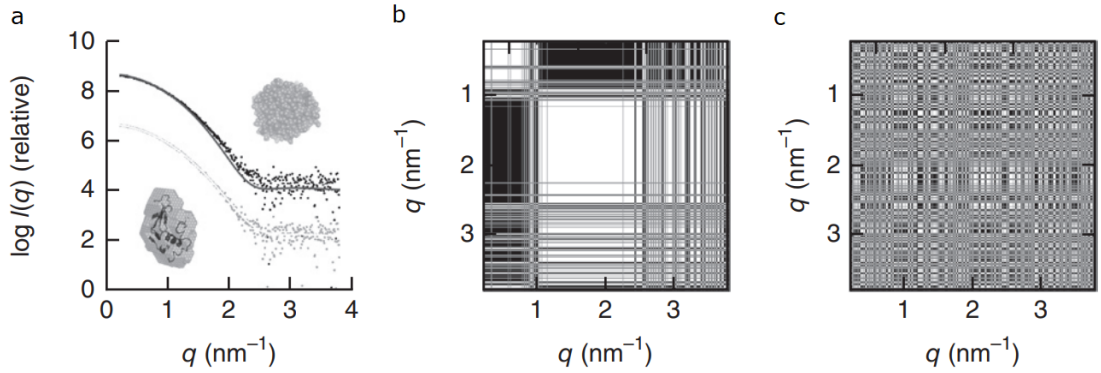

Figura 3.32: Exemplo de uso do CorMap para a avaliação de um ajuste aos dados experimentais: dois ajustes de modelagens diferentes para conjuntos de dados idênticos (preto e cinza claro, apenas diferem por um *offset*) (**a**), CorMap para o ajuste ao conjunto de dados preto, com mapa com características nítidas de desvios sistemáticos de modelagem (**b**) e CorMap para o conjunto cinza, não indicando desvios nítidos (**c**). O CorMap pode ser usado para comparar ajustes a dados experimentais e até mesmo conjuntos de dados diferentes (Franke et al., 2015).

onde o número de caras *não passa de* C. A quantidade $P(R_n > C)$ define a probabilidade de se ter mais do que C caras consecutivas de um total de n lançamentos.

$$P(R_n > C) = 1 - \frac{A_n(C)}{2^n}. \tag{3.114}$$

Para $n \leqslant C$ a probabilidade será sempre nula, de modo que para estes casos teremos necessariamente que $A_n(C) = 2^n$. Já para $n > C$ usemos o exemplo ilustrativo do artigo original para $C = 3$ para entender a ideia. Para que o início da sequência não tenha mais do que 3 caras consecutivas deve-se ter os seguintes resultados (com H sendo cara e T coroa): T, HT, HHT ou HHHT. O resto dos lançamentos, claro, também não deve ter mais do que 3 caras seguidas. Porém pensar assim permite obter uma expressão recursiva para $A_n(C)$ nestes casos, dada por

$$A_n(3) = A_{n-1}(3) + A_{n-2}(3) + A_{n-3}(3) + A_{n-4}(3). \tag{3.115}$$

Utilizando como exemplo, para $n = 4$ todos os termos à direita se encaixam no caso $n \leqslant C$, sendo possível a obtenção de $A_4(3) = 15/16 \approx 0.94$ e, portanto, $P(R_4 > 3) \approx 0.06$. Assim, a chance de se obter uma sequência de mais de 3 caras de um total de 4 lançamentos (neste caso, todos os lançamentos sendo caras) é de aproximadamente 6%. Assim, considerando todos os casos descritos, pode-se escrever $A_n(C)$ como

$$A_n(C) = \begin{cases} \sum_{j=0}^{C} A_{n-1-j}(C) & \text{para } n > C \\ 2^n & \text{para } n \leqslant C \end{cases}. \tag{3.116}$$

A ideia utilizada pelo CorMap é análoga: sendo n o número de pontos experimentais e C a maior sequência de pontos com mesma direção de correlação, a probabilidade P para a qual aleatoriamente se obtêm a maior sequência é calculada. Conforme menor o valor de P, maior a chance da ocorrência de erros sistemáticos no ajuste, e a região de mais pontos com correlação na mesma direção é indicada e colorida pelo programa de acordo com o valor de P de modo a mostrar intuitivamente ao usuário a região onde o ajuste desvia dos dados e quão potencialmente problemática ela pode ser.

O CorMap é uma metodologia alternativa de análise de ajustes baseada em probabilidades para quando as incertezas não são bem determinadas. Porém, mesmo que solidamente fundamentado, o uso de métodos de análise sem incertezas corretamente estimadas, como o CorMap, é potencialmente arriscado, e na tomada de dados deve-se priorizar a obtenção de incertezas que sejam o melhor determinadas quanto possível. Recomenda-se cautela ao utilizar este tipo de metodologia.

3.6 Aquisição de medidas experimentais

3.6.1 Preparação de amostras

As amostras foram produzidas a partir de proteínas expressas pelos membros do Laboratório em Bioquímica e Biofísica de Proteínas (LBBP) do Instituto de Química de São Carlos da Universidade de São Paulo (IQSC-USP), coordenado pelo Prof. Dr. Júlio César Borges, e do laboratório coordenado pelo Prof. Dr. Carlos Henrique Inácio Ramos do Instituto de Química da Universidade Estadual de Campinas (IQ-UNICAMP).

3.6.2 Medidas de SAXS

As medidas de SAXS foram realizadas no Laboratório Nacional de Luz Sincrotron (LNLS), em Campinas, na linha SAXS-1. Em todos os trabalhos a distância entre amostra e detector selecionada foi de $\sim$ 1m, implicando em um vetor de espalhamento de aproximadamente 0.1 a 5.0 nm^{-1}. O porta amostra montado é composto por uma armação de metal onde são posicionadas janelas de mica para passagem dos fótons incidentes e peças de teflon para vedação com capacidade para em torno de 300 μL. Após cada coleta de dados se dava por meio de bombeamento de água Milli-Q no interior com posterior secagem com ar comprimido.

Figura 3.33: Diferentes tipos de porta amostra para uso nas linha de SAXS do LNLS (disponível em **https://www.lnls.cnpem.br/facilities/saxs1**).

Para tomadas de dados sempre, ao trocar de amostra, se fazia uma avaliação medindo 10 quadros de 10 segundos cada para avaliar possível dano por radiação, visível por meio de aumento de $I(0)$ conforme as medidas são executadas. Caso não fosse observada alteração das curvas nos diferentes quadros foram tomadas ou três medidas de 100s cada ou uma de 300s. Concentrações foram definidas de acordo com o esperado para que as proteínas não

apresentem agregação e, ao mesmo tempo, sinal suficientemente para análise dos dados. Foram tomados dados para diferentes valores de concentração para garantir a manutenção de estado oligomérico e otimização da razão sinal-ruído. Os dados experimentais tiveram subtração de tampão, este medido separadamente, e correção de atenuação da amostra feitas automaticamente por *software* do próprio laboratório.

Foram feitos três estudos com diferentes proteínas nas seguintes concentrações, determinadas pelos membros dos laboratórios colaboradores utilizando espectrofotômetros:

- Estudo I - Comparação de diferentes Hsp70 de humanos: Hsp70-1A a 0.5 mg/ml, Hsc70 a 1.0 mg/ml e Bip a 0.8 mg/ml. Tampão TKP. O pIs das Hsp70 são de 5.5, 5.4 e 5.1, respectivamente.

- Estudo II - Comparação da GrpE-L1 de humanos em diferentes condições de solvente: GrpE-L1 a 2.0 mg/ml em tampão, β-mercaptoetanol ou H_2O_2. O pI da GrpE-L1 é de 8.2.

- Estudo III - Determinação estrutural da Hsp90 de *Aedes aegypti* em solução: AaHsp90 a 2.5 mg/ml livre em solução ou com os nucleotídeos AMP, ADP ou ATPγS. Tampão Tris-HCl + NaCl. O pI da AaHsp90 é de 4.9.

As amostras foram medidas em pH fisiológico (em torno de 7) a 25°C, e pureza mínima de 95%.

Capítulo 4

Resultados e Discussões

Foram feitos quatro estudos no presente trabalho. O primeiro é uma comparação estrutural de três membros da família Hsp70 de humanos: Hsp70-1A, Hsc70 (*70-kDa heat shock cognate protein*; ambas de citosol) e Bip (*Binding immunoglobulin protein*, Hsp70 humana de retículo endoplasmático). O segundo é uma análise mais aprofundada da Bip utilizando técnicas mais elaboradas de modelagem de SAXS, incluindo estudos de corpo rígido e flexibilidade. O terceiro busca entender a importância de pontes dissulfeto da proteína GrpE-L1 humana na sua estrutura a partir da análise de dados dela em diferentes solventes. Por fim, o quarto é um estudo da Hsp90 de *Aedes aegypti* (AaHsp90) com diferentes nucleotídeos de adenosina, buscando entender o efeito da ligação destes na estrutura da chaperona.

As incertezas de medida foram calculadas de duas possíveis maneiras: quando utilizando apenas um quadro mais longo foi utilizado o método descrito na seção 3.3.5, enquanto quando mais quadros foram tomados utilizou-se a média e o desvio padrão das intensidades dos quadros para cada q. As análises de Guinier foram feitas utilizando o programa AUTORG do ATSAS e as $p(r)$ geradas a partir do GNOM (tabelas com os critérios perceptuais no Apêndice C). Envelopes foram calculados fazendo uso do DAMMIF para gerar um conjunto de dez a vinte estruturas dos quais foi elaborado um modelo médio utilizando o DAMAVER, com um passo final de refinamento feito com o DAMMIN utilizando o modelo médio como volume de busca inicial. De acordo com cada caso também foram feitas outras análises utilizando programas como o BUNCH e o EOM para modelagem utilizando estruturas parciais disponíveis para os domínios.

4.1 Hsp70: comparação entre três membros da família

O estudo de membros da família das Hsp70 buscou comparar diferenças estruturais entre elas. Os dados estão apresentados na Figura 4.1 juntamente com os ajustes do GNOM relacionados à $p(r)$ e os gráficos de Guinier em *inset*, com os parâmetros estruturais obtidos pela linearização apresentados na Tabela 4.1. Os dados para a Hsp70-1A apresentam maior ruído devido a menor concentração, o que se traduz em maior incerteza dos parâmetros obtidos, especialmente nos relacionados à análise de Guinier. Ao mesmo tempo, a incerteza também pode ser devido a algum grau de agregação nos cinco primeiros pontos com certa linearidade, o que faz com que o programa não se decida entre ambos os regimes lineares e aumente a média e a incerteza atribuídos ao R_g.

A partir de $I(0)$ pode-se calcular as MM por diversos métodos, dependentes e independentes das concentrações de proteína, e os valores estão apresentados na Tabela 4.2, sendo estes condizentes com as Hsp70 em estado monomérico. As concentrações foram de 0.5 mg/ml para a Hsp70-1A, 1.0 mg/ml para a Hsc70 e 0.8 mg/ml para a Bip.

Tabela 4.1 - **Parâmetros estruturais das Hsp70 por Guinier**

Proteína	R_g (nm)	$I(0)$ (10^5 UA)	qR_g
Hsp70-1A	4.38 ± 0.73	6.8 ± 0.1	1.30
Hsc70	3.60 ± 0.11	11.3 ± 0.1	1.30
Bip	3.50 ± 0.06	9.6 ± 0.1	1.29

Incertezas dadas pelo AUTORG (seção 3.4.1).

Tabela 4.2 - **Valores de MM para as Hsp70 por diferentes métodos**

Proteína	*Água*	*SAXSMoW*	Q_p	*MoW*	V_c	*S&S*	*Bayes*	***Teórica***
Hsp70-1A	81.3	110.9	100.9	77.4	81.9	94.7	94.2	**70.1**
Hsc70	71.1	80.5	70.7	70.1	65.3	81.5	67.1	**70.9**
Bip	74.1	84.7	75.9	77.5	70.6	84.4	74.3	**72.3**

Metodologias descritas na seção 3.5. Massa teórica calculada pela sequência de aminoácidos utilizando o programa ProtParam (Gasteiger et al., 2005). Valores em kDa.

As $p(r)$ indicam um formato elongado para as proteínas, o que é condizente com o esperado pelo modelo elaborado, de que os domínios se encontram independentes uns dos outros, de modo que em média o formato seja bastante elongado. Os critérios do

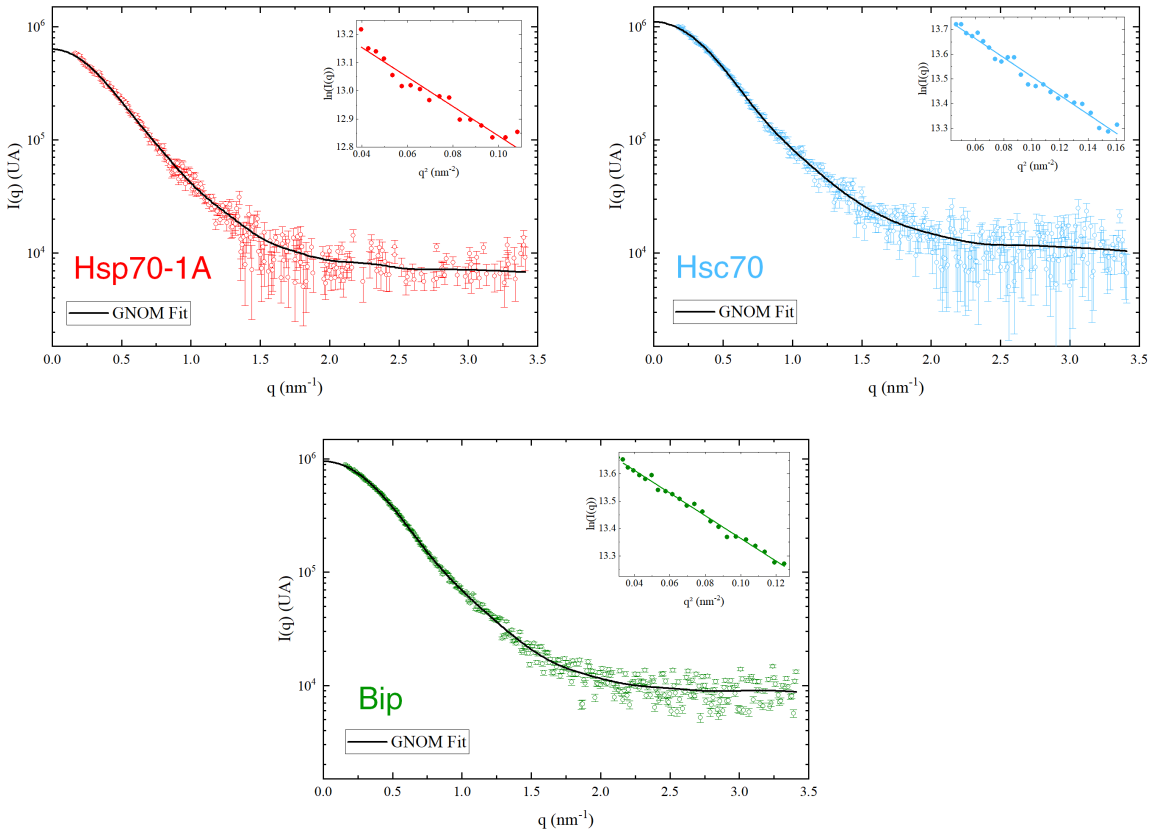

Figura 4.1: Dados de SAXS com gráficos de Guinier em *inset* para as proteínas Hsp70-1A, Hsc70 e Bip.

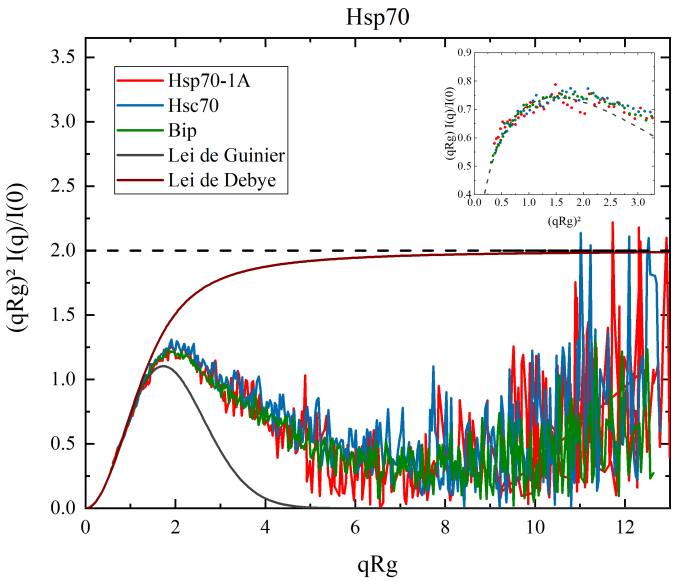

Figura 4.2: Gráfico de Kratky adimensional com GPA no *inset* para as Hsp70, indicando que as proteínas são bastante globulares com alguma flexibilidade entre os domínios.

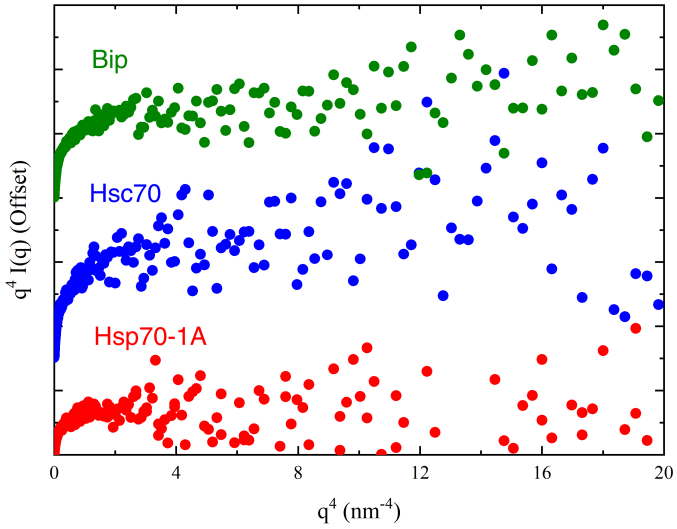

Figura 4.3: Gráficos de Porod-Debye para cada Hsp70. Há aparente presença de platô apesar de alto ruído, indicando rigidez estrutural.

GNOM para as $p(r)$ estão presentes na Tabela C.1 (Vide Apêndice C). Para verificação dos parâmetros obtidos pela análise de Guinier, a Tabela 4.3 apresenta os encontrados a partir da integração da $p(r)$, assim como a MM calculada usando água como padrão de calibração para o valor de $I(0)$ encontrada por este método.

Pode-se constatar que R_g e $I(0)$ calculados pela linearização da região de Guinier e pela integração da $p(r)$ são compatíveis para ambos os métodos dentro de duas incertezas. Porém a MM calculada para a Hsp70-1A mostra um aumento em praticamente todos os métodos por conta do alto valor de $I(0)$ encontrado pelo método de Guinier, o que implica que haja alguma possível agregação na curva que aparece especialmente para ângulos mais baixos, mas ao fazer a análise pela $p(r)$, que utiliza dados da curva inteira, o R_g mostra-se mais próximo das outras duas e a MM torna-se mais condizente com o valor de referência uma vez que o cálculo de parâmetros faz uso de mais dados do que a análise de Guinier.

É interessante comparar as funções $p(r)$ encontradas para as Hsp70 humanas com as referências encontradas na literatura. Foram encontrados dois estudos de SAXS para Hsp70: um de DnaK (Hsp70 de *E. coli*) (Shi et al., 1996) e um de Hsc70 bovina (Wilbanks et al., 1995). Nesses artigos foram estudadas conformações com ADP e ATP para as chaperonas, e pelo modelo atual para Hsp70, esperamos que a $p(r)$ encontrada para ADP ligado seja mais próximo ao encontrado em solução. Como ilustrado na Figura 4.5, as $p(r)$ encontradas nos artigos são similares entre si e apresentam um perfil mais próximo a uma

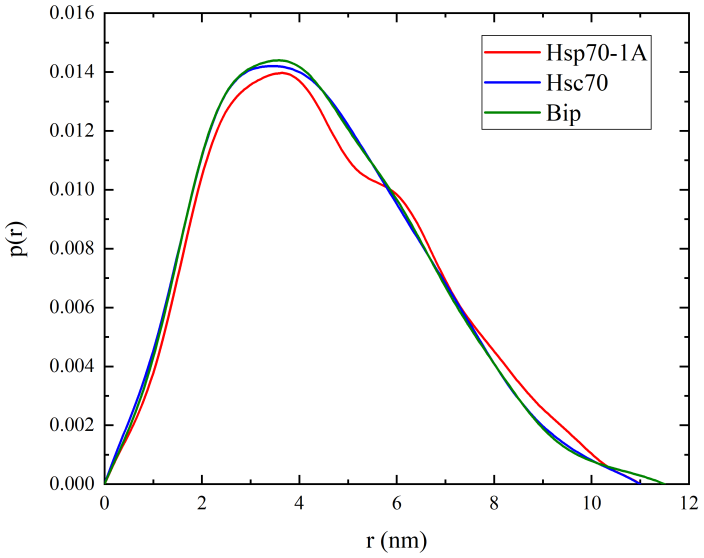

Figura 4.4: Função $p(r)$ gerada pelo GNOM para as diferentes amostras.

topologia mais elongada, como um haltere ou um elipsoide, mesmo apresentando o pico na mesma região e valores próximos de D_{max} aos obtidos para as ortólogas humanas. Uma vez que os perfis são notavelmente diferentes das obtidas para os casos de humanos, este resultado pode indicar que: (i) as Hsp70 humanas diferem consideravelmente de homólogas de outras espécies, mesmo que para um organismo evolucionariamente mais próximo, é o caso da Hsc70 bovina; (ii) a conformação ligada a ADP não corresponde à de uma Hsp70 livre em solução.

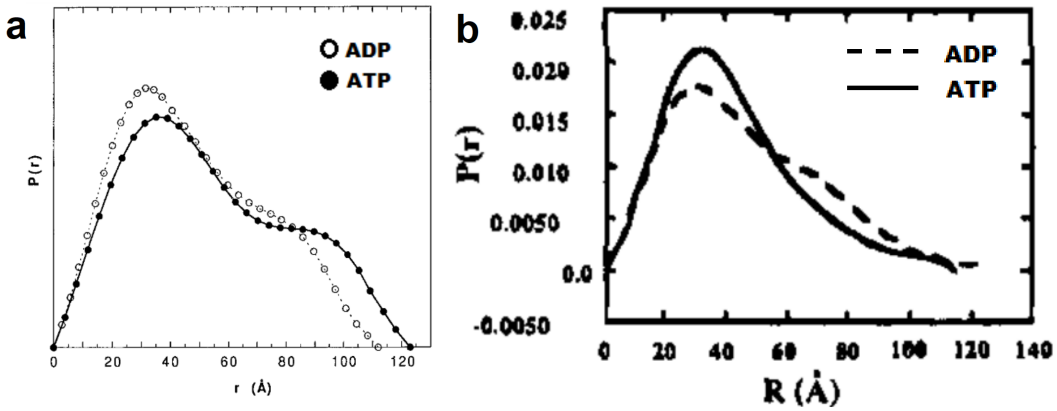

Figura 4.5: Funções $p(r)$ obtidas anteriormente para (**a**) Hsp70 de *E. coli* (DnaK) (adaptado de Shi et al., 1996) e (**b**) Hsc70 bovina (adaptado de Wilbanks et al., 1995) ligadas a ADP ou ATP.

Foi mostrado, assim, que não há variação estrutural a baixa resolução entre membros de

Tabela 4.3 - **Parâmetros estruturais obtidos pela $p(r)$ para as Hsp70**

Proteína	R_g (nm)	$I(0)$ (10^5 UA)	MM (kDa)	***Teórica***	D_{max} (nm)	ER
Hsp70-1A	3.62 ± 0.01	6.04 ± 0.01	71.8	**70.1**	11.0	4.17
Hsc70	3.45 ± 0.02	10.93 ± 0.07	68.9	**70.9**	11.0	4.11
Bip	3.46 ± 0.01	9.49 ± 0.02	72.9	**72.3**	11.5	3.77

Incertezas calculadas pelo GNOM, MM calculada usando água para calibração (seção 3.5.1.2) e ER calculado de acordo com a seção 3.3.2.

Tabela 4.4 - **Parâmetros dos programas SHANUM e AMBIMETER para as Hsp70**

Proteína	N_{op}	q_{op}	$a - score$
Hsp70-1A	12	3.43	2.44
Hsc70	14	4.00	2.53
Bip	18	4.92	2.38

N_{op} e q_{op} calculados de acordo com a seção 3.4.7 e $a - score$ de acordo com a seção 3.4.4.

Hsp70 humana em citosol e retículo endoplasmático, o que aponta que as particularidades de cada proteína com relação a proteínas-cliente, por exemplo, se deve a diferenças mais sutis, que não podem ser medidas a baixa resolução por SAXS. Devido à maior razão sinal-ruído para os dados da Bip em relação aos da Hsc70 (e à baixa qualidade dos dados da Hsp70-1A), a Bip foi utilizada para análises posteriores para modelagem.

4.2 Estudos estruturais da Bip

Como primeira forma de modelagem, para obter um perfil de baixa resolução para a Bip foi feita a modelagem *ab initio* utilizando um modelo de esferas. Uma vez que ao executar o AMBIMETER foi obtido $a - score = 2.38$ e pelo perfil da $p(r)$ ser comum para corpos prolatos, as modelagens se deram impondo anisometria *prolata*.

O programa DAMMIF foi executado 20 vezes e a estrutura média foi obtida por meio do DAMAVER, com refino subsequente fazendo uso do DAMMIN (análise automatizada na interface gráfica da suíte ATSAS). A estrutura com o respectivo ajuste estão presentes na Figura 4.6. O DAMAVER obteve $NSD = 0.986 \pm 0.054$.

Recentemente foram publicadas estruturas cristalográficas da Bip ligada a ATP, a entrada 5E84 do PDB uma dela completa (Yang et al., 2015). Como exercício, essa estrutura foi comparada os dados de SAXS e o envelope obtido. A curva teórica de espalhamento

dessa estrutura foi calculada com o CRYSOL e comparada aos dados experimentais a fim de avaliar o grau de similaridade entre a Bip em seu estado *Apo* e a ligada a ATP, caso a estrutura de cristal realmente ocorra em solução.

Na Figura 4.6 está a tentativa de ajuste da curva teórica aos dados experimentais feito pelo CRYSOL, e é nítido que não há acordo entre ambas. Isso leva a dois possíveis cenários: i) a estrutura da Bip em ATP é muito diferente da livre em solução, ou ii) a estrutura cristalográfica da Bip obtida não é representativa de como ela se encontra em solução e, portanto, uma comparação não é possível.

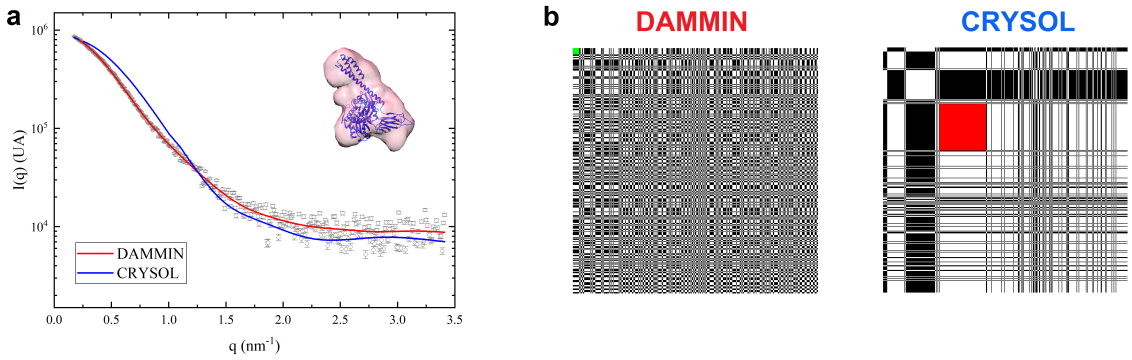

Figura 4.6: Dados experimentais da Bip com ajustes resultantes dos modelos cristalográfico (em azul, calculado pelo CRYSOL) e *ab initio* (em vermelho, calculado pelo DAMMIN) (**a**), e os respectivos CorMap referentes a cada modelagem, indicando o quanto o modelo cristalográfico se desvia do obtido em solução, possivelmente devido a questões de flexibilidade (**b**).

Outra abordagem possível é a modelagem híbrida. Foram utilizadas estruturas cristalográficas para cada domínio (Wisniewska et al., 2010; Yang et al., 2015) (Figura 4.7) e empregado o BUNCH para buscar ajustá-las aos dados incluindo modelagem da conexão entre o NBD e o SBD, incluindo flexibilidade entre os SBDα e SBDβ ao subdividir o segundo arquivo PDB. Este cenário busca avaliar se a baixa resolução pode-se observar a "pinça" do SBD abrindo e fechando para a associação à proteína-cliente.

Para avaliar se o modelo de corpo rígido ainda descreve a partícula considerando a flexibilidade constatada no gráfico de Kratky foi realizado um estudo com o EOM para comparação. Para esta modelagem foram utilizadas as mesmas estruturas de referência que para o BUNCH, e os laços foram gerados sem a imposição de restrições.

Os ajustes para ambas as modelagens estão na Figura 4.8 com os respectivos CorMap. A modelagem do BUNCH manteve os domínios NBD e SBD próximos um ao outro, porém com os SBDα e SBDβ pouco mais afastados que na estrutura de referência. Ao observar

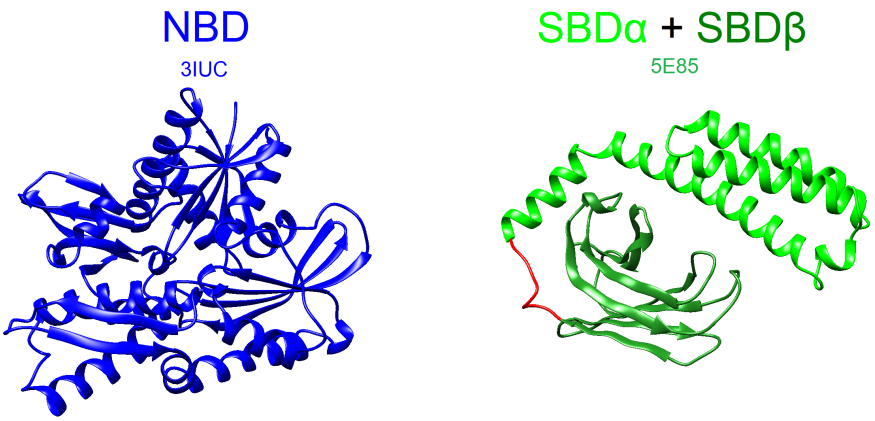

Figura 4.7: Estruturas cristalográficas parciais para os domínios da Bip a serem usadas nos processos de modelagem (Wisniewska et al., 2010; Yang et al., 2015). A região em vermelho foi retirada da estrutura para que fosse modelada *ab initio*.

o CorMap para a modelagem do EOM, constata-se um desvio importante na região de baixíssimo ângulo, o que indica desvios consideráveis no espaço real. Portanto, a análise de *ensemble* deve ser tratada com cuidado. O que foi observado, como constatado nos modelos selecionados (Figura 4.9) e na Tabela 4.5, são estruturas mais compactas. Pelos histogramas de R_g e D_{max} (Figura 4.10), comparando a *pool* ao *ensemble* de estruturas geradas pode-se notar que, de fato, não se observa flexibilidade alta da estrutura, inclusive com esta sendo razoavelmente compacta e rígida ao observar as distribuições. Os valores obtidos para R_{flex} e R_σ foram 54.9% (comparado a 85.0% da pool) e 0.24, respectivamente, o que indica alguma flexibilidade. Este resultado é condizente com o modelo de domínios bem enovelados com *linkers* flexíveis. Porém, ao observar com cuidado os valores de R_g do *ensemble* é que não são compatíveis com os valores obtidos por outras análises.

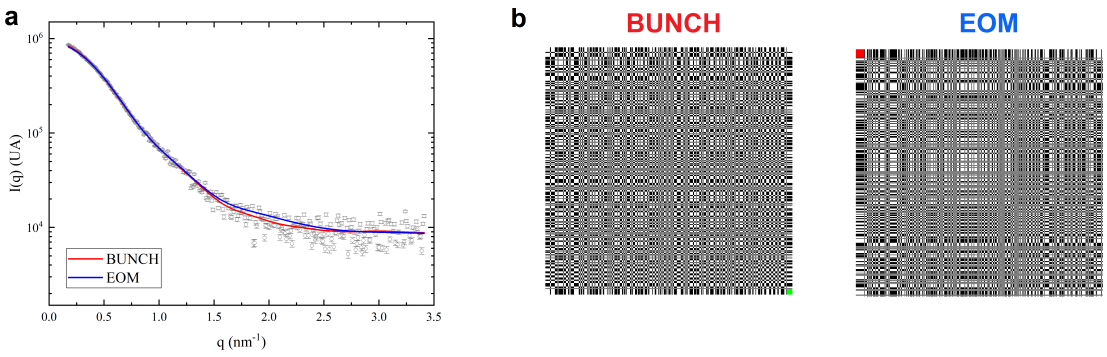

Figura 4.8: Comparação dos ajustes dos modelos gerados pelo BUNCH (vermelho) e pelo EOM (azul) a partir da plotagem com os dados experimentais (**a**) e dos CorMap (**b**).

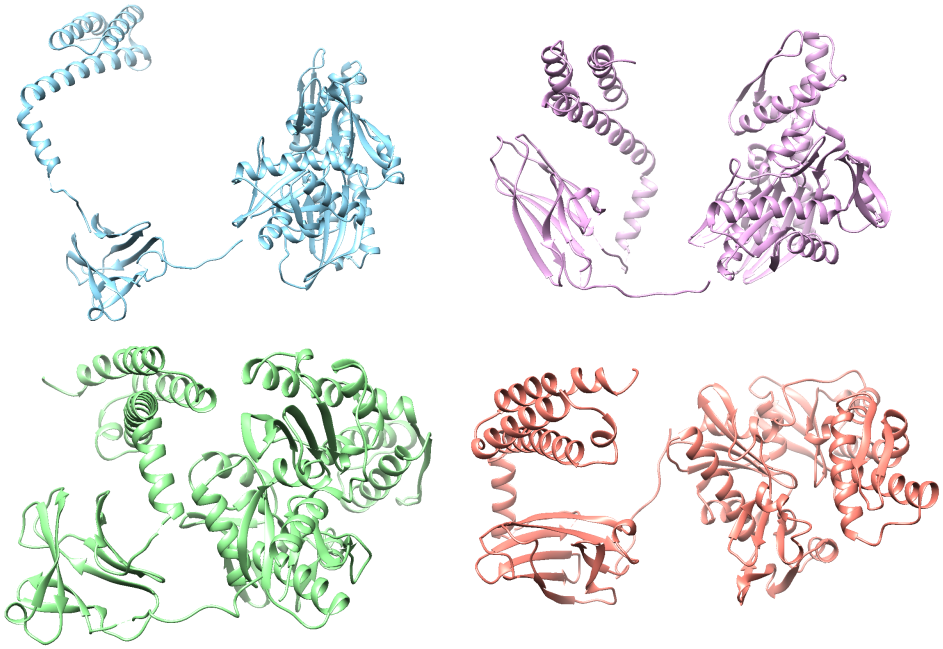

Figura 4.9: Ensemble de estruturas selecionadas a partir de um *pool* de 10 mil estruturas pelo EOM.

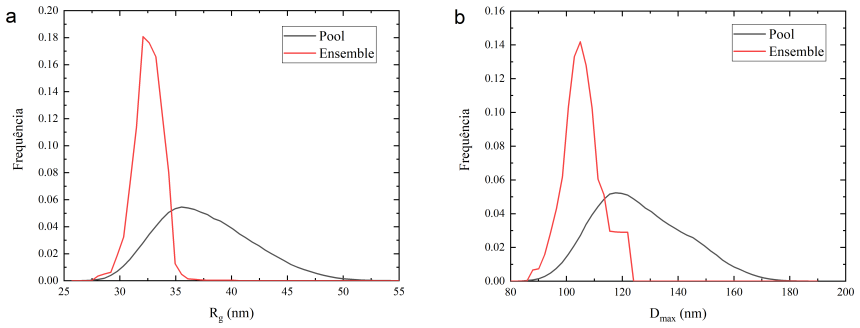

Figura 4.10: Distribuição de R_g (**a**) e D_{max} (**b**) para o *pool* inicial (preto) e para ensemble de estruturas selecionadas (vermelho).

Tabela 4.5 - R_g**,** D_{max} **e frações populacionais de *ensemble* calculadas pelo EOM**

Estrutura	R_g	D_{max}	*Fração populacional*
Azul	3.90	11.2	18.2%
Rosa	3.27	9.9	36.4%
Verde	2.94	9.8	36.4%
Laranja	3.47	10.5	9.1%
ENSEMBLE FINAL	**3.28**	**10.2**	**100%**

R_g e D_{max} das conformações representativas do *ensemble* selecionado. Os valores são mais baixos que os obtidos tanto pela linearização da região de Guinier quanto pela $p(r)$.

A Tabela 4.6 contém os parâmetros de ajuste para cada método de análise estrutural e a Figura 4.11 contém o envelope obtido pelo DAMMIN e o modelo de corpo rígido resultado do BUNCH alinhados para comparação. Interessante notar que o valor de χ^2_{red} para o obtido pelo DAMMIN é menor, porém que o valor de p para o modelo do BUNCH é maior. Ou seja, o envelope contém menor discrepância com os dados, porém o modelo de corpo rígido apresenta menor probabilidade de haverem erros sistemáticos e ao mesmo tempo indica desvio na região de maiores valores de q, o que implica que as diferenças são relativas a distâncias menores no espaço real.

Tabela 4.6 - **Parâmetros de avaliação de ajuste da Bip para cada metodologia**

Modelagem	χ^2_{red}	p
CRYSOL	72.9	0.00
DAMMIN	14.0	0.49
BUNCH	15.1	0.76
EOM	22.06	0.01

Valores de p calculados de acordo com a seção 3.5.5.2.

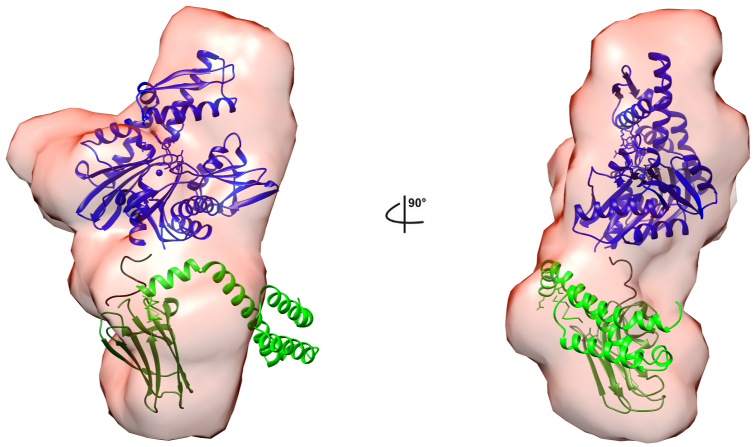

Figura 4.11: Sobreposição dos modelos gerados *ab initio* com o DAMMIN e por modelagem híbrida com o BUNCH.

O modelo para a Bip, portanto, apresenta baixa flexibilidade mesmo com a ligação entre os domínios tendo comprimento apreciável, o que indica que ele possivelmente interage bastante com os próprios domínios e não tanto com o solvente, o que implicaria na observação de maior flexibilidade e maior movimento entre os domínios com aumento artificial do volume do envelope calculado como reflexo de alteração no perfil de Kratky.

Também foi mostrado que há diferenças importantes entre a estrutura da Bip medida em solução e em cristal, possivelmente devido à baixa, porém sensível, flexibilidade entre os domínios ou a interações artificiais nas estruturas dos próprios domínios devido ao processo de cristalização da proteína.

4.3 GrpE

O estudo das GrpE foi feito visando identificar mudanças estruturais com mudança de solvente, o que pode favorecer ou desfavorecer a formação de pontes dissulfeto. Para as GrpE, surgiram questões relacionadas à agregação para as amostras de GrpE-L2 em todos os solventes, portanto apenas se comparou a GrpE-L1 em tampão, β-mercaptoetanol e H_2O_2. Os dados a 2.0 mg/ml estão presentes na Figura 4.12, juntamente com os gráficos de Guinier (em *inset*) e ajustes do programa GNOM, usado para gerar as funções $p(r)$.

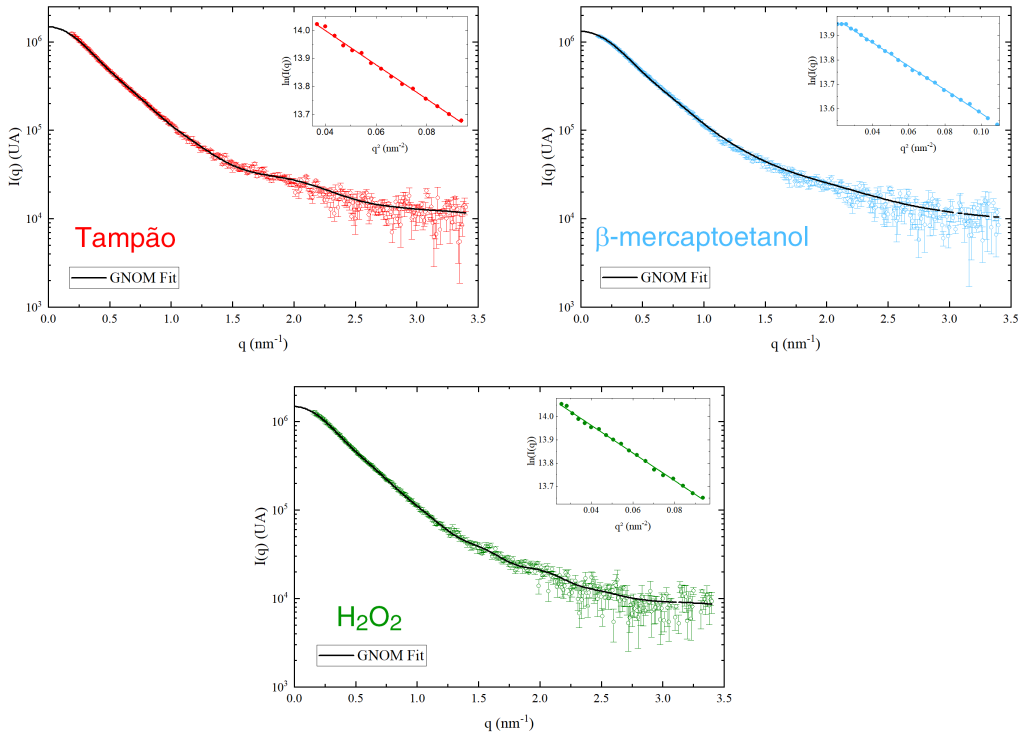

Figura 4.12: Dados para a GrpE-L1 em diferentes solventes com gráficos de Guinier em *inset*: tampão (**a**), β-mercaptoetanol (**b**) e H_2O_2 (**c**).

Os gráficos de Kratky indicam bastante flexibilidade na estrutura, assim como os gráficos de Porod-Debye (que não contém platô), e estão indicados nas Figuras 4.13 e

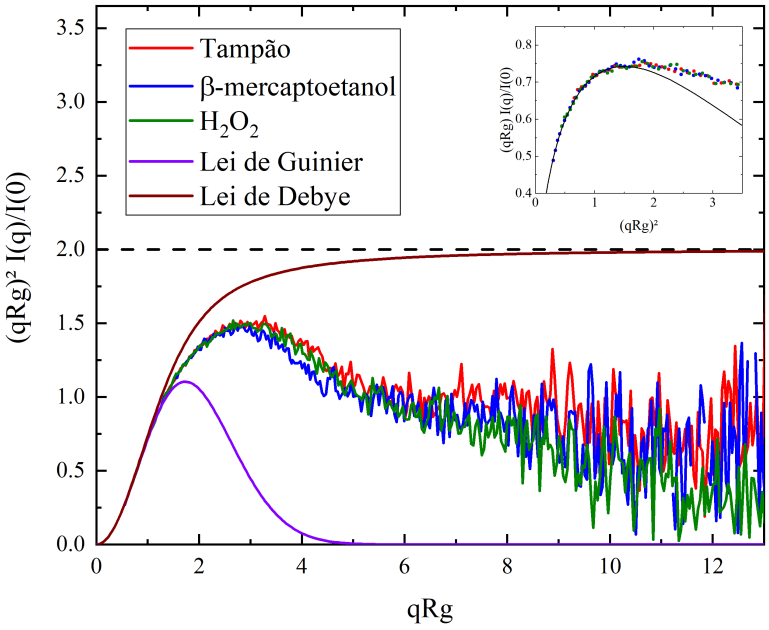

Figura 4.13: Gráfico de Kratky adimensional para a GrpE-L1 em diferentes solventes, indicando alta flexibilidade, e de GPA no *inset*, validando os parâmetros extraídos pela análise de Guinier.

4.14, respectivamente. A análise de Guinier mostra que ambas as proteínas apresentam variação sensível de R_g (Tabela 4.7), e a linearização é validada pelo gráfico de GPA (Figura 4.13, *inset*). Os valores de MM (Tabela 4.8) são condizentes com dímeros para todos os métodos utilizados. Uma vez que os perfis de Kratky indicam alta flexibilidade, o valor de MM obtido por metodologias que consistem no cálculo do invariante de Porod Q (como SAXSMoW) devem ser considerados com cuidado. De maneira similar, o alto grau de elongamento também prejudica o cálculo da MM por métodos independentes de concentração, o que torna o método de cálculo com água como padrão o mais indicado para determinar a MM, o que é confirmado pela maior proximidade dos valores calculados ao de referência. Curiosamente os métodos que utilizam o gráfico de Kratky mostraram resultados menos discrepantes para a proteína em β-mercaptoetanol, mesmo que a diferença de perfil de Kratky desta amostra para as outras não seja grande. O gráfico de Porod-Debye indica alta flexibilidade devido à ausência de platô, como observado na Figura 4.14.

As $p(r)$ para todas as amostras são bastante similares, com forma bastante elongada, como mostrado na Figura 4.15. Isso é esperado e condizente com a estrutura cristalográfica e estudo anterior de GrpE por SAXS (Nakamura et al., 2010; Borges et al., 2003). Os

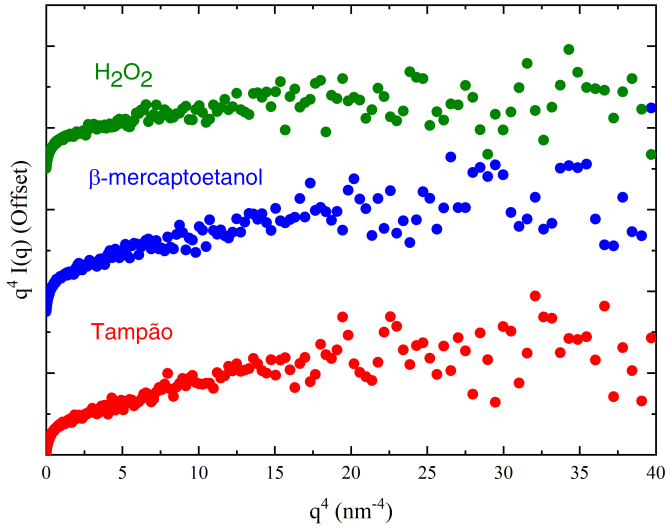

Figura 4.14: Gráfico de Porod-Debye para a GrpE-L1 em diferentes solventes, mostrando alto grau de flexibilidade por conta da ausência de platô.

Tabela 4.7 - **Parâmetros estruturais da GrpE-L1 por Guinier**

Solvente	R_g (nm)	$I(0)$ (10^6 UA)	qR_g
Tampão	4.23 ± 0.26	1.52 ± 0.01	1.29
β-mercaptoetanol	3.82 ± 0.05	1.28 ± 0.01	1.26
H_2O_2	4.17 ± 0.11	1.46 ± 0.01	1.28

Incertezas dadas pelo AUTORG (seção 3.4.1).

Tabela 4.8 - **Valores de MM da GrpE-L1 por diferentes métodos**

Solvente	*Água*	*SAXSMoW*	Q_p	*MoW*	V_c	*S&S*	*Bayes*	***Teórica***
Tampão	54.3	71.3	69.4	52.7	57.1	90.1	53.2	48.6
β-mercaptoetanol	45.9	59.2	54.9	50.6	49.0	75.6	48.7	48.6
H_2O_2	52.2	79.7	69.8	72.7	63.4	78.0	68.8	48.6

Metodologias descritas na seção 3.5. Massa teórica calculada pela sequência de aminoácidos utilizando o programa ProtParam (Gasteiger et al., 2005). A referência é da GrpE-L1 em estado dimérico. Valores em kDa.

critérios de avaliação do GNOM estão presentes na Tabela C.2 e os parâmetros estruturais obtidos pela $p(r)$ se encontram na Tabela 4.7.

Para avaliar como fazer a modelagem, foram utilizados os programas SHANUM para determinar o intervalo de dados úteis para modelagem e AMBIMETER para determinar o grau de ambiguidade das reconstruções tridimensionais *ab initio*. O SHANUM validou o

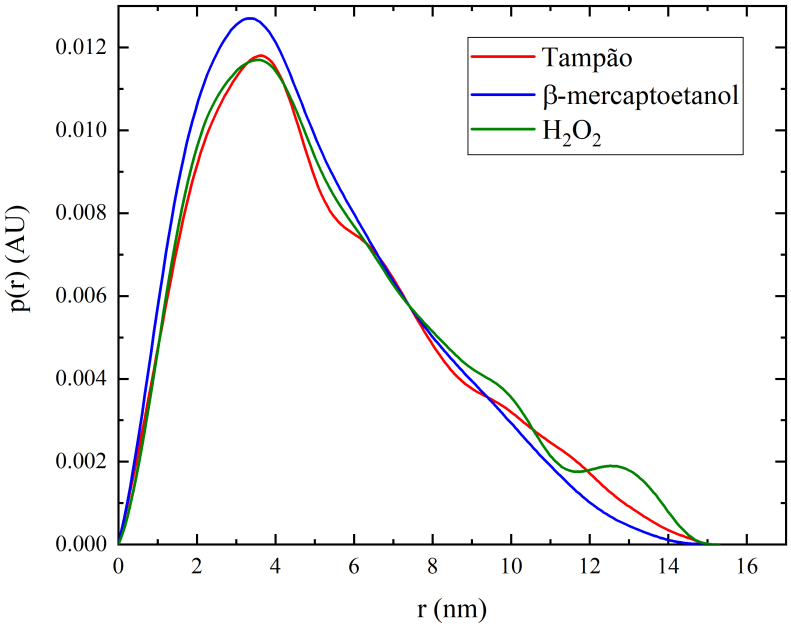

Figura 4.15: Funções $p(r)$ calculadas com o GNOM para a GrpE-L1, apresentando um caráter bastante elongado.

Tabela 4.9 - **Parâmetros estruturais obtidos pela $p(r)$ para a GrpE-L1**

Solvente	R_g (nm)	$I(0)$ (10^6 UA)	MM (kDa)	***Teórica***	D_{max} (nm)	ER
Tampão	4.30 ± 0.03	1.471 ± 0.012	52.6	**48.6**	14.9	1.93
β-mercaptoetanol	4.02 ± 0.01	1.305 ± 0.003	46.7	**48.6**	15.0	2.05
H_2O_2	4.44 ± 0.03	1.469 ± 0.001	52.6	**48.6**	15.3	-

Incertezas calculadas pelo GNOM, MM calculada usando água para calibração (3.5.1.2) e ER calculado de acordo com a seção 3.3.2, exceto para a amostra com H_2O_2 devido a oscilações e aumento próximo a D_{max}.

intervalo de q utilizado para o cálculo das $p(r)$ ($q \approx 3.5nm^{-1}$), permitindo o uso do mesmo intervalo para a modelagem tridimensional, e o AMBIMETER trouxe valores próximos de 2, o que traz a recomendação de uso de programas como o DAMCLUST ou imposição de simetria e/ou anisometria.

As estruturas a baixa resolução foram calculadas utilizando o DAMMIF com simetria prolata, seguido de filtragem pelo DAMAVER e refinamento pelo DAMMIN. Os envelopes *ab initio* calculados indicam bastante similaridade com a estrutura cristalográfica de *Thermus thermophilus* (entrada 3A6M do PDB) (Nakamura et al., 2009) ao se fazer o alinhamento utilizando o SUPCOMB. As Figuras 4.17, 4.19 e 4.21 mostram o modelo filtrado

pelo DAMFILT sobreposto a todos os modelos gerados pelo DAMMIF e o modelo refinado do DAMMIN alinhado à estrutura cristalográfica de referência. Os ajustes e CorMap estão presentes nas Figuras 4.16, 4.18 e 4.20, indicando potenciais desvios sistemáticos em regiões de q suficientemente baixo para que haja efeito apreciável no envelope obtido apenas para o caso da amostra em tampão. A Tabela 4.11 traz os parâmetros de ajuste, sendo um o χ^2_{red} e o outro o valor p determinado a partir do CorMap, e dispersão dos modelos gerados utilizando o NSD do DAMAVER.

Tabela 4.10 - **Parâmetros para modelagem por SHANUM e AMBIMETER para a GrpE-L1**

Solvente	N_{op}	q_{op}	$a-score$
Tampão	23	4.85	2.07
β-mercaptoetanol	24	5.03	2.29
H_2O_2	17	3.49	1.93

N_{op} e q_{op} calculados de acordo com a seção 3.4.7 e $a-score$ de acordo com a seção 3.4.4.

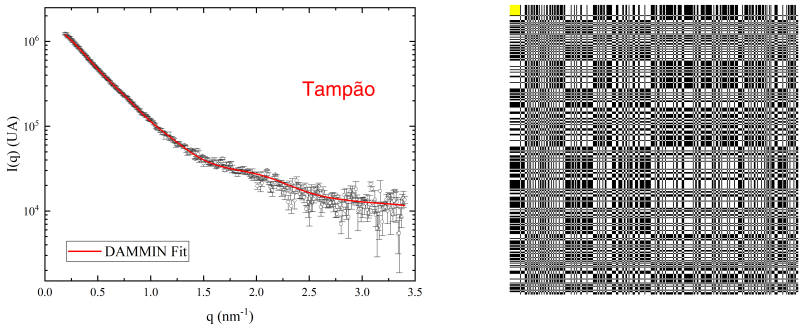

Figura 4.16: Ajuste da modelagem *ab initio* aos dados da GrpE-L1 em tampão, com o respectivo CorMap.

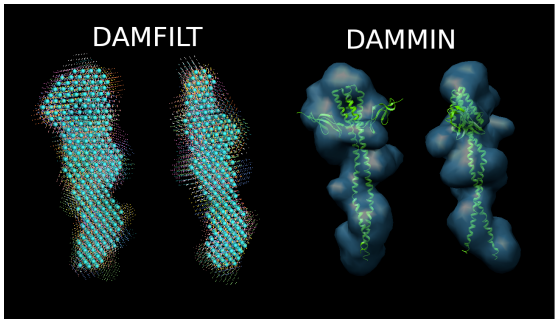

Figura 4.17: Modelo filtrado pelo DAMFILT (esferas) a partir do alinhamento dos modelos obtidos pelo DAMMIF (pontos, cada modelo de uma cor) e modelo final obtido pelo refinamento do DAMMIN (superfície) alinhado à estrutura cristalográfica 3A6M em tampão.

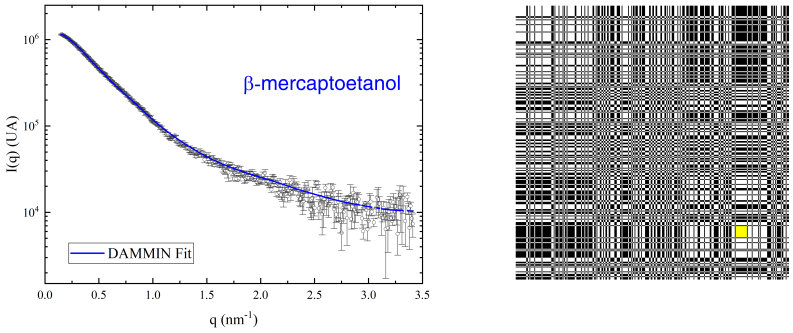

Figura 4.18: Ajuste da modelagem *ab initio* aos dados da GrpE-L1 em β-mercaptoetanol, com o respectivo CorMap.

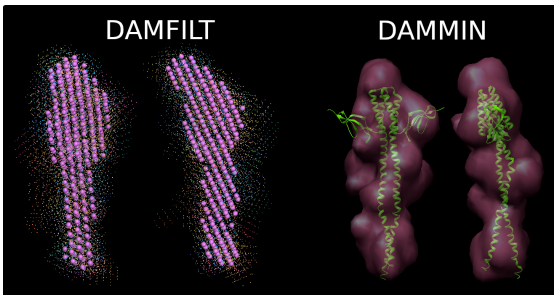

Figura 4.19: Modelo filtrado pelo DAMFILT (esferas) a partir do alinhamento dos modelos obtidos pelo DAMMIF (pontos, cada modelo de uma cor) e modelo final obtido pelo refinamento do DAMMIN (superfície) alinhado à estrutura cristalográfica 3A6M em β-mercaptoetanol.

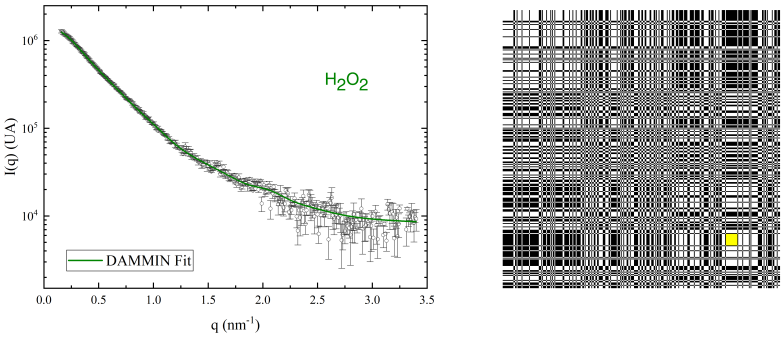

Figura 4.20: Ajuste da modelagem *ab initio* aos dados da GrpE-L1 em H_2O_2, com o respectivo CorMap.

Os resultados apontam que o rompimento das pontes dissulfeto em β-mercaptoetanol não tem efeito apreciável na estrutura, com envelope calculado similar ao das outras proteínas, e a análise de Guinier aponta um R_g compatível por teste-Z com o de outros solventes, implicando que essas interações não parecem ser fatores preponderantes para a manutenção da estrutura de coiled-coil encontrada entre as cadeias.

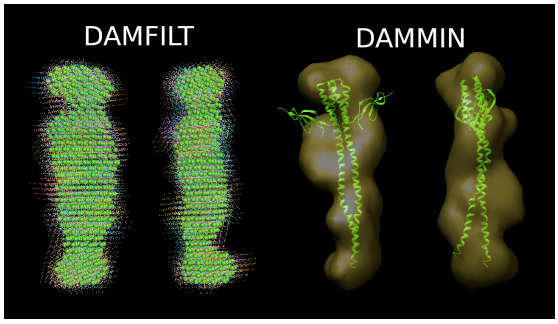

Figura 4.21: Modelo filtrado pelo DAMFILT (esferas) a partir do alinhamento dos modelos obtidos pelo DAMMIF (pontos, cada modelo de uma cor) e modelo final obtido pelo refinamento do DAMMIN (superfície) alinhado à estrutura cristalográfica 3A6M em H_2O_2.

Tabela 4.11 - **Parâmetros de ajuste e de reconstrução *ab initio* para a GrpE-L1**

Solvente	χ^2_{red}	p	*NSD*
Tampão	5.30	0.04	1.10 ± 0.05
β-mercaptoetanol	7.51	0.00	0.95 ± 0.10
H_2O_2	6.19	0.02	0.94 ± 0.03

Valores de p calculados de acordo com a seção 3.5.5.2.

4.4 AaHsp90

O objetivo do estudo de SAXS da AaHsp90 foi obter a estrutura a baixa resolução da chaperona, avaliando como essa informação se relaciona com informações derivadas por outras técnicas, como microscopia eletrônica (Southworth e Agard, 2008), que indicou diferentes equilíbrios entre estruturas abertas e fechadas com cada nucleotídeo a depender da espécie, indicando que co-chaperonas podem estar mais fortemente associadas à atividade chaperona.

Os dados coletados estão presentes para os diferentes estados, todos a 2.5 mg/ml, com ajustes do GNOM, na Figura 4.22, mostrando pouca diferença estrutural entre as formas ligadas a AMP, ADP, ATPγS, ou na mesmo forma apo (ou seja, sem nucleotídeos), como ilustrado no gráfico com *offset* da Figura 4.23. Os dados são a média com desvio padrão da tomada de 10 quadros de 10s. Os gráficos de Guinier, nos insets da Figura 4.22 mostram também muita similaridade entre si, o que se percebe nos parâmetros de saída da linearização do gráfico (Tabela 4.12), e a análise de Guinier é validada pelo gráfico de GPA, no *inset* da Figura 4.24. As MM mostram que, dentro do erro esperado para SAXS (de aproximadamente 10%), a proteína, como esperado pela literatura, se encontra como

dímero em solução (Tabela 4.13).

Os perfis de Kratky (Figura 4.24) mostram que a AaHsp90 apresenta certa flexibilidade, porém seus domínios são bem enovelados. Os perfis de Porod-Debye, encontrados na Figura 4.25, também apontam para considerável rigidez estrutural uma vez que apresentam um platô bem definido, indicando boa descrição da estrutura pelo modelo da lei de Porod.

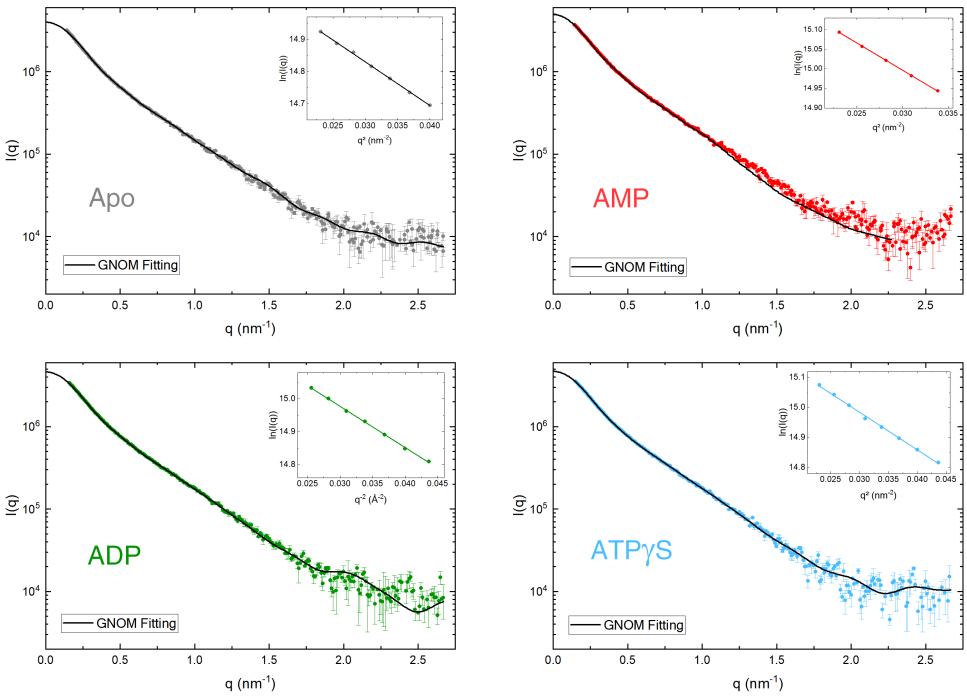

Figura 4.22: Dados referentes à AaHsp90 ligada a diferentes nucleotídeos de adenosina, com gráficos de Guinier em *inset* e ajustes do GNOM.

Tabela 4.12 - **Parâmetros estruturais da AaHsp90 por Guinier**

Nucleotídeo	R_g (nm)	$I(0)$ (10^6 UA)	qR_g
Apo	6.36 ± 0.11	4.13 ± 0.02	1.27
AMP	6.51 ± 0.26	4.97 ± 0.03	1.20
ADP	6.11 ± 0.22	4.64 ± 0.03	1.28
ATPγS	6.17 ± 0.26	4.71 ± 0.02	1.29

Incertezas dadas pelo AUTORG (seção 3.4.1).

Uma vez que as funções $p(r)$ mostraram certa instabilidade para maiores valores de q_{max} para uso no GNOM, foi utilizado o programa SHANUM para estabelecer o intervalo

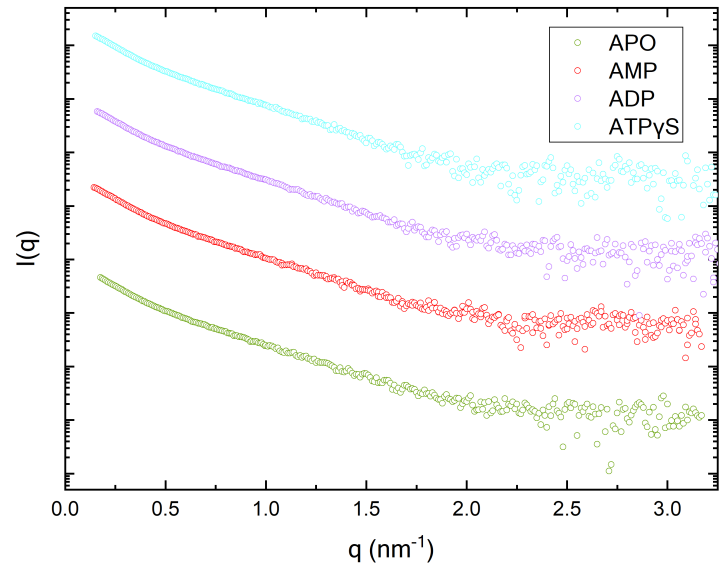

Figura 4.23: Gráfico de offset com dados da AaHsp90 com diferentes nucleotídeos. Não há variação estrutural aparente para nenhum dos nucleotídeos.

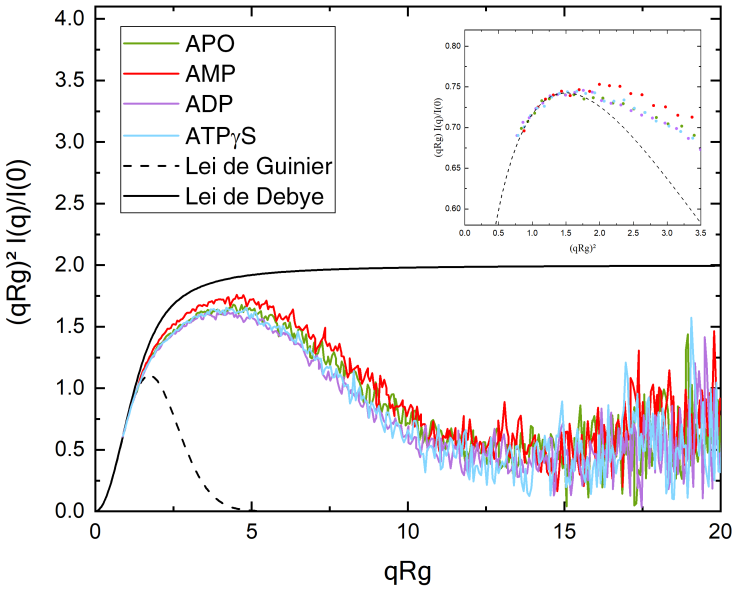

Figura 4.24: Gráficos de Kratky adimensional indicando domínios bem enovelados com grau de flexibilidade considerável. No *inset*, o gráfico de GPA, validando a análise de Guinier feita para a AaHsp90.

apropriado de q para o cálculo da $p(r)$. Os ajustes estão presentes na Figura 4.22 e as $p(r)$ normalizadas estão na Figura 4.26. Os perfis calculados apresentam um caráter bastante elongado, com formas e valores de D_{max} similares, como esperado por terem curvas similares

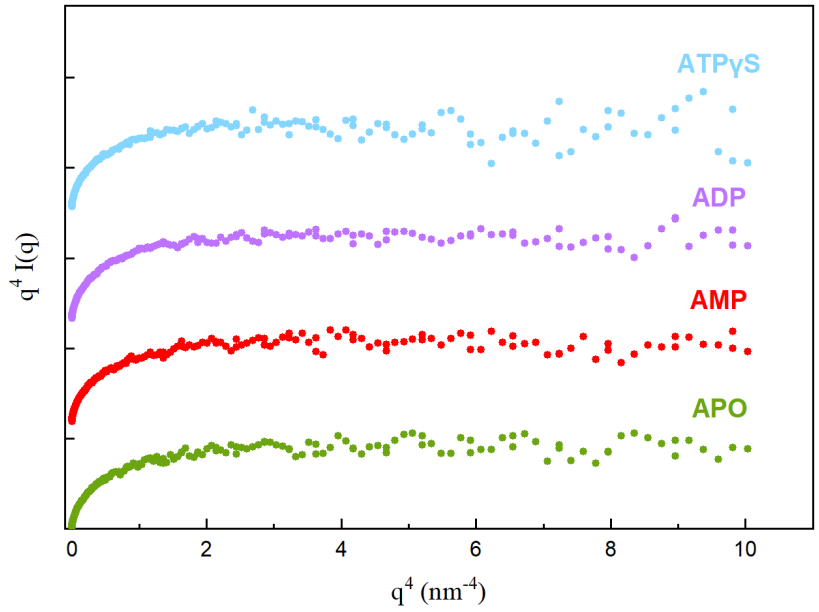

Figura 4.25: Gráfico de Porod-Debye, com a presença de platô constatando rigidez estrutural para a AaHsp90.

Tabela 4.13 - **Valores de MM da AaHsp90 por diferentes métodos**

Nucleotídeo	*Água*	*SAXSMoW*	Q_p	*MoW*	V_c	*S&S*	*Bayes*	***Teórica***
Apo	130.4	190.0	197.6	141.6	144.7	195.3	169.6	163.2
AMP	156.5	192.1	201.0	142.5	141.6	208.7	169.6	163.2
ADP	146.1	179.1	180.8	161.0	141.3	213.2	157.1	163.2
ATP	148.3	181.7	182.4	161.4	141.7	225.4	157.1	163.2

Metodologias descritas na seção 3.5. Massa teórica calculada pela sequência de aminoácidos utilizando o programa ProtParam (Gasteiger et al., 2005). A referência é da AaHsp90 em estado dimérico. Valores em kDa.

de espalhamento. Os critérios obtidos pelo GNOM para avaliação das $p(r)$ obtidas estão na Tabela C.3 (vide Apêndice C), e os parâmetros estruturais estão na Tabela 4.14. O cálculo da MM a partir do $I(0)$ dado pela $p(r)$ é o de calibração utilizando água como padrão. Aqui se observa, assim como para os casos que fazem uso de $I(0)$ obtido por Guinier, um desvio sistemático dos valores com relação ao teórico, o que pode indicar, por exemplo, influência da camada de hidratação da proteína no espalhamento e, portanto, um desvio do comportamento descrito pelo modelo de duas fases, e não devido apenas a flutuações estatísticas. Outra possibilidade é um desvio devido a problemas relacionados à determinação da concentração da proteína, com desvios sutis de concentração podendo levar a desvios consideráveis na MM calculada.

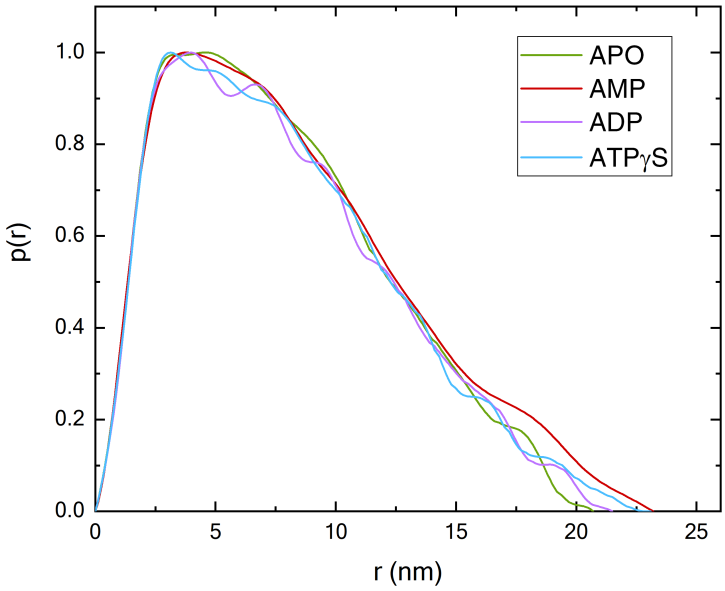

Figura 4.26: Funções $p(r)$ calculadas para a AaHsp90 com os diferentes nucleotídeos. Percebe-se uma forma elongada e a presença de oscilações pode tanto ser artefato do conjunto de dados quanto sinal de bastante rigidez estrutural.

Tabela 4.14 - **Parâmetros estruturais obtidos pela $p(r)$ para a AaHsp90**

Nucletídeo	R_g (nm)	$I(0)$ (10^6 UA)	MM (kDa)	***Teórica***	D_{max} (nm)	ER
Apo	6.32 ± 0.03	3.99 ± 0.02	125.8	**163.2**	20.7	4.33
AMP	6.72 ± 0.02	4.92 ± 0.01	155.1	**163.2**	23.2	4.10
ADP	6.36 ± 0.04	4.61 ± 0.03	145.2	**163.2**	21.5	3.33
ATPγS	6.43 ± 0.05	4.70 ± 0.04	147.9	**163.2**	23.1	5.48

Incertezas calculadas pelo GNOM, MM calculada usando água para calibração (seção 3.5.1.2) e ER calculado de acordo com a seção 3.3.2.

Tabela 4.15 - **Parâmetros para modelagem da AaHsp90 por SHANUM e AMBIMETER**

Nucleotídeo	N_{op}	q_{op}	$a-score$
Apo	26	3.95	2.72
AMP	35	4.74	2.47
ADP	24	3.51	2.65
ATPγS	25	3.41	2.58

N_{op} e q_{op} calculados de acordo com 3.4.7 e $a-score$ de acordo com a seção 3.4.4.

Por conta de maior razão sinal-ruído e melhor $p(r)$ gerada a partir dos critérios do GNOM, os dados da proteína sem nucleotídeo foram utilizados para modelagem. Para de-

terminação da estrutura da AaHsp90 em solução foram utilizados dois métodos: o *ab initio* por esferas e o método híbrido de corpo-rígido e *linkers* como esferas. Primeiramente a ambiguidade da reconstrução foi estabelecida utilizando o AMBIMETER, obtendo o valor de 2.68, considerado elevado, levando à necessidade de imposição de restrições à modelagem, sendo imposta simetria P2 para as reconstruções uma vez que a MM é condizente com AaHsp90 sendo um dímero em solução. Foram elaborados 20 modelos pelo DAMMIF, sendo tomado o modelo médio com o DAMAVER e em seguida refinado utilizando o DAMMIN com a estrutura filtrada como volume de busca inicial.

Para a modelagem híbrida foi utilizado o BUNCH. A partir da sequência primária da AaHsp90 foram modeladas estruturas de alta resolução pelo SWISSMODEL a partir da sequência de aminoácidos e de duas estruturas cristalográficas do PDB de outras espécies: uma para o NTD (entrada 1BYQ, de humanos, Obermann et al., 1998) e outra para os MD e CTD (entrada 3HJC, de *Leishmania major*, depositada em 2009 porém não publicada). No BUNCH, a imposição de simetria é automática quando modelando estados oligoméricos acima de monômeros, portanto foi imposta simetria P2.

Os ajustes estão na Figura 4.27**a**, sendo visualmente bastante similares, e ambas as estruturas obtidas foram alinhados utilizando o SUPCOMB, sendo a representação de cartum referente ao modelo do BUNCH e o envelope ao *ab initio*. Os valores de χ^2_{red} e do valor p foram, respectivamente, 3.38 e 0.231 para o modelo do DAMMIN e 3.58 e 0.009 para o do BUNCH. Os modelos estruturais propostos pelos programas estão na Figura 4.27**b**, e os CorMaps relativos aos ajustes são apresentados na Figura 4.28. A partir do DAMAVER foi obtidoo valor de NSD $= 1.391 \pm 0.411$.

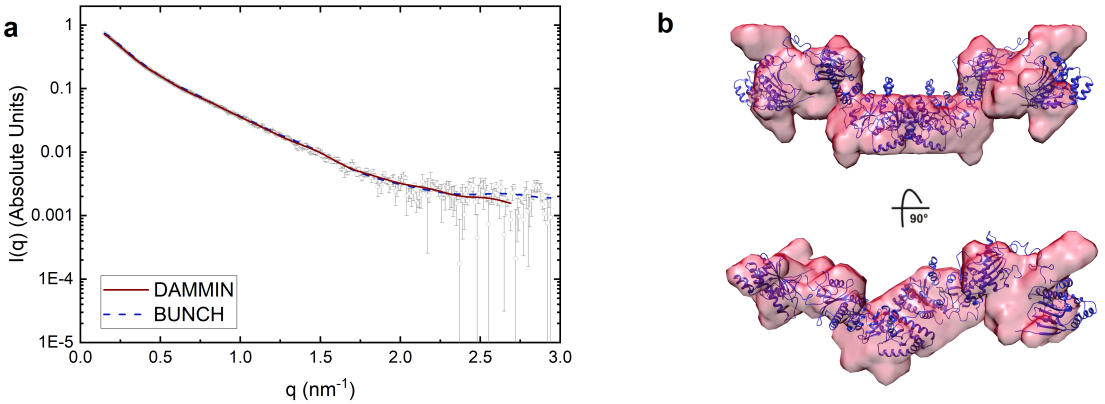

Figura 4.27: Ajuste dos modelos *ab initio* (vermelho) e de corpo-rígido (azul) aos dados da AaHsp90 em estado apo (**a**) e modelos obtidos por cada metodologia alinhados (**b**). Há forte concordância entre ambos.

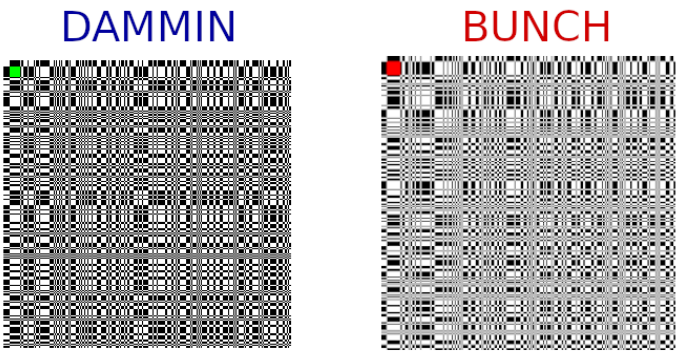

Figura 4.28: CorMaps relativos aos ajustes dos modelos propostos por DAMMIN e BUNCH.

O grupo do Prof. Dr. Carlos Ramos mostrou que a estrutura primária da AaHsp90 é altamente conservada desde *Saccharomyces cerevisiae* (cana-de-açúcar) a humanos. Também foi mostrado por dicroísmo circular (CD) que a proteína tem alto conteúdo de α-hélice, indo de acordo com o esperado por outras espécies na literatura, e estudos hidrodinâmicos feitos por ultracentrifugação analítica (AUC) indicam uma estrutura elongada a partir do desvio do raio de Stokes (R_S) e do coeficiente de difusão translacional (D) de previsto por meio da massa molecular, calculada por cromatografia de exclusão por tamanho acoplado a espalhamento de luz (SEC-MALS) como 160.1 ± 2.2 kDa, condizente com os 163.2 kDa esperados para um dímero. Assim como ocorreu para os dados de SAXS, não houve variação de R_S e D com ligação de nucleotídeos. Curiosamente, apesar de não passar por grandes variações conformacionais, a AaHsp90 hidroliza ATP e com eficiência de catálise similar a Hsp90s de outras espécies.

A Tabela 4.16 mostra as porcentagens de identidade e similaridade de sequência de aminoácidos entre membros de Hsp90 de diferentes espécies, calculados pelos membros do laboratório do Prof. Dr. Carlos Ramos. Um trabalho anterior do nosso grupo de pesquisa com o grupo do Prof. Dr. Júlio Borges mostrou que a Hsp90 de *Leishmania brasiliensis* (LbHsp90) apresenta flexibilidade estrutural entre os domínios, sendo, portanto, um resultado que, unido a este estudo, indica a importância de estudos de Hsp90 para diferentes espécies em busca de melhor compreensão do funcionamento dessa proteína central na manutenção da proteostase celular.

A estrutura de AaHsp90 obtida por SAXS, portanto, apresenta alguma flexibilidade mas conserva um perfil altamente elongado em solução sem variação ao haver ligação de nu-

Tabela 4.16 - **Identidade e similaridade de sequência da AaHsp90 para diferentes espécies**

Proteína	***Organismo***	***Identidade (%)***	***Similaridade (%)***
DmHsp90	*Drosophila melanogaster*	85	92
hHsp90α	*Homo sapiens*	79	88
hHsp90β	*Homo sapiens*	79	88
SsHsp90	*Saccharum sp.*	69	83
LbHsp90	*Leishmania brasiliensis*	64	79
yHsp82	*Saccahromyces cerevisiae*	63	78

Valores estimados pelos colaboradores do laboratório do Prof. Dr. Carlos Ramos.

cleotídeos de adenosina, se encaixando no modelo de que Hsp90s de eucariotos apresentam menor flexibilidade estrutural devido à interação com nucleotídeos, possivelmente indicando maior dependência de co-chaperonas ou da interação com proteínas-cliente para que possa atingir o estado conformacional necessário para que possa desempenhar sua função biológica auxiliando no enovelamento de outras proteínas, mesmo conseguindo hidrolisar ATP. Também é possível que o processo ocorra em uma escala de tempo muito curta em relação ao tempo em que a estrutura fica em conformação aberta. Essa predominância de estruturas abertas foi constatada por criomicroscopia eletrônica para Hsp90 humana (Southworth e Agard, 2008), e a importância de interações da Hsp90 humana tanto com proteínas-cliente quanto com co-chaperonas para o aumento nas taxas de hidrólise ATP foi demonstrada anteriormente (McLaughlin et al., 2002), o que pode indicar que essa mudança de atividade esteja relacionada justamente com as mudanças conformacionais esperadas. Ao observar a Tabela 4.16, indicando alta similaridade entre a AaHsp90 e Hsp90 de humanos, esse resultado passa a ser esperado. Tendo isso em vista, de maneira caricata, é como se o modelo de tesoura para Hsp90 tivesse, em *Aedes aegypti*, o caso de uma tesoura enferrujada, precisando de mais energia para que a sua função seja desempenhada por meio da interação com outras proteínas, sejam elas co-chaperonas ou as próprias proteínas-cliente.

Capítulo 5

Conclusões

Neste trabalho, foi feita uma apresentação sobre métodos de análise de dados de SAXS de proteínas em soluções diluídas, desde metodologias bem estabelecidas e amplamente utilizadas até trabalhos recentemente publicados e potencialmente promissores. Foram mostradas formas de análise que têm como objetivo validação de parâmetros estruturais, proposição de novas grandezas com significado físico, modelagem de diferentes tipos de proteína (em especial fazendo uso de procedimentos computacionais) e validação dessas modelagens. A bibliografia de SAXS em português sobre metodologias de análise é escassa e este trabalho pode ter utilidade como referência tanto para novos usuários da técnica quanto para atualização dos conhecimentos de já iniciados. As metodologias de análise apresentadas foram aplicadas para o estudo de proteínas chaperonas em solução das famílias Hsp70, incluindo o fator de troca de nucleotídeo GrpE-L1, e Hsp90.

Foi mostrado que as Hsp70 humanas de citoplasma (Hsp70-1A e Hsc70) e retículo endoplasmático (Bip) se apresentam como monômeros em solução e são bem enoveladas, mostrando alta similaridade estrutural entre si, indicando que as diferentes especificidades de proteínas-cliente e funcionalidade de cada uma se deve a variações estruturais mais sutis, a ponto de não poderem ser observadas por SAXS. As topologias obtidas indicam potenciais desvios com relação a resultados anteriores de DnaK (Hsp70 de *E. coli*) e Hsc70 bovina ligados a ADP, que deveriam, de acordo com o atual modelo, coincidir com a estrutura livre em solução. Por conta da maior razão sinal-ruído, a Bip foi escolhida como modelo de Hsp70 humana para análises posteriores. Foi possível fazer modelagem de corpo-rígido juntamente com metodologia *ab initio* e de *ensemble*. Os modelos *ab initio* e de corpo rígido, este último feito utilizando estruturas parciais de cada domínio, se ajustaram aos dados experimentais sem desvios sistemáticos. O estudo de *ensemble* não se ajustou aos

dados apesar de já apontar uma estrutura mais rígida. Isso mostrou que o modelo atual de Hsp70 livre em solução, elaborado a partir de estudos com DnaK, de grau de flexibilidade mais elevado, não foi observado por SAXS fazendo uso de estruturas conhecidas, indicando que, possivelmente, flexibilidade entre os domínios não tem um papel tão dominante em humanos. Mais estudos podem ser feitos comparando as variações de conformação para membros de Hsp70 humana na ligação de diferentes nucleotídeos para avaliar mais as fundo a validade do modelo proposto a partir da DnaK.

As GrpE-L1 de humanos foram medidas em diferentes condições de solvente e foi constatado em todos os meios que se trata de uma proteína bastante elongada e com significativo grau de desordem. As reconstruções *ab initio* resultaram em estrutura média de aspecto comparável à obtida por cristalografia para *Thermus thermophilus*, porém a curva teórica calculada para esta estrutura não descreveu bem os dados, indicando que trata-se, de fato, de uma estrutura média dentro de um *ensemble* de conformações possíveis. Não foi constatada compactação estrutural apreciável em presença de meio que quebra pontes dissulfeto com relação a meios de formação dessas ligações, podendo indicar que em humanos essas ligações não são o único fator para a manutenção da estrutura de α-hélices paralelas entre as cadeias.

As AaHsp90 não mostraram variação estrutural apreciável na resolução acessível por SAXS quando ligada a diferentes nucleotídeos de adenosina, e apresentaram baixa flexibilidade estrutural e perfil altamente elongado, porém foi possível modelar os dados por meio de simulações de corpo-rígido utilizando estruturas cristalográficas parciais para diferentes domínios e de procedimento *ab initio* com resultados de ajuste similares. Ambas as metodologias reproduziram o perfil elongado observado pela função de distribuição de pares. Os resultados apresentam contraste com estudo anterior feito com LbHsp90, que apresenta flexibilidade conformacional quando livre em solução, e, por outro lado, traços em comum com estudos publicados de Hsp90 humana, que mostra características estruturais similares à AaHsp90 e dependência de co-chaperonas e proteínas-cliente para aumento nas taxas de hidrólise de ATP. Esses resultados podem indicar que as Hsp90 de eucariotos apresentam, dependendo do espécie, uma estrutura mais ou menos maleável na ligação a nucleotídeos de adenosina e, portanto, maior ou menor dependência da ação de co-chaperonas ou de interações com proteínas-cliente para que seja alcançada a conformação necessária para o desempenho de sua função biológica.

Este trabalho ilustra o amplo leque de metodologias de análise disponível e o potencial da técnica de SAXS para o estudo de proteínas em solução em conjunto com outras metodologias de obtenção de estruturas a fim de que se obtenha um retrato mais detalhado sobre o funcionamento dessas macromoléculas nos processos biológicos.

Referências Bibliográficas

Andrews S. S., Using rotational averaging to calculate the bulk response of isotropic and anisotropic samples from molecular parameters, Journal of chemical education, 2004, vol. 81, p. 877

Anfinsen C. B., Principles that govern the folding of protein chains, Science, 1973, vol. 181, p. 223

Apweiler R., Bairoch A., Wu C. H., Barker W. C., Boeckmann B., Ferro S., Gasteiger E., Huang H., Lopez R., Magrane M., Martin M. J., Natale D. A., O'Donovan C., Redaschi N., Yeh L. L., UniProt: the Universal Protein knowledgebase, Nucleic Acids Research, 2004, vol. 32, p. D115

Badger J., A new algorithm for the reconstruction of protein molecular envelopes from X-ray solution scattering data, Journal of Applied Crystallography, 2019, vol. 52

Barbosa L. R. S., Spinozzi F., Mariani P., Itri R., Small-angle X-ray scattering applied to proteins in solution, Proteins in Solution and at Interfaces. John Wiley & Sons, Inc, 2013, pp 49–72

Beirlant J., Dudewicz E. J., Györfi L., Van der Meulen E. C., Nonparametric entropy estimation: An overview, International Journal of Mathematical and Statistical Sciences, 1997, vol. 6, p. 17

Berman H. M., Westbrook J., Feng Z., Gilliland G., Bhat T. N., Weissig H., Shindyalov I. N., Bourne P. E., The Protein Data Bank, Nucleic Acids Research, 2000, vol. 28, p. 235

Bernadó P., Effect of interdomain dynamics on the structure determination of modular proteins by small-angle scattering, European Biophysics Journal, 2010, vol. 39, p. 769

Bernadó P., Blanchard L., Timmins P., Marion D., Ruigrok R. W., Blackledge M., A structural model for unfolded proteins from residual dipolar couplings and small-angle x-ray scattering, Proceedings of the National Academy of Sciences, 2005, vol. 102, p. 17002

Bernadó P., Mylonas E., Petoukhov M. V., Blackledge M., Svergun D. I., Structural characterization of flexible proteins using small-angle X-ray scattering, Journal of the American Chemical Society, 2007, vol. 129, p. 5656

Best R. B., Emerging consensus on the collapse of unfolded and intrinsically disordered proteins in water, Current Opinion in Structural Biology, 2020, vol. 60, p. 27

Best R. B., Zheng W., Borgia A., Buholzer K., Borgia M. B., Hofmann H., Soranno A., Nettels D., Gast K., Grishaev A., et al., Comment on "Innovative scattering analysis shows that hydrophobic disordered proteins are expanded in water", Science, 2018, vol. 361, p. eaar7101

Biebl M. M., Buchner J., Structure, Function, and Regulation of the Hsp90 Machinery, Cold Spring Harbor perspectives in biology, 2019, p. a034017

Borges J. C., Fischer H., Craievich A. F., Hansen L. D., Ramos C. H., Free human mitochondrial GrpE is a symmetric dimer in solution, Journal of Biological Chemistry, 2003, vol. 278, p. 35337

Borges J. C., Seraphim T. V., Dores-Silva P. R., Barbosa L. R., A review of multi-domain and flexible molecular chaperones studies by small-angle X-ray scattering, Biophysical reviews, 2016, vol. 8, p. 107

Bracher A., Verghese J., The nucleotide exchange factors of Hsp70 molecular chaperones, Frontiers in molecular biosciences, 2015, vol. 2, p. 10

Brünger A. T., Free R value: a novel statistical quantity for assessing the accuracy of crystal structures, Nature, 1992, vol. 355, p. 472

Burger V. M., Arenas D. J., Stultz C. M., A structure-free method for quantifying conformational flexibility in proteins, Scientific reports, 2016, vol. 6, p. 29040

Casey L. W., Mark A. E., Kobe B., Small-Angle X-Ray Scattering for the Discerning Macromolecular Crystallographer, Australian Journal of Chemistry, 2014, vol. 67, p. 1786

Daugaard M., Rohde M., Jäättelä M., The heat shock protein 70 family: Highly homologous proteins with overlapping and distinct functions, FEBS letters, 2007, vol. 581, p. 3702

Debye P., Zerstreuung von röntgenstrahlen, Annalen der Physik, 1915, vol. 351, p. 809

Debye P., Anderson Jr H., Brumberger H., Scattering by an inhomogeneous solid. II. The correlation function and its application, Journal of Applied Physics, 1957, vol. 28, p. 679

Debye P., Bueche A., Scattering by an inhomogeneous solid, Journal of Applied Physics, 1949, vol. 20, p. 518

Debye P. J. W., Fuoss R. M., The Collected Papers of PJW Debye. Interscience Publishers New York, 1954

Dill K. A., Polymer principles and protein folding, Protein Science, 1999, vol. 8, p. 1166

Dores-Silva P. R., Barbosa L. R. S., Ramos C. H. I., Borges J. C., Human mitochondrial Hsp70 (mortalin): shedding light on ATPase activity, interaction with adenosine nucleotides, solution structure and domain organization, PLoS One, 2015, vol. 10, p. e0117170

Dos Reis M. A., Aparicio R., Zhang Y., Improving protein template recognition by using small-angle x-ray scattering profiles, Biophysical journal, 2011, vol. 101, p. 2770

Durand D., Vivès C., Cannella D., Pérez J., Pebay-Peyroula E., Vachette P., Fieschi F., NADPH oxidase activator p67phox behaves in solution as a multidomain protein with semi-flexible linkers, Journal of structural biology, 2010, vol. 169, p. 45

Fienup J. R., Reconstruction of an object from the modulus of its Fourier transform, Optics letters, 1978, vol. 3, p. 27

Fischer H., Oliveira Neto M. d., Napolitano H., Polikarpov I., Craievich A. F., Determination of the molecular weight of proteins in solution from a single small-angle X-ray scattering measurement on a relative scale, Journal of applied crystallography, 2010, vol. 43, p. 101

Fischer H., Polikarpov I., Craievich A. F., Average protein density is a molecular-weight-dependent function, Protein Science, 2004, vol. 13, p. 2825

Flory P. J., Principles of polymer chemistry. Cornell University Press, 1953

Franke D., Jeffries C. M., Svergun D. I., Correlation Map, a goodness-of-fit test for one-dimensional X-ray scattering spectra, Nature methods, 2015, vol. 12, p. 419

Franke D., Jeffries C. M., Svergun D. I., Machine learning methods for X-ray scattering data analysis from biomacromolecular solutions, Biophysical journal, 2018, vol. 114, p. 2485

Franke D., Svergun D. I., DAMMIF, a program for rapid ab-initio shape determination in small-angle scattering, Journal of applied crystallography, 2009, vol. 42, p. 342

Fuertes G., Banterle N., Ruff K. M., Chowdhury A., Mercadante D., Koehler C., Kachala M., Girona G. E., Milles S., Mishra A., et al., Decoupling of size and shape fluctuations in heteropolymeric sequences reconciles discrepancies in SAXS vs. FRET measurements, Proceedings of the National Academy of Sciences, 2017, vol. 114, p. E6342

Fuertes G., Banterle N., Ruff K. M., Chowdhury A., Pappu R. V., Svergun D. I., Lemke E. A., Comment on “Innovative scattering analysis shows that hydrophobic disordered proteins are expanded in water”, Science, 2018, vol. 361, p. eaau8230

Funari S. S., Rapp G., Perbandt M., Dierks K., Vallazza M., Betzel C., Erdmann V. A., Svergun D. I., Structure of free Thermus flavus 5 S rRNA at 1.3 nm resolution from synchrotron X-ray solution scattering, Journal of Biological Chemistry, 2000, vol. 275, p. 31283

Gasteiger E., Hoogland C., Gattiker A., Wilkins M. R., Appel R. D., Bairoch A., et al.,, 2005 in , The proteomics protocols handbook. Springer pp 571–607

Glatter O., A new method for the evaluation of small-angle scattering data, Journal of Applied Crystallography, 1977, vol. 10, p. 415

Glatter O., Kratky O., Small angle X-ray scattering. Academic press, 1982

Goldstein H., Poole C., Safko J., Classical Mechanics. Addison Wesley, 2002

Grant T. D., Ab initio electron density determination directly from solution scattering data, Nature methods, 2018, vol. 15, p. 191

Guin D., Gruebele M., Weak Chemical Interactions That Drive Protein Evolution: Crowding, Sticking, and Quinary Structure in Folding and Function, Chemical reviews, 2019, vol. 119, p. 10691

Guinier A., Fournet G., Small angle scattering of X-rays. John Wiley & Sons New York, 1955

Hajizadeh N. R., Franke D., Jeffries C. M., Svergun D. I., Consensus Bayesian assessment of protein molecular mass from solution X-ray scattering data, Scientific reports, 2018, vol. 8, p. 7204

Harrison C., GrpE, a nucleotide exchange factor for DnaK, Cell stress & chaperones, 2003, vol. 8, p. 218

Harrison C. J., Hayer-Hartl M., Di Liberto M., Hartl F.-U., Kuriyan J., Crystal structure of the nucleotide exchange factor GrpE bound to the ATPase domain of the molecular chaperone DnaK, Science, 1997, vol. 276, p. 431

Hartl F. U., Bracher A., Hayer-Hartl M., Molecular chaperones in protein folding and proteostasis, Nature, 2011, vol. 475, p. 324

Hura G. L., Budworth H., Dyer K. N., Rambo R. P., Hammel M., McMurray C. T., Tainer J. A., Comprehensive macromolecular conformations mapped by quantitative SAXS analyses, Nature methods, 2013, vol. 10, p. 453

Jacques D. A., Trewhella J., Small-angle scattering for structural biology—Expanding the frontier while avoiding the pitfalls, Protein Science, 2010, vol. 19, p. 642

Johnson B. D., Schumacher R. J., Ross E. D., Toft D. O., Hop modulates Hsp70/Hsp90 interactions in protein folding, Journal of Biological Chemistry, 1998, vol. 273, p. 3679

Kellner R., Hofmann H., Barducci A., Wunderlich B., Nettels D., Schuler B., Single-molecule spectroscopy reveals chaperone-mediated expansion of substrate protein, Proceedings of the National Academy of Sciences, 2014, vol. 111, p. 13355

Kikhney A. G., Borges C. R., Molodenskiy D. S., Jeffries C. M., Svergun D. I., SASBDB: Towards an automatically curated and validated repository for biological scattering data, Protein Science, 2019

Kirkpatrick Scott C. D. G., Vecchi M. P., Optimization by simulated annealing, Science, 1983, vol. 220, p. 671

Kleywegt G. J., Validation of protein models from Cα coordinates alone, Journal of molecular biology, 1997, vol. 273, p. 371

Konarev P. V., Svergun D. I., A posteriori determination of the useful data range for small-angle scattering experiments on dilute monodisperse systems, IUCrJ, 2015, vol. 2, p. 352

Kozin M. B., Svergun D. I., Automated matching of high-and low-resolution structural models, Journal of applied crystallography, 2001, vol. 34, p. 33

Levitt M., A simplified representation of protein conformations for rapid simulation of protein folding, Journal of molecular biology, 1976, vol. 104, p. 59

Luengo T. M., Kityk R., Mayer M. P., Rüdiger S. G., Hsp90 breaks the deadlock of the Hsp70 chaperone system, Molecular cell, 2018, vol. 70, p. 545

Luengo T. M., Mayer M. P., Rüdiger S. G., The Hsp70–Hsp90 chaperone cascade in protein folding, Trends in cell biology, 2018

McLaughlin S. H., Smith H. W., Jackson S. E., Stimulation of the weak ATPase activity of human hsp90 by a client protein, Journal of molecular biology, 2002, vol. 315, p. 787

Marchesini S., He H., Chapman H. N., Hau-Riege S. P., Noy A., Howells M. R., Weierstall U., Spence J. C., X-ray image reconstruction from a diffraction pattern alone, Physical Review B, 2003, vol. 68, p. 140101

Mayer M. P., Hsp70 chaperone dynamics and molecular mechanism, Trends in biochemical sciences, 2013, vol. 38, p. 507

Mayer M. P., Gierasch L. M., Recent advances in the structural and mechanistic aspects of Hsp70 molecular chaperones, Journal of Biological Chemistry, 2019, vol. 294, p. 2085

Mertens H. D., Svergun D. I., Structural characterization of proteins and complexes using small-angle X-ray solution scattering, Journal of structural biology, 2010, vol. 172, p. 128

Moore P. B., Small-angle scattering. Information content and error analysis, Journal of Applied Crystallography, 1980, vol. 13, p. 168

Murzin A. G., Brenner S. E., Hubbard T., Chothia C., SCOP: a structural classification of proteins database for the investigation of sequences and structures, Journal of molecular biology, 1995, vol. 247, p. 536

Mylonas E., Svergun D. I., Accuracy of molecular mass determination of proteins in solution by small-angle X-ray scattering, Applied Crystallography, 2007, vol. 40, p. s245

Nakamura A., Takumi K., Miki K., Crystal structure of a thermophilic GrpE protein: insight into thermosensing function for the DnaK chaperone system, Journal of molecular biology, 2010, vol. 396, p. 1000

Nelson D. L., Lehninger A. L., Cox M. M., Lehninger principles of biochemistry. Macmillan, 2008

Obermann W. M., Sondermann H., Russo A. A., Pavletich N. P., Hartl F. U., In vivo function of Hsp90 is dependent on ATP binding and ATP hydrolysis, The Journal of cell biology, 1998, vol. 143, p. 901

Orengo C. A., Michie A. D., Jones S., Jones D. T., Swindells M. B., Thornton J. M., CATH–a hierarchic classification of protein domain structures, Structure, 1997, vol. 5, p. 1093

Orthaber D., Bergmann A., Glatter O., SAXS experiments on absolute scale with Kratky systems using water as a secondary standard, Journal of Applied Crystallography, 2000, vol. 33, p. 218

Pelikan M., Hura G. L., Hammel M., Structure and flexibility within proteins as identified through small angle X-ray scattering, General physiology and biophysics, 2009, vol. 28, p. 174

Perkins S. J., Protein volumes and hydration effects: The calculations of partial specific volumes, neutron scattering matchpoints and 280-nm absorption coefficients for proteins and glycoproteins from amino acid sequences, European Journal of Biochemistry, 1986, vol. 157, p. 169

Petoukhov M. V., Franke D., Shkumatov A. V., Tria G., Kikhney A. G., Gajda M., Gorba C., Mertens H. D., Konarev P. V., Svergun D. I., New developments in the ATSAS program package for small-angle scattering data analysis, Journal of applied crystallography, 2012, vol. 45, p. 342

Petoukhov M. V., Konarev P. V., Kikhney A. G., Svergun D. I., ATSAS 2.1–towards automated and web-supported small-angle scattering data analysis, Applied crystallography, 2007, vol. 40, p. s223

Petoukhov M. V., Svergun D. I., Global rigid body modeling of macromolecular complexes against small-angle scattering data, Biophysical journal, 2005, vol. 89, p. 1237

Petoukhov M. V., Svergun D. I., Ambiguity assessment of small-angle scattering curves from monodisperse systems, Acta Crystallographica Section D: Biological Crystallography, 2015, vol. 71, p. 1051

Piiadov V., Ares de Araújo E., Oliveira Neto M., Craievich A. F., Polikarpov I., SAXSMoW 2.0: Online calculator of the molecular weight of proteins in dilute solution from experimental SAXS data measured on a relative scale, Protein Science, 2019, vol. 28, p. 454

Putnam C. D., Guinier peak analysis for visual and automated inspection of small-angle X-ray scattering data, Journal of applied crystallography, 2016, vol. 49, p. 1412

Putnam C. D., Hammel M., Hura G. L., Tainer J. A., X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution, Quarterly reviews of biophysics, 2007, vol. 40, p. 191

Ramakrishnan C., Ramachandran G., Stereochemical criteria for polypeptide and protein chain conformations: II. Allowed conformations for a pair of peptide units, Biophysical journal, 1965, vol. 5, p. 909

Rambo R. P., Tainer J. A., Characterizing flexible and intrinsically unstructured biological macromolecules by SAS using the Porod-Debye law, Biopolymers, 2011, vol. 95, p. 559

Rambo R. P., Tainer J. A., Accurate assessment of mass, models and resolution by small-angle scattering, Nature, 2013, vol. 496, p. 477

Riback J. A., Bowman M. A., Zmyslowski A., Knoverek C. R., Jumper J., Kaye E. B., Freed K. F., Clark P. L., Sosnick T. R., Response to comment on "Innovative scattering analysis shows that hydrophobic disordered proteins are expanded in water", Science, 2018, vol. 361, p. eaar7949

Riback J. A., Bowman M. A., Zmyslowski A. M., Knoverek C. R., Jumper J. M., Hinshaw J. R., Kaye E. B., Freed K. F., Clark P. L., Sosnick T. R., Innovative scattering analysis shows that hydrophobic disordered proteins are expanded in water, Science, 2017, vol. 358, p. 238

Riback J. A., Bowman M. A., Zmyslowski A. M., Plaxco K. W., Clark P. L., Sosnick T. R., Commonly used FRET fluorophores promote collapse of an otherwise disordered protein, Proceedings of the National Academy of Sciences, 2019, vol. 116, p. 8889

Roe R.-J., Methods of X-ray and neutron scattering in polymer science. vol. 739, Oxford University Press New York, 2000

Schilling M. F., The longest run of heads, The College Mathematics Journal, 1990, vol. 21, p. 196

Schneidman-Duhovny D., Hammel M., Tainer J. A., Sali A., Accurate SAXS profile computation and its assessment by contrast variation experiments, Biophysical journal, 2013, vol. 105, p. 962

Schopf F. H., Biebl M. M., Buchner J., The HSP90 chaperone machinery, Nature reviews Molecular cell biology, 2017, vol. 18, p. 345

Sedlak S. M., Bruetzel L. K., Lipfert J., Quantitative evaluation of statistical errors in small-angle X-ray scattering measurements, Journal of applied crystallography, 2017, vol. 50, p. 621

Seraphim T. V., Silva K. P., Dores-Silva P. R., Barbosa L. R., Borges J. C., Insights on the structural dynamics of Leishmania braziliensis Hsp90 molecular chaperone by small angle X-ray scattering, International journal of biological macromolecules, 2017, vol. 97, p. 503

Shannon C. E., Weaver W., , 1949 The mathematical theory of communication (Urbana, IL)

Shi L., Kataoka M., Fink A. L., Conformational characterization of DnaK and its complexes by small-angle X-ray scattering, Biochemistry, 1996, vol. 35, p. 3297

Sousa R., Lafer E. M., The Physics of Entropic Pulling: A Novel Model for the Hsp70 Motor Mechanism, International journal of molecular sciences, 2019, vol. 20, p. 2334

Southworth D. R., Agard D. A., Species-dependent ensembles of conserved conformational states define the Hsp90 chaperone ATPase cycle, Molecular cell, 2008, vol. 32, p. 631

Spinozzi F., Ferrero C., Ortore M. G., De Maria Antolinos A., Mariani P., GENFIT: software for the analysis of small-angle X-ray and neutron scattering data of macromolecules in solution, Journal of applied crystallography, 2014, vol. 47, p. 1132

Stuhrmann H. B., Interpretation of small-angle scattering functions of dilute solutions and gases. A representation of the structures related to a one-particle scattering function, Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography, 1970, vol. 26, p. 297

Svergun D., Determination of the regularization parameter in indirect-transform methods using perceptual criteria, Journal of applied crystallography, 1992, vol. 25, p. 495

Svergun D., Barberato C., Koch M. H., CRYSOL–a program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates, Journal of applied crystallography, 1995, vol. 28, p. 768

Svergun D. I., Restoring low resolution structure of biological macromolecules from solution scattering using simulated annealing, Biophysical journal, 1999, vol. 76, p. 2879

Svergun D. I., Koch M. H., Small-angle scattering studies of biological macromolecules in solution, Reports on Progress in Physics, 2003, vol. 66, p. 1735

Svergun D. I., Koch M. H., Timmins P. A., May R. P., Small angle X-ray and neutron scattering from solutions of biological macromolecules. vol. 19, Oxford University Press, 2013

Svergun D. I., Petoukhov M. V., Koch M. H., Determination of domain structure of proteins from X-ray solution scattering, Biophysical journal, 2001, vol. 80, p. 2946

Tame J. R., Vallone B., The structures of deoxy human haemoglobin and the mutant Hb Tyrα42His at 120 K, Acta Crystallographica Section D: Biological Crystallography, 2000, vol. 56, p. 805

Tiroli-Cepeda A. O., Ramos C. H. I., An overview of the role of molecular chaperones in protein homeostasis, Protein and peptide letters, 2011, vol. 18, p. 101

Trewhella J., Duff A. P., Durand D., Gabel F., Guss J. M., Hendrickson W. A., Hura G. L., Jacques D. A., Kirby N. M., Kwan A. H., et al., 2017 publication guidelines for structural modelling of small-angle scattering data from biomolecules in solution: an update, Acta Crystallographica Section D: Structural Biology, 2017, vol. 73, p. 710

Tria G., Mertens H. D., Kachala M., Svergun D. I., Advanced ensemble modelling of flexible macromolecules using X-ray solution scattering, IUCrJ, 2015, vol. 2, p. 207

Tuukkanen A. T., Kleywegt G. J., Svergun D. I., Resolution of ab initio shapes determined from small-angle scattering, IUCrJ, 2016, vol. 3, p. 440

Uversky V. N., Unusual biophysics of intrinsically disordered proteins, Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics, 2013, vol. 1834, p. 932

Valentini E., Kikhney A. G., Previtali G., Jeffries C. M., Svergun D. I., SASBDB, a repository for biological small-angle scattering data, Nucleic acids research, 2014, vol. 43, p. D357

Volkov V. V., Svergun D. I., Uniqueness of ab initio shape determination in small-angle scattering, Journal of applied crystallography, 2003, vol. 36, p. 860

Wilbanks S. M., Chen L., Tsuruta H., Hodgson K. O., McKay D. B., Solution small-angle X-ray scattering study of the molecular chaperone Hsc70 and its subfragments, Biochemistry, 1995, vol. 34, p. 12095

Wisniewska M., Karlberg T., Lehtiö L., Johansson I., Kotenyova T., Moche M., Schüler H., Crystal structures of the ATPase domains of four human Hsp70 isoforms: HSPA1L/Hsp70-hom, HSPA2/Hsp70-2, HSPA6/Hsp70B', and HSPA5/BiP/GRP78, PloS one, 2010, vol. 5, p. e8625

Yang J., Nune M., Zong Y., Zhou L., Liu Q., Close and allosteric opening of the polypeptide-binding site in a human Hsp70 chaperone BiP, Structure, 2015, vol. 23, p. 2191

Zheng W., Best R. B., An extended Guinier analysis for intrinsically disordered proteins, Journal of molecular biology, 2018, vol. 430, p. 2540

Apêndices

Apêndice A

Demonstrações das leis de Guinier e Porod

O primeiro passo consiste na aproximação do seno em uma série de Maclaurin em torno de a (Svergun et al., 2013), dada por

$$\text{sen}(x) = \text{sen}(a) + \frac{\cos(a)}{1!}(x-a) - \frac{\text{sen}(a)}{2!}(x-a)^2 - \frac{\cos(a)}{3!}(x-a)^3 + \mathcal{O}(qr^4). \quad \text{(A.1)}$$

Por simplificação, trabalharemos com a aproximação em ângulos muito baixos, e os termos seno são muito pequenos, e também deixaremos a série truncada no segundo termo não-nulo (no caso, o cúbico) pois, uma vez que trabalha-se em ângulos muito baixos em SAXS, o termo quártico torna-se desprezível rapidamente. Assim, ficaremos com

$$\text{sen}(qr) \approx qr - \frac{1}{3!}(qr)^3. \quad \text{(A.2)}$$

Voltando à equação de Debye:

$$I(q) \approx 4\pi \int_0^{Dmax} p(r)\, \frac{1}{qr}\left(qr + \frac{(qr)^3}{6}\right) dr, \quad \text{(A.3)}$$

da qual, simplificando os termos com qr ficamos com

$$I(q) \approx 4\pi \left(\int_0^{Dmax} p(r)dr + \frac{q^2}{6} \int_0^{Dmax} r^2 p(r)dr \right), \quad \text{(A.4)}$$

obtendo, ao deixar $4\pi \int_0^{Dmax} p(r)dr$ em evidência,

$$I(q) \approx 4\pi \int_0^{Dmax} p(r)dr \left(1 - \frac{q^2}{3} \frac{\int_0^{Dmax} r^2 p(r)dr}{2\int_0^{Dmax} p(r)dr}\right), \quad \text{(A.5)}$$

onde, como já visto, os termos correspondem em verde a $I(0)$ e em amarelo a R_g^2. Fazendo as devidas substituições ficamos com

$$I(q) \approx I(0)\left(1 - \frac{R_g^2}{3}q^2\right). \tag{A.6}$$

Há mais um passo possível de ser feito: lembrando da série de Taylor de e^x, dada por

$$\exp(x) = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \mathcal{O}(x^4), \tag{A.7}$$

é possível encaixar o termo de $I(q)$ entre parêntesis fazendo que $x = -\frac{R_g^2}{3}q^2$, nos deixando com uma intensidade de espalhamento que decai exponencialmente de acordo com

$$I(q) \approx I(0)\exp\left[-\left(\frac{R_g^2}{3}\right)q^2\right], \tag{A.8}$$

que é a *lei de Guinier*. Essa expressão pode tomar uma forma mais prática tomando o ln de ambos os lados, resultando

$$\ln I(q) = \ln I(0) - \frac{R_g^2}{3}q^2, \tag{A.9}$$

que é a expressão que baseia as análises por meio de linearizações de gráficos de $\ln I(q)$ vs q^2.

A dedução da lei de Porod baseia-se na assunção da existência de duas fases em solução. Para um modelo de duas fases ideal, ou seja, com dois valores fixos para densidade eletrônica ρ_1 e ρ_2, cada um ocupando as respectivas frações volumétricas ϕ_1 e $\phi_2 = 1-\phi_1$, a densidade eletrônica média $\langle\rho\rangle$ é dada por $\langle\rho\rangle = \phi_1\rho_1+\phi_2\rho_2$. Desse modo, seja $\Delta\rho = \rho_1-\rho_2$, as flutuações com relação à densidade média são dadas por

$$\begin{aligned}
\eta_1 &= \rho_1 - \langle\rho\rangle \\
&= \rho_1 - (\phi_1\rho_1 + \phi_2\rho_2) \\
&= \rho_1 - [\phi_1\rho_1 + (1-\phi_1)\rho_2] \\
&= (1-\phi_1)(\rho_1 - \rho_2),
\end{aligned}$$

e, portanto,

$$\eta_1 = \phi_2\Delta\rho. \tag{A.10}$$

Analogamente,

$$\eta_2 = -\phi_1 \Delta\rho, \tag{A.11}$$

e, desse modo, definindo $\Delta\eta = \eta_1 - \eta_2$ e fazendo as respectivas substituições, tem-se que $\Delta\eta = \Delta\rho$.

Para a dedução da lei de Porod (Roe, 2000) pode-se subdividir um elemento de volume da amostra em duas fases: uma com densidade ρ_1 (de volume $\phi_1 V$) e outra com densidade ρ_2 (de volume $\phi_2 V$). Cada fase pode, por sua vez, ser subdividida em região de *bulk* (B) e de superfície (S), definindo as quatro regiões para descrição do elemento de volume: 1B, 1S, 2B e 2S. Assim, a fase 1 é dividida em uma parte superficial de volume Sr e uma de bulk com volume $\phi_1 V - Sr$, com o mesmo ocorrendo para a fase 2 (analogamente, com volume de bulk $\phi_2 V - Sr$), onde S é a área superficial entre ambas as fases. Dados dois vetores de posição arbitrários $\vec{r'}$ e $\vec{r''}$, a espessura das regiões de superfície é dada por $r = |\vec{r'} - \vec{r''}|$ e é tomada como sendo muito menor que o menor raio de curvatura da superfície que delimita ambas as fases. Para um sistema constituído somente por proteínas, uma fase se refere à proteína contida no volume e outra, ao solvente. A Figura A.1 ilustra as quatro regiões descritas.

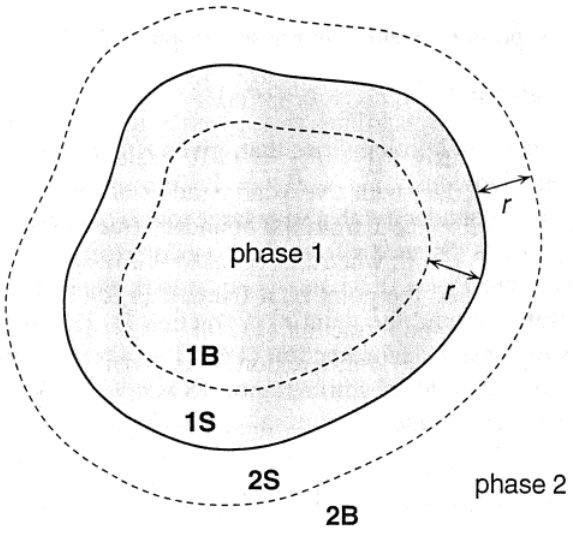

Figura A.1: Divisão de um elemento de volume da amostra em fases, sendo 1 referente à partícula e 2 referente ao meio. B e S indicam duas subregiões de cada fase, sendo *bulk* e superfície com espessura r, respectivamente (Roe, 2000).

Para esta dedução não é necessário que a função $\gamma(r)$ seja normalizada, de modo que podemos generalizá-la por uma função $\Gamma(r)$ tal que

$$\gamma(r) = \frac{\Gamma(r)}{\Gamma(0)}, \tag{A.12}$$

sendo, portanto, $\Gamma(0)$ um fator de normalização. Também pode-se escrever que $\Gamma(r) = V\langle\eta(\vec{r'})\eta(\vec{r''})\rangle$. Considerando a divisão em quatro regiões, pode-se fazer a decomposição $V\Gamma(r) = V_{1B}\Gamma_{1B}(r) + V_{1S}\Gamma_{1S}(r) + V_{2S}\Gamma_{2S}(r) + V_{2B}\Gamma_{2B}(r)$.

Dado que r é suficientemente pequeno, para um $\vec{r'}$ na região 1B temos que, necessariamente, $\vec{r''}$ pertence ou à região 1B ou à 1S. Dessa forma os valores de $\eta(\vec{r'})$ e $\eta(\vec{r''})$ serão iguais. Assim, fica-se com

$$\Gamma_{1B}(r) = (\phi_1 V - Sr)\eta_1^2, \tag{A.13}$$

e, analogamente,

$$\Gamma_{2B}(r) = (\phi_2 V - Sr)\eta_2^2. \tag{A.14}$$

Quando tratando regiões de superfície a questão fica um pouco mais complexa. Há de se considerar o caso em que $\vec{r'}$ está em uma fase e $\vec{r''}$ em outra. Seja P_{ij} a probabilidade média de $\vec{r'}$ estar na fase i e $\vec{r''}$ na fase j tem-se que, para $i \neq j$

$$\Gamma_{iS}(r) = Sr(\eta_i^2 P_{ii} + \eta_i\eta_j P_{ij}), \tag{A.15}$$

e, assim, o problema se reduz a determinar as probabilidades. Como se trata de probabilidades, fixa-se $P_{ii} + P_{ij} = 1$. Imaginando uma esfera de raio r e distância x entre seu centro e a superfície entre as fases (Figura A.2), pode-se inferir que para $x = r$ tem-se $P_{ii} = 1$ e $P_{ij} = 0$, com ambos os vetores estando na mesma fase. Para $x = 0$ as probabilidades são iguais, ou seja, $P_{ii} = P_{ij} = 1/2$. Dessa maneira, em geral pode-se dizer que uma fração $1/2(1 + x/r)$ da superfície da esfera está na fase i e $1/2(1 - x/r)$ está na j. Com essa informação pode-se calcular P_{ii} como

$$P_{ii} = \frac{1}{r}\int_0^r \frac{1}{2}\left(1 + \frac{x}{r}\right)dx = \frac{1}{2r}\left(r + \frac{r^2}{2r}\right) = \frac{3}{4}, \tag{A.16}$$

e, portanto, $P_{ij} = 1/4$. Dessa maneira ficamos com

$$\Gamma_{1S}(r) = Sr\left(\frac{3}{4}\eta_1^2 + \frac{1}{4}\eta_1\eta_2\right), \tag{A.17}$$

$$\Gamma_{2S}(r) = Sr\left(\frac{1}{4}\eta_1\eta_2 + \frac{3}{4}\eta_2^2\right). \tag{A.18}$$

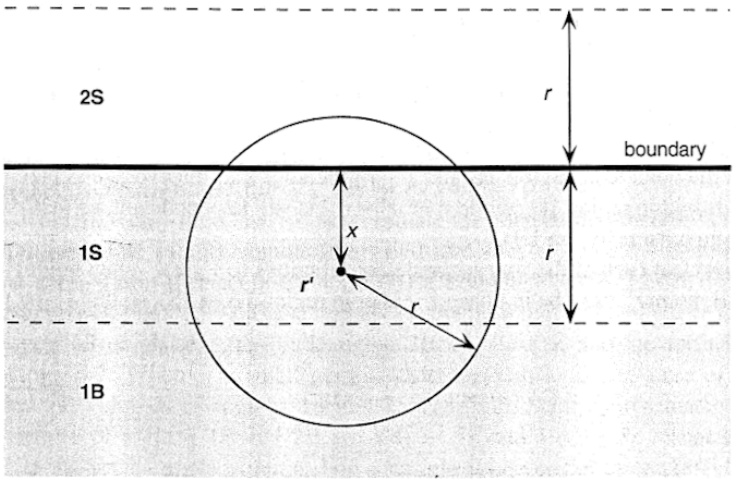

Figura A.2: Esquematização do método de determinação das probabilidades P_{ii} e P_{ij}. Imagina-se uma esfera de raio r a uma distância x da superfície e a partir dessa distância as distâncias são calculadas (Roe, 2000).

Assim, somando cada contribuição temos que

$$\begin{aligned}\Gamma(r) &= V(\phi_1\eta_1^2 + \phi_2\eta_2^2) - \frac{Sr}{4}\left(\eta_1^2 + \eta_2^2 - 2\eta_1\eta_2\right) \\ &= V\langle\eta^2\rangle - \frac{Sr}{4}(\Delta\eta)^2 \\ &= V\langle\eta^2\rangle\left[1 - \frac{1}{4}\frac{S}{V}\frac{(\Delta\rho)^2}{\langle\eta^2\rangle}r\right],\end{aligned} \tag{A.19}$$

e, uma vez que o fator S/V tem dimensão de inverso de comprimento, pode-se definir um comprimento l_P tal que $l_P = 4(V/S)(\langle\eta^2\rangle/(\Delta\rho)^2)$ para simplificar a expressão, resultando em

$$\Gamma(r) = V\langle\eta^2\rangle\left(1 - \frac{r}{l_P}\right). \tag{A.20}$$

Para $r << l_P$ pode-se aproximar a função em série de Taylor como

$$\Gamma(r) = V\langle\eta^2\rangle\, exp\left(-\frac{r}{l_P}\right). \tag{A.21}$$

O último passo consiste no cálculo da intensidade a partir da $\Gamma(r)$. Desse modo, como $\Gamma(r)$ é basicamente a $\gamma(r)$, tem-se que

$$
\begin{aligned}
I(q) &= \int_0^\infty 4\pi V\langle\eta^2\rangle \, exp\Big(- \frac{r}{l_P}\Big) r^2 \frac{sen(qr)}{qr} dr \\
&= V\langle\eta^2\rangle \frac{8\pi l_P^3}{\Big(1 + l_P^2 q^2\Big)^2},
\end{aligned}
\tag{A.22}
$$

de modo que, para $q \to \infty$, tem-se

$$
I(q) \to \frac{8\pi V\langle\eta^2\rangle}{l_P q^4}, \tag{A.23}
$$

com a lei de Porod tendo a forma mais reconhecida como

$$
I(q) \to \frac{2\pi S(\Delta\rho)^2}{q^4}, \tag{A.24}
$$

com algumas variações sendo representada em função de grandezas como razão S/V e $I(0)$ (Rambo e Tainer, 2011).

O formalismo acima é desenvolvido a partir da assunção de que a superfície entre as fases é bem definida, e, portanto, a lei só vale para proteínas bem enoveladas, pois quando se trata de uma cadeia muito flexível, a superfície vista por SAXS não é tão bem delimitada por se tratar de uma estrutura média, e isso faz com que a lei não consiga descrever o espalhamento para este tipo de partícula a ângulos mais altos (para uma melhor descrição desses casos deve-se utilizar modelos de polidispersão).

Apêndice B

Obtenção dos máximos dos gráficos de Kratky adimensional e GPA

Há dois casos extremos para mostrar no caso do Kratky adimensional: o de proteínas perfeitamente globulares e de proteínas de cadeia aleatória. Para o primeiro caso basta multiplicar ambos os lados por $(qR_g)^2/I(0)$ para permitir encontrar o máximo do gráfico de Kratky adimensional. Multiplicando, tem-se que

$$(qR_g)^2\frac{I(q)}{I(0)} = (qR_g)^2\exp\left[-\left(\frac{R_g^2}{3}\right)q^2\right], \tag{B.1}$$

portanto, encontrando o ponto extremo

$$2(qR_g)\exp\left[\frac{R_g^2}{3}q^2\right] + (qR_g)^2\exp\left[-\left(\frac{R_g^2}{3}\right)q^2\right]\left(-\frac{2qR_g}{3}\right) = 0, \tag{B.2}$$

$$2(qR_g)\exp\left[-\left(\frac{R_g^2}{3}\right)q^2\right]\left(1-\frac{qR_g}{3}\right) = 0, \tag{B.3}$$

$$qR_g = \sqrt{3}. \tag{B.4}$$

Para esse valor na abscissa tem-se $(qR_g)^2I(q)/I(0) \approx 1.1$. O segundo caso apresenta um plateau de $I(q)/I(0) \to 2$ para valores maiores de qR_g.

Para encontrar a expressão utilizada para a GPA, multiplica-se ambos os lados da lei de Guinier por q:

$$qI(q) = I(0)\sqrt{q^2}\exp\left[-\left(\frac{R_g^2}{3}\right)q^2\right]. \tag{B.5}$$

Para encontrar o máximo, basta derivar e igualar a zero para encontrar os pontos extremos.

$$\frac{d}{dq}\left\{ I(0)\sqrt{q^2}\exp\left[-\left(\frac{R_g^2}{3}\right)q^2\right]\right\} = 0, \tag{B.6}$$

$$I(0)\exp\left[-\left(\frac{R_g^2}{3}\right)q^2\right]\left(1 - 2\frac{R_g^2}{3}q^2\right) = 0, \tag{B.7}$$

$$1 - 2\frac{R_g^2}{3}q^2 = 0, \tag{B.8}$$

$$R_g^2 q^2 = 1.5, \tag{B.9}$$

ou, mais simplesmente

$$qR_g \approx 1.22. \tag{B.10}$$

Apêndice C

Critérios perceptuais do GNOM obtidos

Neste apêndice constam os critérios calculados pelo GNOM nos ajustes feitos para a obtenção das funções $p(r)$ de cada amostra.

	Hsp70-1A	**Hsc70**	**Bip**	***Ideal***
α	204.2	20.68	120.0	-
DISCRP	4.854	0.987	2.965	*0.700*
OSCILL	1.236	1.250	1.319	*1.100*
STABIL	0.014	0.013	0.012	*0.000*
SYSDEV	0.828	0.831	0.870	*1.000*
POSITV	1.000	1.000	1.000	*1.000*
VALCEN	0.887	0.725	0.870	*0.950*
SMOOTH	1.248	0.711	0.533	*0.000*
ESTIMATE	**0.6129**	**0.6596**	**0.6634**	***1.000***

Tabela C.1 - Valores dos critérios do GNOM para a avaliação das $p(r)$ geradas.

	Tampão	**β-mercaptoetanol**	**H_2O_2**	***Ideal***
α	63.9	35.2	50.8	-
DISCRP	2.146	2.607	2.410	*0.700*
OSCILL	1.561	1.617	1.605	*1.100*
STABIL	0.013	0.017	0.019	*0.000*
SYSDEV	0.782	0.640	0.730	*1.000*
POSITV	1.000	1.000	1.000	*1.000*
VALCEN	0.755	0.730	0.742	*0.950*
SMOOTH	0.572	0.068	0.128	*0.000*
ESTIMATE	**0.4777**	**0.4915**	**0.4947**	***1.000***

Tabela C.2 - Critérios do GNOM para as $p(r)$ encontradas da GrpE-L1.

	Apo	**AMP**	**ADP**	**ATPγS**	***Ideal***
α	36.2	304.2	58.3	49.7	-
DISCRP	1.669	2.183	1.445	1.895	*0.700*
OSCILL	1.502	1.638	1.631	1.725	*1.100*
STABIL	0.009	0.011	0.012	0.012	*0.000*
SYSDEV	0.784	0.628	0.688	0.693	*1.000*
POSITV	1.000	1.000	1.000	1.000	*1.000*
VALCEN	0.782	0.744	0.762	0.741	*0.950*
SMOOTH	0.981	0.711	0.418	0.161	*0.000*
ESTIMATE	**0.4789**	**0.4321**	**0.4651**	**0.4589**	***1.000***

Tabela C.3 - Critérios perceptuais do GNOM para avaliação das $p(r)$ geradas para os conjuntos de dados da AaHsp90.