MARCELO AUGUSTO DOS REIS

APLICAÇÃO DE TÉCNICAS DE ESPALHAMENTO DE RAIOS X NA CARACTERIZAÇÃO ESTRUTURAL DE PROTEÍNAS E MODELAGEM COMPUTACIONAL UTILIZANDO VÍNCULOS EXPERIMENTAIS OBTIDOS POR SAXS

CAMPINAS

2013

UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE QUÍMICA

MARCELO AUGUSTO DOS REIS

APLICAÇÃO DE TÉCNICAS DE ESPALHAMENTO DE RAIOS X NA CARACTERIZAÇÃO ESTRUTURAL DE PROTEÍNAS E MODELAGEM COMPUTACIONAL UTILIZANDO VÍNCULOS EXPERIMENTAIS OBTIDOS POR SAXS

ORIENTADOR: PROF. DR. RICARDO APARICIO

TESE DE DOUTORADO APRESENTADA AO
INSTITUTO DE QUÍMICA DA UNICAMP PARA
OBTENÇÃO DO TÍTULO DE DOUTOR EM CIÊNCIAS.

ESTE EXEMPLAR CORRESPONDE À VERSÃO FINAL DA TESE DEFENDIDA
POR MARCELO AUGUSTO DOS REIS E ORIENTADA PELO PROF. DR. RICARDO APARICIO.

Assinatura do Orientador

CAMPINAS
2013

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Química
Danielle Dantas de Sousa - CRB 8/6490

R277a

Reis, Marcelo Augusto dos, 1978-
Aplicação de técnicas de espalhamento de raios X na caracterização estrutural de proteínas e modelagem computacional utilizando vínculos experimentais obtidos por SAXS / Marcelo Augusto dos Reis. – Campinas, SP : [s.n.], 2013.

Orientador: Ricardo Aparicio.
Tese (doutorado) – Universidade Estadual de Campinas, Instituto de Química.

1. Espalhamento de raios X. 2. SAXS. 3. Cristalografia de proteínas. 4. Bioinformática. 5. Modelagem 3D. I. Aparicio, Ricardo. II. Universidade Estadual de Campinas. Instituto de Química. III. Título.

Informações para Biblioteca Digital

Título em outro idioma: Application of X-ray scattering techniques in protein structure characterization and computational modeling using experimental restraints obtained by SAXS
Palavras-chave em inglês:
X-ray scattering
SAXS
Protein crystallography
Bioinformatics
3D modeling
Área de concentração: Físico-Química
Titulação: Doutor em Ciências
Banca examinadora:
Ricardo Aparicio [Orientador]
Rosângela Itri
Mário Tyago Murakami
Munir Salomão Skaf
Adalberto Maurizio Sacchi Bassi
Data de defesa: 13-12-2013
Programa de Pós-Graduação: Química

Para aquela que me apoiou,

nos momentos cinzentos e de glória.

Dedico à Stela, minha amada esposa e companheira.

AGRADECIMENTOS

- Ao Ricardo Aparicio, amigo e orientador, por ter me recebido gentilmente em seu grupo e por ter me oferecido a oportunidade de trabalho em uma área onde a Física, a Química, a Biologia e a Computação se encontram fortemente. Ficarei eternamente grato pela confiança depositada em mim na execução de parte do projeto em terras estrangeiras. Fica aqui o meu muito obrigado.

- Ao Prof. Dr. Silvio Vitiello do IFGW-UNICAMP que apoiou a minha decisão e vontade de trabalhar com ciência multidisciplinar e ao Prof. Dr. Carlos Giles do mesmo instituto, que me auxiliou formalmente sendo o meu orientador no período de transição do doutorado na física para o doutorado na química.

- Ao Instituto de Física (IFGW-UNICAMP) e ao Instituto de Química (IQ-UNICAMP) por terem me proporcionado a formação científica mais abrangente e profunda que eu poderia ter.

- Ao Laboratório Nacional de Luz Síncrotron (LNLS) pelo apoio nas coletas de cristalografia e SAXS.

- Ao Center for Computational Medicine and Bioinformatics (CCMB) da University of Michigan por terem me recebido com muito profissionalismo.

- Ao ex-membro do LABEC, Marcelo Leite pela ajuda nos primeiros experimentos de cristalização e coleta de dados e aos novos membros Valéria Scorsato e Emanuella Fonseca por me auxiliarem com todos os trâmites “fim de tese”.

- À Prof.^{a} Dr.^{a} Anete Pereira da Silva do CBMEG-UNICAMP e ao seu grupo, em especial ao Antonio Saraiva, responsável pela clonagem, expressão e purificação da SurE e companheiro de coletas no LNLS.

- Ao Prof. Dr. Yang Zhang, o qual eu considero um mestre do *pipeline* em bioinformática, por ter me recebido em seu grupo durante a realização do doutorado sanduíche.

- A todos os membros do Zhang's lab. Especialmente Andrea Bazzoli, Huisun Lee, James, Dong Xu, Sava, Srayanta e Ambrish Roy por todos os momentos "lab meeting" e pelas primeiras lições sobre o I-TASSER.

- Aos amigos de Ann Arbor, Katie e Chad, Bebete Martins e Sueann Caulfield e Lenny Urena pela sua inestimável ajuda no mapeamento da área.

- Ao Prof. Dr. Jorge Iulek e seu grupo na UEPG, por finalizar o refinamento das estruturas cristalográficas da SurE devido ao meu estágio na Universidade de Michigan.

- Aos servidores da CPG do IQ, em especial à Bel que sempre com presteza atendeu todos os meus trâmites burocráticos.

- Aos membros da banca do exame de qualificação geral Prof. Dr. Adalberto Bassi, Prof. Dr. Marcelo Ganzarolli e Prof. Dr. Roy Bruns, pelas saudáveis discussões.

- Aos Prof. Dr. Fábio Gozzo, Prof. Dr. Michel Yamagishi e Prof. Dr. Rogério Custodio pelas sugestões sobre os trabalhos de tese durante o exame de qualificação de área.

- Ao CNPq e IFSULDEMINAS pelo apoio financeiro.

- Aos meus alunos que compreendem a delicada tarefa de ocupar dois lugares ao mesmo tempo.

- À comunidade software livre, que sem ela, não haveria doutorado.

- Ao meu pai Joaquim, por me ensinar desde pequeno que não precisamos ter títulos acadêmicos para construir coisas sofisticadas e que funcionem, bastam a ideia e o trabalho. À minha mãe Marcia, uma professora por ofício e responsável por despertar a minha paixão e vocação pela educação. Aos meus irmãos André e Marcia Andrea. O primeiro por ter me colocado nos caminhos da informática e a segunda por me inspirar e me lembrar que ainda existe a Arte nesse mundo, aliás, meu muito obrigado pela criação do logotipo do SAXSTER!

- E para o final.... à grande companheira de jornada, àquela que desde os tempos idos esteve ao meu lado desbravando esse mundo fascinante. Obrigado por tudo meu amor, sem você seria bem mais difícil a caminhada. Obrigado Stela.

CURRICULUM

Marcelo Augusto dos Reis

LATTES: http://lattes.cnpq.br/7681935841953063

Formação

2010 - 2013 Doutorado em Ciências (Físico-Química)
Instituto de Química - Universidade Estadual de Campinas

2010 - 2011 Pesquisador visitante
Universidade de Michigan (Ann Arbor – Michigan, EUA) (Maio 2010-Agosto 2011)

2008 - 2010 Doutorado em Física (interrompido e transferido para o doutorado em Físico-Química)
Instituto de Física *Gleb Wataghin* - Universidade Estadual de Campinas
Bolsa: Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)

2005 - 2008 Mestrado em Física
Instituto de Física *Gleb Wataghin* - Universidade Estadual de Campinas
Bolsa: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES)

2005 - 2006 Licenciatura em Física
Universidade Estadual de Campinas

2001 - 2004 Bacharelado em Física
Universidade Estadual de Campinas
Bolsa: Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)

Ocupação profissional atual

2012 - Professor de Física
Instituto Federal de Educação, Ciência e Tecnologia do Sul de Minas Gerais
Câmpus Inconfidentes

Publicações

1. **Reis, M. A.**; Saraiva, A. M., dos Santos; M. L.; de Souza, A. P.; Aparicio, R. Crystallization and preliminary X-ray analysis of stationary phase survival protein E (SurE) from *Xylella fastidiosa* in two crystals forms. Acta Crystallographiica Section F, v.68, 464-467 (2012)

2. **Reis, M. A.**; Aparicio, Ricardo; Zhang, Yang. Improving protein template recognition by using small angle X-ray scattering profiles. Biophysical Journal, v. 101, 2770-2781 (2011)

3. Saraiva, Antonio M.; **Reis, M. A.**; Tada, Susely F.; Rosselli-Murai, Luciana K.; Schneider, Dilaine R. S.; Pelloso, Alexandre C.; Toledo, Marcelo A. S.; Giles, Carlos; Aparicio, Ricardo; de Souza, Anete P. Functional and small-angle X-ray scattering studies of a new stationary phase survival protein E (SurE) from - evidence of allosteric behaviour . The FEBS Journal , v.276, 6751-6762 (2009)

4. **Reis, M. A.**, VITIELLO, S. A.. Método Variacional com Monte Carlo aplicado ao oscilador harmônico quântico In Revista Brasileira de Ensino de Física (São Paulo). , v.28, 45-50 (2006)

Resumos em Congressos e Eventos

1. **Reis, M. A.**; Aparicio, R.; Zhang, Y. SAXSTER: A Web Server for protein template recognition aided by threading and small-angle X-ray scaterring data In Reunião Anual de Usuários do LNLS, Campinas, SP, 2012.

2. **Reis, M. A.**; Aparicio, R. X-ray diffraction techniques and Computational Modeling for the study of protein structure-function relationships In São Paulo School of Advanced Science, EMBRAPA, Campinas-SP, 2012

3. **Reis, M. A.**; MURAI, M J, Lopes-Cendes, I; Aparicio, Ricardo. 2009. Saxs solution studies and structure prediction of EFHC1, a protein related to juvenile myoclonic epilepsy In *X-Meeting 5th international conference of the Brazilian association for bioinformatics and computational biology Angra dos Reis 2009*

4. **Reis, M. A.**; Saraiva, Antonio M.; Tada, Susely F.; Rosselli-Murai, Luciana K.; Toledo, Marcelo A. S.; de Souza, Anete P.; Aparicio, Ricardo. 2009. The Oligomeric State of the SurE Protein from Xylella fastidiosa: a Study by Small-Angle X-ray Scattering In *Congress Abstract CD-ROM XXXVIII Annual Meeting of the Brazilian Society for Biochemistry and Molecular Biology (SBBq) Águas de Lindóia - SP 2009*

5. **Reis, M. A**.; VITIELLO, S. A.. 2008. The supersolid phase of 4He In *Caderno de Resumos XXXI Encontro Nacional de Física da Matéria Condensada Águas de Lindóia-SP 2008*

6. **Reis, M. A.**; Vitiello, S. A.. 2007. The Solid Phase of Systems of 4He Atoms In *Caderno de Resumos XXX Encontro Nacional de Física da Matéria Condensada São Lourenço 2007*

7. **Reis, M. A.**; Vitiello, S. A.. 2006. Variational theory for the solid phase of systems of helium atoms In *Caderno*

Resumo

Neste trabalho de tese, o problema da caracterização estrutural de proteínas foi abordado de maneira contextualizada e com um viés em modelagem computacional utilizando vínculos experimentais obtidos com a técnica de espalhamento de raios X a baixos ângulos (SAXS, *Small-Angle X-ray Scattering*).

Parte das pesquisas foram concentradas na caracterização estrutural da proteína SurE de *Xylella fastidiosa* (XfSurE) por técnicas experimentais e computacionais. Estudos estruturais da XfSurE realizados com a técnica de SAXS apontaram para um arranjo tetramérico da enzima apo e, do nosso conhecimento, foi a primeira estrutura em solução descrita na literatura para esta família de proteínas. Quando associada às técnicas computacionais — como, por exemplo, análise de modos normais de vibração — a interpretação das análises por SAXS foi realçada. Neste caso, o vínculo experimental imposto pela curva $I(q)$ possibilitou que uma estrutura em solução fosse modelada apenas com o uso de um único modo normal, cujo efeito estaria relacionado com as possíveis transições alostéricas de XfSurE.

Em outra frente de trabalho, um novo programa denominado SAXSTER foi desenvolvido. SAXSTER tem a habilidade de gerar modelos estruturais mais prováveis para uma proteína-alvo a partir de alinhamentos ótimos obtidos por *threading* e de estruturas similares identificadas em um banco de dados, com o auxílio de SAXS. A partir dos dados de entrada, é realizada uma busca no *Protein Data Bank* para que a estrutura da proteína-alvo possa ser predita. O programa foi testado para 553 proteínas não redundantes. Foi demonstrado que SAXSTER pode melhorar consistentemente o resultado global da classificação dos alinhamentos, com p-valores que variam de 10^{-6} a 10^{-8}. De acordo com TM-score médio, conclui-se que SAXSTER tende a melhorar o desempenho preditivo conforme a estrutura da proteína-alvo se afasta da forma globular.

Abstract

In this work, the protein structure problem was approached from two different perspectives: from the computational modeling to the experimental data mainly collected by Small-Angle X-ray Scattering technique (SAXS) whose data were also used as constraint for modelling.

Part of the research was focused on the structural characterization of the protein SurE of *Xylella fastidiosa* (XfSurE) by experimental and computational techniques. Structural studies of XfSurE performed by SAXS technique indicated a tetrameric arrangement of the apo enzyme and to our knowledge, this was the first solution structure of a SurE protein described in the literature. When combined with computational techniques — for instance, normal mode analysis — the interpretation of SAXS analysis was enhanced. In that case, the experimental constraints imposed by the $I(q)$ curve allowed to reach a new structure model that fits the SAXS profile using only a single normal mode. This effect would be associated with the possible allosteric transitions of the XfSurE.

It was also developed a new program called SAXSTER (SAXS-assisted multi-source ThreadER). SAXSTER has the ability to generate more likely structural models for the target protein from optimal alignments obtained by threading and similar structures identified in the Protein Data Bank aided by SAXS. The program was tested on 553 nonredundant proteins. It was shown that SAXSTER can consistently improve the overall classification of the alignments, with p-values ranging from 10^{-6} to 10^{-8}. According to average TM-score, a more promising use of the SAXSTER algorithm would be to improve the template recognition results for protein whose structure is more rod-like than globular-like ones.

Índice

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

É esperado que o título deste trabalho de tese, — “Aplicação de técnicas de espalhamento de raios X na caracterização estrutural de proteínas e modelagem computacional utilizando vínculos experimentais obtidos por SAXS” — cumpra o seu papel em oferecer ao leitor uma ideia geral do que está por vir. Por se tratar de apenas um título, todas as balizas, limitações e potencialidades do trabalho aqui descrito terão a chance de serem descritas e discutidas somente a partir de agora.

O problema da caracterização estrutural de proteínas é extremamente vasto. Diversos níveis estruturais podem ser explorados utilizando-se várias técnicas experimentais e computacionais. O foco deste trabalho de tese está na caracterização da estrutura terciária e quaternária das proteínas. Para atingir este objetivo, métodos computacionais tradicionais são aplicados na resolução de problemas reais e também novos são desenvolvidos, com o intuito de aplicá-los na predição estrutural tridimensional de uma proteína, a partir da sua sequência de aminoácidos e de vínculos experimentais. Estes vínculos decorrem de curvas unidimensionais medidas pela técnica de espalhamento de raios X a baixos ângulos, uma técnica tradicional que permite o estudo de proteínas em solução. Outra técnica experimental utilizada e também referida genericamente no título por “espalhamento de raios X” é a cristalografia por difração de raios X, bastante consagrada na determinação da estrutura tridimensional de proteínas em alta resolução. No entanto, a aplicação desta técnica é mais pontual na tese.

De um ponto de vista mais amplo, a abordagem escolhida segue um fio condutor que permite explorar nuances entre resultados experimentais e computacionais em problemas de biologia es-

trutural, combinando-os quando possível. Especificamente, problemas independentes são abordados. Por esta razão, a divisão em capítulos contempla a respectiva descrição metodológica peculiar de cada assunto, os respectivos resultados e também as discussões apropriadas.

A seguir, os principais conceitos que permeiam todo o trabalho são apresentados de forma sucinta. Detalhes adicionais podem ser encontrados nas seções pertinentes em cada capítulo.

1.1 – Proteínas

Proteínas são macromoléculas biológicas que essencialmente desempenham funções vitais em um organismo [1]. O paradigma vigente preconiza a ideia da atividade estar intrinsecamente relacionada com a conformação tridimensional assumida de uma proteína, podendo até ocorrer a total perda de sua função biológica dependendo do grau de distorção de sua estrutura. Também tem sido discutida a intrigante possibilidade da função biológica proceder a partir de proteínas nativamente desenoveladas [2, 3].

A estrutura proteica é geralmente dividida em quatro níveis:

i) Estrutura primária ou sequência de aminoácidos: onde cada aminoácido é associado a outro por meio de ligações peptídicas (covalentes);

ii) Estrutura secundária ou conformações locais (hélices-α e fitas-β): estabilizadas ao longo da cadeia principal por ligações de hidrogênio (não-covalentes);

iii) Estrutura terciária: caracterizada pelo enovelamento da cadeia principal sobre si mesma resultando em um núcleo hidrofóbico e uma superfície hidrofílica;

iv) Estrutura quaternária ou estado oligomérico: formação de complexos pela associação de estruturas terciárias.

1.2 – Estrutura proteica: perspectiva histórica

As bases experimentais para o estudo estrutural de proteínas foram estabelecidas em meados do século XX. Com o advento da descoberta dos raios X por William Röntgen em 1895, Max Von Laue demonstrou em 1912, que cristais possuem a propriedade de apresentar o fenômeno da difração quando submetidos aos raios-X e com essa técnica pioneira, diversos estudos foram

iniciados na área de cristalografia.

Mais tarde, Linus Pauling foi o primeiro a propor motivos estruturais em proteínas como, por exemplo, as estruturas hélices-α e fitas-β [4]. Anos depois, Max Perutz demonstrou [5] que o problema das fases inerentes aos experimentos de difração poderia ser resolvido com substituição isomorfa múltipla — que basicamente trata da comparação dos padrões de difração de cristais de proteína na forma nativa com os padrões de difração de cristais de proteína embebidos em íons de metais pesados e em diferentes condições.

Michael Rossmann e colegas abordaram o problema das fases por uma técnica conhecida até hoje como substituição molecular [6] possibilitando o crescimento dos bancos de dados de estruturas de proteínas em uma velocidade jamais vista até então. A partir disto, novos experimentos puderam ser beneficiados por utilizarem estruturas já resolvidas como ponto de partida para o modelamento de novas.

Atualmente, o *Protein Data Bank* [7] — considerado o banco de dados mais popular — contém em seus registros cerca de 90 mil estruturas, das quais 88 $\%$ são estruturas que foram determinadas através da técnica de Cristalografia, enquanto Ressonância Magnética Nuclear (NMR) e Microscopia Eletrônica (EM) representam 11,5 $\%$ e 0,5 $\%$, respectivamente.

1.3 – Espalhamento de raios X a baixos ângulos (SAXS)

Espalhamento de raios X a baixos ângulos (SAXS, *Small-Angle X-ray Scattering*) é uma técnica experimental baseada na interação de raios X com a matéria [8, 9, 10, 11, 12], compartilhando o mesmo princípio físico básico da técnica de cristalografia por difração de raios X [13].

Diferentemente da cristalografia, SAXS fornece apenas um contraste do sinal entre a densidade eletrônica média da partícula e do seu ambiente químico sem qualquer informação sobre as posições atômicas. Consequentemente, SAXS fornece informação de baixa resolução sobre a molécula em estudo, muito embora seja uma técnica poderosa na determinação de estados oligoméricos em solução, mecanismos de enovelamento e desenovelamento e na elucidação de arranjos moleculares entre proteína-proteína.

A montagem experimental esquemática de um experimento de SAXS está ilustrada na Figura 1.1. Após um feixe de raios X incidir na amostra, ele é espalhado por um ângulo 2θ. A intensidade

espalhada próxima da direção do feixe incidente é mais intensa e à medida que o ângulo 2θ cresce, a intensidade decai rapidamente, dando sentido ao termo “baixos ângulos” presente no nome da técnica. O registro dos eventos de espalhamento é coletado por um detector onde após a redução dos dados, se obtém a informação da intensidade coletada em função do módulo do vetor de espalhamento definido por $q = 4\pi sin\theta/\lambda$, com λ representando o comprimento de onda da radiação incidente.

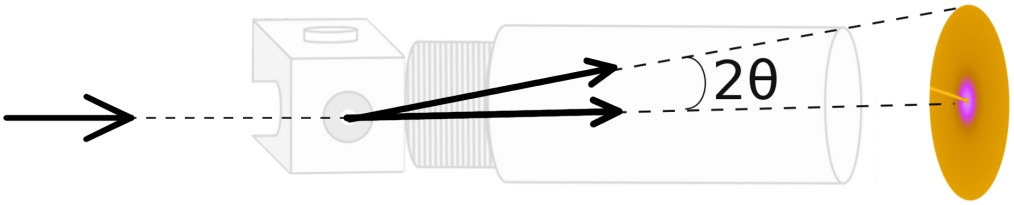


Figura 1.1: Aparato experimental esquemático de um experimento de SAXS. A seta à esquerda ilustra o feixe incidente em direção ao porta-amostra (cubo) e as setas à direita representam feixes espalhados na direção incidente e em um dado ângulo 2θ da direção original de incidência. Após percorrer a câmara de vácuo, a radiação espalhada é coletada por um detector.

Dado que as orientações moleculares são aleatoriamente distribuídas em solução, o padrão de intensidade é isotrópico e em última instância corresponde à uma curva unidimensional da intensidade coletada em função do módulo do vetor de espalhamento $I(q)$.

Se em uma dada condição experimental as moléculas espalhadoras tiverem a mesma forma tridimensional, assim como as distâncias intermoleculares serem grandes o suficiente para que efeitos de interferência inter-partícula possam ser considerados desprezíveis, a amostra é dita uma solução diluída. Se ambos estes requisitos são satisfeitos, é possível escrever a intensidade espalhada $I(q)$ para um dado módulo de vetor de espalhamento q em função da forma geométrica da molécula por

$$I(q) = 4\pi \int_0^{D_{max}} p(r) \frac{sin(qr)}{qr} dr, \tag{1.1}$$

onde q representa o módulo do vetor de espalhamento, 2θ o respectivo ângulo e λ o comprimento de onda da radiação incidente. O parâmetro D_{max} representa a máxima distância intramolecular

e a $p(r)$ é a função de distribuição de distância entre pares (PDDF, *Pair Distance Distribution Function*[1]).

Na prática, os experimentos de SAXS fornecem diretamente a curva $I(q)$ e, em princípio, uma transformada de Fourier inversa da Eq. (1.1) pode levar a obtenção da $p(r)$:

$$p(r) = \frac{r}{2\pi^2} \int_0^\infty I(q) q sin(qr) dq. \tag{1.2}$$

A curva unidimensional $p(r)$ pode ser entendida como uma função assinatura da forma geométrica do conjunto dos espalhadores arranjados no espaço tridimensional. Qualitativamente, a partir da forma de $p(r)$ pode-se avaliar a anisotropia do objeto espalhador como sendo, por exemplo, uma esfera, um elipsoide ou um cilindro.

1.4 – Bioinformática estrutural: um novo paradigma

Apesar dos esforços experimentais na determinação estrutural de proteínas, a velocidade de deposição ao longo dos anos comparada com as sequências proteicas determinadas por grandes centros de genômica e proteômica é altamente discrepante (Figura 1.2).

Faz-se necessário o uso de métodos computacionais frente a essa avalanche de informações atualmente disponíveis para que a investigação acerca das características estruturais de proteínas sejam exploradas. Um dos nichos de pesquisas e também de serviços do ramo de atividade conhecido por bioinformática estrutural, é a predição de enovelamentos a partir apenas da sequência de aminoácidos.

A predição computacional de uma estrutura proteica se divide em três categorias [15]: modelagem comparativa, *threading* e predição de novos enovelamentos. Na modelagem comparativa ou por homologia, a estrutura da proteína é predita pelo alinhamento da sequência alvo com sequências evolutivamente relacionadas e que tenham um modelo estrutural já resolvido. *Threading* vai

[1] Os termos "Função de pares", "Função radial de pares" ou mesmo "Função de distribuição de distâncias entre pares" muitas vezes se confudem. Em física da matéria condensada, por exemplo, a função radial de pares — muitas vezes representada por $g(r)$ — refere-se à distribuição de partículas do *bulk*, seja na fase sólida, líquida ou gasosa. Portanto, $g(r)$ trata da distribuição intermolecular da matéria. No entanto, o termo empregado aqui para a $p(r)$ — PDDF, sigla para *Pair Distance Distribution Function* — refere-se às distâncias intramoleculares de uma proteína.

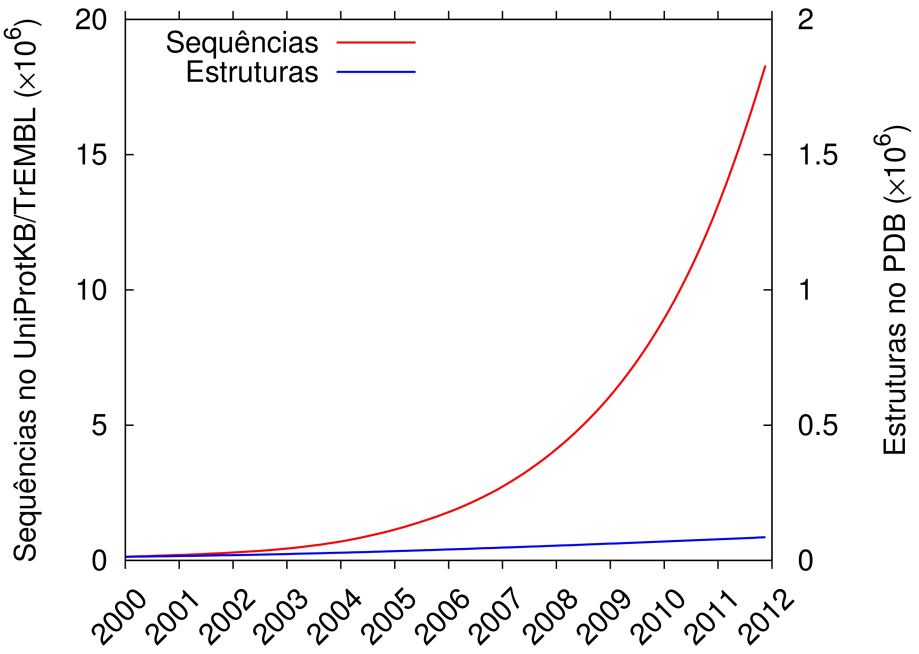


Figura 1.2: Total acumulado do número de sequências e estruturas de proteínas depositadas nos bancos de dados UniProt [14] e PDB [7] até o final do ano de 2012.

mais além da modelagem comparativa, pois as sequências usadas nos alinhamentos não precisam necessariamente estar evolutivamente relacionadas o que permite verificar enovelamentos semelhantes mesmo com baixa identidade sequencial. Já a predição de novos enovelamentos não se baseia em estruturas conhecidas e por isso é um tipo de modelagem que leva em consideração apenas princípios físicos e/ou bioquímicos assim como dados estatísticos coletados em bancos de dados, e por isso é muitas vezes chamada de modelagem *ab initio*[2].

1.5 – A tese e seus objetivos

O tema deste trabalho de doutorado é centrado na Biologia Estrutural e possui duas abordagens complementares: uma experimental e outra computacional.

[2] O termo *ab initio* utilizado aqui é normalmente entendido como uma abordagem computacional onde não se utiliza nenhuma informação adicional sobre a proteína de interesse que possa guiar o processo de modelagem. Há, portanto, uma sobreposição de significado com o tradicional termo utilizado no âmbito das ciências básicas onde a abordagem ao problema requereria o uso de leis fundamentais apenas.

A partir de técnicas experimentais de espalhamento de raios X, tais como SAXS e cristalografia por difração de raios X, estudamos características estruturais de uma proteína de interesse.

O aspecto computacional abrange desde a aplicação de métodos e algoritmos existentes ao desenvolvimento de novos programas na resolução de problemas em Biologia Estrutural.

Objetivos específicos

- Caracterização estrutural da proteína XfSurE de *Xylella fastidiosa* relacionada à doença CVC dos citros por SAXS, cristalografia e análise de modos normais de vibração;
- Desenvolvimento de um programa com abordagem híbrida que utiliza *threading* (alinhamento de sequências de aminoácidos em um estrutura protéica) e vínculos experimentais de SAXS para predição do enovelamento de proteínas.

1.6 – Organização

A divisão em capítulos contempla a respectiva descrição metodológica peculiar de cada assunto, os respectivos resultados e também as discussões apropriadas.

O Capítulo 2 descreve a caracterização estrutural de uma proteína da bactéria *Xylella fastidiosa* através de duas técnicas experimentais: espalhamento de raios X a baixos ângulos e cristalografia, muito embora as análises em solução tenham sido predominantes. Modelagem computacional também foi realizada à luz dos modos normais de vibração das estruturas obtidas.

O Capítulo 3 — fruto do conjunto de pesquisas paralelas ao trabalho de tese — também serve como um preâmbulo para o Capítulo 4. Neste estudo, foram aplicados métodos de recuperação da informação no reconhecimento de formas de objetos tridimensionais com o propósito de ilustrar uma estratégia computacional de busca por formas semelhantes, dado um objeto-alvo, a partir de métricas inspiradas em SAXS.

O objetivo do Capítulo 4 foi o desenvolvimento de um programa, denominado SAXSTER, que tem a habilidade de gerar modelos estruturais mais prováveis para uma proteína-alvo, a partir de alinhamentos ótimos obtidos por *threading* e de estruturas similares identificadas no *Protein Data Bank*, com o auxílio de dados experimentais obtidos pela técnica de espalhamento de raios X a baixos ângulos. As considerações finais são apresentadas no Capítulo 5.

Capítulo 2

Caracterização estrutural da proteína SurE de *Xylella fastidiosa*

O trabalho descrito neste capítulo foi fruto de um projeto colaborativo entre o nosso grupo do IQ com o grupo da Prof^{a}. Dr^{a}. Anete Pereira de Souza do Centro de Biologia Molecular e Engenharia Genética (CBMEG) da UNICAMP, responsável pela clonagem, expressão e purificação da proteína de interesse e também com o grupo do Prof. Dr. Jorge Iulek da Universidade Estadual de Ponta Grossa, onde os refinamentos das estruturas cristalográficas foram finalizados.

O presente estudo consistiu na caracterização estrutural da proteína SurE de *Xylella fastidiosa* (XfSurE) através de duas técnicas experimentais: espalhamento de raios X a baixos ângulos (SAXS) e cristalografia, muito embora as análises em solução tenham sido predominantes. Os dados de SAXS apontaram para um arranjo tetramérico de XfSurE e, do nosso conhecimento, foi a primeira estrutura em solução descrita na literatura para esta família de proteínas. Parte das análises em solução e a caracterização cristalográfica preliminar foram publicadas nos artigos *Functional and small-angle X-ray scattering studies of a new stationary phase survival protein E (SurE) from Xylella fastidiosa-evidence of allosteric behaviour* [16] e *Crystallization and preliminary X-ray analysis of stationary phase survival protein E (SurE) from Xylella fastidiosa in two crystal forms* [17].

Diferenças conformacionais foram observadas entre as estruturas cristalográficas obtidas motivando-nos a realizar análises computacionais à luz dos modos normais de vibração e SAXS, com o objetivo de caracterizar os rearranjos das subunidades de XfSurE. Estes achados tornam-

se importantes à medida que XfSurE é sabidamente uma proteína alostérica e movimentos de domínios podem estar relacionados ao mecanismo de regulação.

2.1 – Introdução

Bactéria *Xylella fastidiosa*

A bactéria *Xylella fastidiosa* é o agente causador de diversas doenças em plantas economicamente importantes para diversos países. Entre tais doenças, destaca-se a Clorose Variegada dos Citrus (CVC) [18, 19, 20] mais conhecida como “amarelinho” e atinge todas as variedades comerciais de citros. O modo de transmissão para a planta se dá por insetos hospedeiros — várias espécies de cigarrinhas — que, ao se alimentarem no xilema (tecido condutor) da planta contaminada, transmitem a bactéria para plantas sadias. Como a bactéria é restrita ao xilema, sua presença, consequentemente, promove a obstrução dos vasos responsáveis pelo transporte de nutrientes da raiz até o fruto. Do ponto de vista econômico, a CVC é a doença mais prejudicial à citricultura brasileira afetando principalmente pomares de laranjas doces (Figura 2.1) que uma vez contaminados pela doença, desenvolvem frutos pequenos, rígidos e com amadurecimento precoce, impróprios para a comercialização.

Figura 2.1: Comparação entre laranjas sadias e laranjas afetadas pela bactéria *Xylella fastidiosa* mediante a doença clorose variegada em citrus (CVC), conhecida popularmente por “amarelinho”. Fonte: `http://www.ars.usda.gov/Aboutus/docs.htm?docid=16790`, Acesso em 11/12/2013.

Proteína SurE

A sequência genômica desta bactéria [21] revela que diversas proteínas estão associadas aos possíveis mecanismos de patogenicidade. Entre as diversas proteínas possivelmente envolvidas com a doença, encontra-se a *stationary phase survival protein E* (SurE), ou simplesmente XfSurE. O gene *surE* é amplamente distribuído entre archaea, eubactérias, eucariotos e aparentemente é bem conservado [22].

Não existe um consenso a respeito do estado oligomérico das proteínas SurE. A partir de estruturas cristalográficas de *Thermotoga maritima*, foram encontrados dímeros [23] e dímeros + tetrâmeros [24] na unidade assimétrica (ASU). Já em *Pyrobaculum aerophylum*, *Campylobacter jejuni* e *Salmonella typhimurium* dímeros são encontrados na ASU [22, 25, 26]. Também foi descrito em *Thermus thermophylus* a coexistência entre dímeros e tetrâmeros [27]. Para a bactéria *Escherichia coli*, ensaios de gel filtração mostraram a presença de um oligômero tetramérico de SurE [28].

O alinhamento sequencial da SurE de *Xylella* (Figura 2.2) revela diversas regiões conservadas, especialmente o aminoácido ASP-8 que estaria diretamente relacionado à catálise. Já na Figura 2.3, o arranjo tetramérico cristalográfico de proteínas SurEs em diferentes organismos pode ser observado, em particular, a região do sítio ativo destas proteínas que está localizada no centro do tetrâmero.

Função da SurE

A função da proteína SurE ainda não é completamente entendida, entretanto ela é usualmente classificada como nucleotidase devido à grande especificidade a nucleosídios monofosfatados. Nucleotidases ou nucleosídeo monofosfato fosfohidrolases (EC 3.1.3.5 ou 3.1.3.6) são fosfatases que especificamente desfosforilam nucleosídeos monofosfatados em nucleosídeos e fosfato inorgânico [28]. Tal função contribui, entre outras coisas, para a manutenção do balanço correto dos *pools* de nucleotídeos na célula [29]. Assim, especula-se que a enzima SurE estaria envolvida ou na regulação da síntese de DNA e RNA ou no catabolismo de nucleosídios não-canônicos.

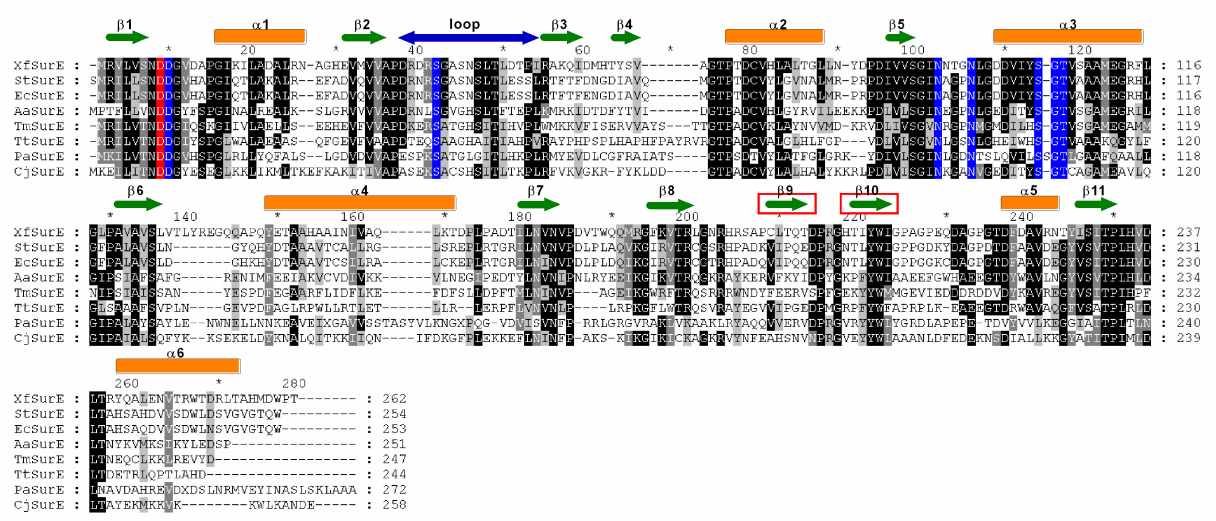


Figura 2.2: Alinhamento sequencial de XfSurE com outras proteínas SurE de diferentes organismos. Em destaque (vermelho) encontra-se o aminoácido ASP-8, envolvido no mecanismo de catálise desempenhado pelas SurEs.

Proteínas alostéricas

Enzimas são seletivas no reconhecimento das moléculas (substrato) e específicas na reação que catalisa. Em muitos casos, dependem de compostos não proteicos (cofatores) para desempenhar sua função, como íons metálicos ou moléculas orgânicas. No caso de íons metálicos, estes atuam como catalisadores eletrofílicos, estabilizando o aumento da densidade de elétrons ou a carga negativa que se desenvolve durante a reação. Outra provável função é prover um poderoso nucleófilo em pH neutro. Em enzimas alostéricas [30, 31, 32], a ligação do substrato ao sítio ativo pode afetar a propriedade de ligação de outro sítio ativo da enzima. Isto possivelmente resulta da interação entre suas subunidades, tornando a ligação do substrato cooperativa.

No modelo clássico de Monod-Wyman-Changeux (MWC) [33] para proteínas alostéricas, uma enzima é modelada a partir de duas configurações: a relaxada (R) e a tensionada (T). O estado R está associado à ausência de ligantes no sítio ativo e o estado T possui uma alteração conformacional que induziria a cooperatividade, mediada por transições alostéricas que por sua vez, modulariam a afinidade do substrato no sítio. O objetivo do modelo de MWC é estabelecer uma equação que associe a fração do número médio de ligantes aderidos ao sítio em termos

da concentração de ligantes livres em termos das constantes de equilíbrio. Consequentemente, o modelo MWC acaba sendo uma ferramenta muito útil no contexto da cinética enzimática para distinguir proteínas alostéricas de não alostéricas.

Uma caracterização preliminar mostrou que a XfSurE também apresenta propriedade alosté-

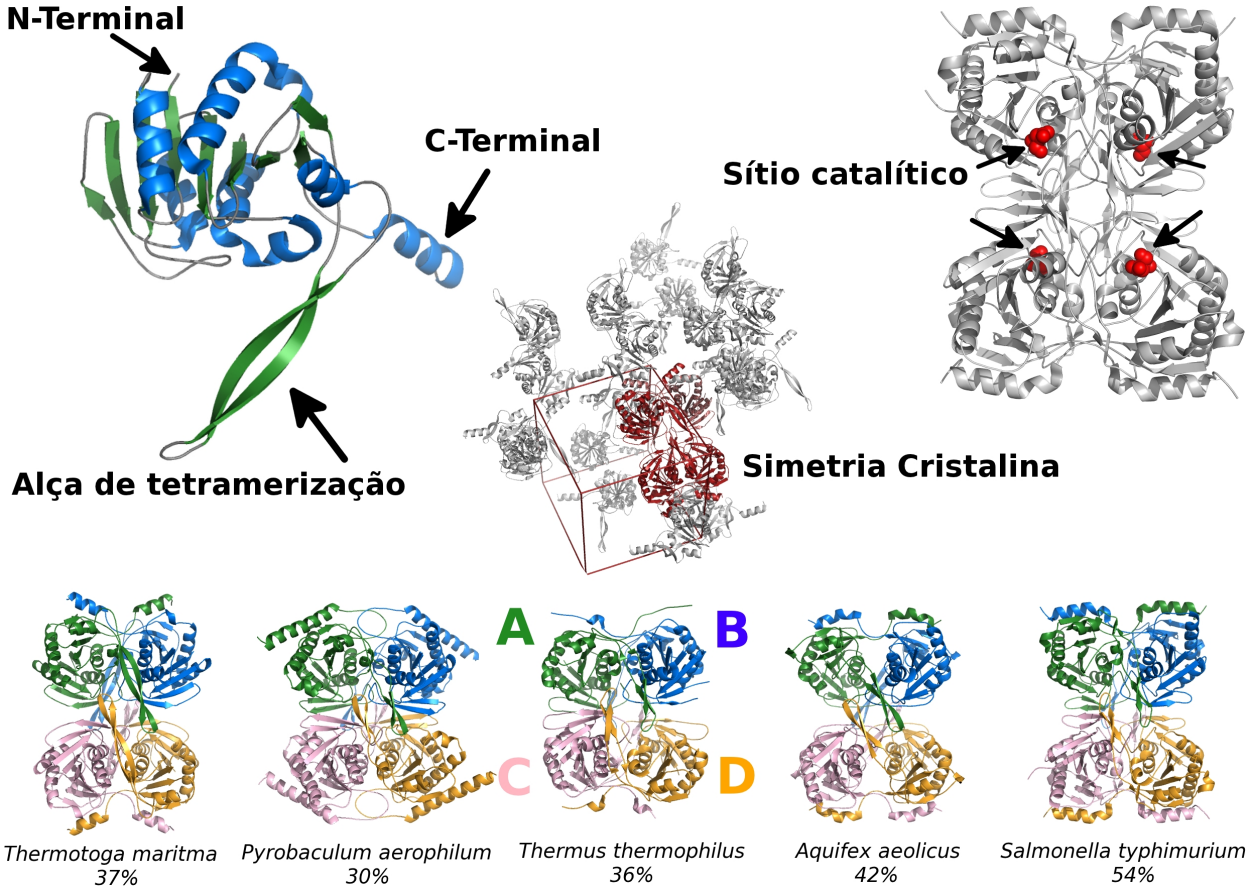


Figura 2.3: Estrutura das proteínas SurE em diversos organismos. Destaque à esquerda para o monômero de SurE composto por três domínios típicos: enovelamento tipo Rossmann, alça β-*hairpin* e hélice-α no C-terminal. A unidade biológica putativa é construída por operações de simetria a partir da estrutura cristalográfica cujos códigos PDB a partir da estrutura de *Thermotoga maritma* são: 1J9J, 1L5X, 2E69, 2PHJ e 2V4O. As identidades sequenciais percentuais com XfSurE também são mostradas. As subunidades foram nomeadas arbitrariamente de A-D. À direita, o destaque dos quatro sítios ativos do tetrâmero onde cada conjunto de esferas vermelhas representa o aminoácido ASP na posição 8 da sequência.

ricas [16]. No caso da enzima SurE de *Xylella fastidiosa*, o cofator metálico é o íon Mn^{2+} e o substrato de maior afinidade é o 3'-AMP. Em geral, cada monômero de SurE possui um sítio ativo, que na estrutura tetrâmerica é localizado na região central do tetrâmero (Figura 2.3).

Estudos de conformações de proteínas alostéricas por SAXS

Diversos estudos são conduzidos na investigação dos efeitos alostéricos de proteínas pela técnica de SAXS [34, 35, 36, 37]. Esta técnica é poderosa à medida que a enzima encontra-se em solução e deste modo seria possível detectar possíveis arranjos entre subunidades ou deformações locais que estariam relacionadas com transições alostéricas. Neste sentido, estas abordagens têm cunho mais estrutural que cinético e procuram abrir novas frentes de trabalho a partir da caracterização estrutural de estados do tipo "relaxado" e "tensionado".

Análise por modos normais de vibração

A dinâmica de uma proteína abrange uma ampla gama de escalas temporais: femtossegundos a segundo, e escalas espaciais: flutuações atômicas a movimentos coletivos de domínios. Uma das técnicas mais bem difundidas ao estudo computacional destes eventos espaciais e temporais é a Dinâmica Molecular (DM) [38]. No entanto, as simulações de DM atomística de uma proteína são tradicionalmente investigadas em uma escala temporal tipicamente compreendida no intervalo de dezenas a centenas de nanossegundos, principalmente por limitações dos recursos computacionais disponíveis para lidar com o grande número de graus de liberdade do sistema proteína + solvente.

Movimentos de grande amplitude podem estar relacionados a transições conformacionais do tipo estado aberto$\leftrightarrow$estado fechado. Em diversos estudos deste tipo com outras proteínas, os movimentos relacionados às transições alostéricas estão associados à escala temporal da ordem de nanossegundos - milissegundos [39, 40].

Diante disto, modelos simplificados usualmente conhecidos por *coarse-grained* ou grosseiros têm sido desenvolvidos alternativamente como uma ferramenta para o estudo de movimentos de grandes amplitudes de domínios constituintes da estrutura proteica que a princípio, necessitariam de um tempo computacional proibitivo para serem observados em simulações de DM. Por outro lado, estes modelos são demasiadamente simplificados e os resultados devem ser analisados

com cautela.

Um desses modelos computacionais simplificados utiliza a Análise de Modos Normais (AMN) de vibração da estrutura de uma proteína a partir de uma rede elástica formada apenas pelos carbonos alfa dos resíduos de aminoácidos [41, 42, 43]. Os modos vibracionais abrangem uma extensa gama de frequências dependendo do tamanho do sistema. Os modos associados a vibrações de baixa frequência estão associados aos movimentos coletivos de domínios estruturais e os de alta frequência geralmente estão associados a movimentos não correlacionados de domínios. Contudo, devido à simplicidade desta técnica, somente correlações espaciais podem ser analisadas e nada pode ser inferido na escala temporal.

2.1.1 – Objetivos

Neste capítulo, o objetivo geral foi de caracterizar estruturalmente a proteína XfSurE de *Xylella fastidiosa* relacionada à doença CVC dos citros por SAXS, cristalografia e análise de modos normais de vibração. Os objetivos específicos são:

- Determinação do estado oligomérico em solução da proteína XfSurE na forma nativa;
- Cristalização e coleta de dados da XfSurE;
- Descrição e correlação dos modos normais de oscilação da XfSurE a partir de sua estrutura cristalográfica e dados de SAXS de XfSurE na forma nativa e com ligantes.

2.2 – Metodologia

2.2.1 – Obtenção das amostras

O protocolo de expressão e purificação da XfSurE foi estabelecido pelo grupo da Prof^{a}. Dr^{a}. Anete Pereira de Souza (IB/CBMEG/UNICAMP) durante o trabalho de doutorado de Antonio Saraiva [44, 16] o qual também realizou as análises funcionais e de cinética enzimática mencionadas ao longo do texto. A partir das amostras fornecidas, procedemos às análises por SAXS e cristalografia.

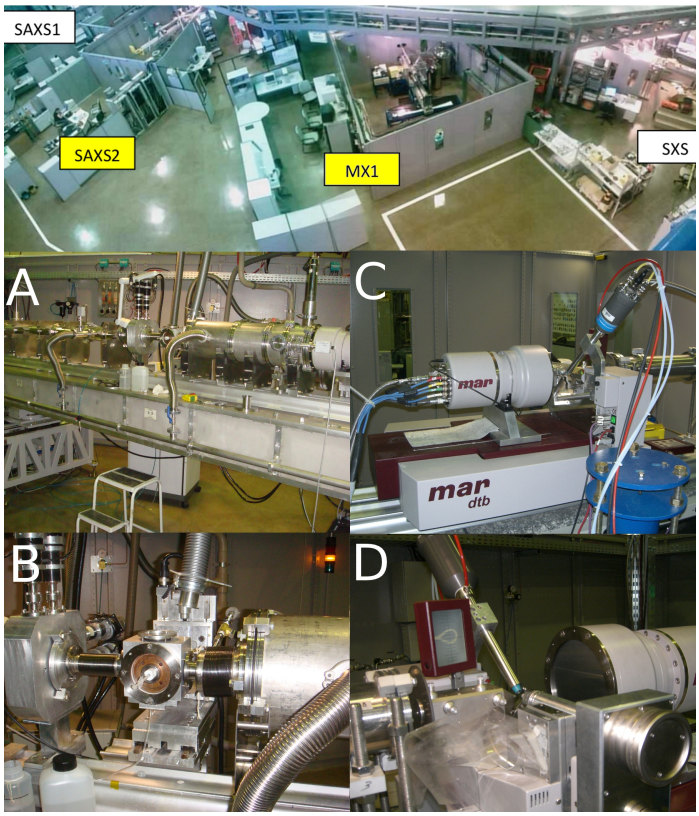


Figura 2.4: Linhas de luz SAXS2 e MX1 do LNLS. (A) Aspecto geral da linha SAXS2, com detalhe do porta amostra para líquidos que são injetados por uma seringa (centro da foto B). (C) Detector MarCCD da linha MX1 e soprador de nitrogênio, com detalhe da região onde o cristal é afixado no goniômetro (centro da foto D) por um *loop* de *nylon*. O monitor de visualização do *loop* também pode ser visto na foto.

2.2.2 – Procedimento experimental para as medidas de SAXS

Preparação das amostras

Para as medidas de SAXS, as amostras de XfSurE na forma nativa foram preparadas com concentrações no intervalo 2,0-12,1 mg/mL em tampão 25 mM de Tris-HCl, pH 7,5, 50 mM de NaCl e 1 mM de ditiotreitol (DTT[1]). As curvas de SAXS foram coletadas na linha de luz D02A-SAXS2 do Laboratório Nacional de Luz Síncrotron (LNLS, Campinas-SP, Brasil), utilizando um detector

[1]Para manter o ambiente redutor similar ao ambiente celular e assim, evitando ligações inespecíficas tipo S-S.

bidimensional MarCCD (Marresearch, Hamburgo, Alemanha). Os dados de XfSurE para a forma nativa foram coletados com a distância amostra-detector de 1306,85 mm e comprimento de onda dos raios X de 1,488 $\mathrm{\AA}$. O intervalo coberto para o vetor de espalhamento foi de $0,01 < q < 0,24\mathrm{\AA}^{-1}$. As medidas foram realizadas com 3 minutos de exposição aos raios X na temperatura de 20^0C. Para cada concentração, 5 medidas sucessivas foram registradas. A redução dos dados foi realizada pelo programa Fit2D [45]. Como nenhum dano por radiação foi observado, realizamos médias de cinco medidas com o intuito de otimizar a razão sinal-ruído experimental.

As medidas para XfSurE na presença de aditivos foram realizadas em outro momento e com condições experimentais sutilmente diferentes: concentração 10 mg/mL de XfSurE no mesmo tampão já descrito (controle) e XfSurE com um dos três aditivos: 0,5 mM Mn^{2+}, 1 mM PO_4^{3-} e 100 mM 3'-AMP. A distância amostra-detector foi de 1028,37 mm, $0,028 < q < 0,22\mathrm{\AA}^{-1}$. Processamentos preliminares dos dados de SAXS da XfSurE na presença de 3'-AMP foram realizados considerando médias de cinco medidas sucessivas a cada 3 minutos de exposição. Isso significa dizer que a quinta medida foi exposta aos raios X durante 15 minutos. Curiosamente, algumas medidas apenas para a XfSurE na presença de 3'-AMP com mais de 9 minutos de exposição apresentaram sinais semelhantes a danos por radiação, devido ao surgimento de uma mudança sistemática apenas no final da curva $I(q)$ em torno de $q \approx 0,2\mathrm{\AA}^{-1}$ embora o início da curva — relacionada à região de Guinier — não tenha apresentado variações. É importante enfatizar que este tipo de efeito não foi visto em amostras de XfSurE nativa ou com os íons Mn^{2+} e $PO_{4^{3-}}$ mesmo com 15 minutos de exposição aos raios X. Logo, todas as análises feitas com a amostra XfSurE + 3'-AMP foram realizadas com curvas coletadas com apenas 3 minutos de exposição (amostra e tampão) de forma a evitar o possível efeito de dano por radiação.

Correção das medidas devido à absorção de raios X e ao espalhamento do solvente

Pretende-se que o sinal medido seja apenas proveniente da proteína. Para lidar com o espalhamento devido ao solvente e também com absorção da amostra (proteína + solvente), uma medida de SAXS é feita em duas etapas. Primeiro mede-se a amostra e em seguida o "branco", que é justamente o tampão onde a proteína se encontra e que portanto, deve ter o seu sinal de espalhamento subtraído do sinal medido para a amostra. Para descontar as contribuições de

espalhamento não desejadas, fazemos

$$I(q) = \left(\frac{I_{am}(q) \cdot at_{am}}{I_{am}^{int}} - \frac{I_{sol}(q) \cdot at_{sol}}{I_{sol}^{int}} \right), \tag{2.1}$$

onde $I_{am}(q)$ e $I_{sol}(q)$ são as intensidades coletadas em função de q, para a amostra e o solvente, respectivamente. Visto que a intensidade do feixe incidente pode variar durante o período de coleta, I_{am}^{int} e I_{sol}^{int} são fatores de normalização integrados no tempo da respectiva medida e at_{am} e at_{sol} são os fatores de atenuação do feixe de raios X pela absorção da amostra. Devido à natureza do detector utilizado, um sinal é adicionado ao CCD (*Charge-Coupled Device*) antes mesmo da medida ocorrer. No caso da linha SAXS2 do LNLS este sinal é arbitrado em um valor constante igual a dez [46]. Logo, antes de aplicarmos a correção imposta pela Equação 2.1, subtraímos esta constante das medidas $I_{am}(q)$ e $I_{sol}(q)$.

Parâmetros estruturais a partir das curvas de SAXS

O raio de giro foi obtido pela análise de Guinier referente à região do início da curva $I(q)$. Partindo da lei de Guinier [8]:

$$I(q) = I(0) exp\left(-\frac{R_G^2}{3} q^2 \right), \tag{2.2}$$

obtemos R_G e também a intensidade extrapolada para $q = 0$ ($I(0)$) pela linearização da curva em um gráfico $LnI(q)$ vs q^2 para "ângulos" que satisfazem $q \cdot R_G \leq 1.3$.

Estimativas de R_G e $I(0)$ também foram feitas pelo programa GNOM [47] que calcula a transformada de Fourier indireta da curva $I(q)$ resultando na função conhecida por distribuição de distâncias $p(r)$. Assim, a partir da curva $p(r)$ o raio de giro pode ser calculado pela expressão:

$$R_G^2 = \frac{\int_0^{D_{max}} r^2 p(r) dr}{2 \int_0^{D_{max}} p(r) dr}, \tag{2.3}$$

com D_{max} sendo uma estimativa para a máxima distância intramolecular obtida pela $p(r)$.

Outros critérios qualitativos também são muito úteis na caracterização por SAXS da proteína em solução [8]. A Figura 2.5-A mostra o comportamento esperado da curva $p(r)$ para diversos objetos tridimensionais que fazem o papel de formas idealizadas assumidas pelas proteínas. Consequentemente, a partir da "assinatura" da $p(r)$, pode-se inferir qualitativamente a anisometria da

partícula espalhadora de raios X. Por exemplo, se a partícula em solução for esférica, a $p(r)$ será altamente simétrica em relação ao centro[2]; se a partícula for prolata, a $p(r)$ correspondente terá o máximo global deslocado para esquerda ($r \to 0$) e um decaimento linear até ($r \to D_{max}$), enquanto que uma partícula com uma cavidade em seu interior, o máximo da $p(r)$ será deslocado para a direita.

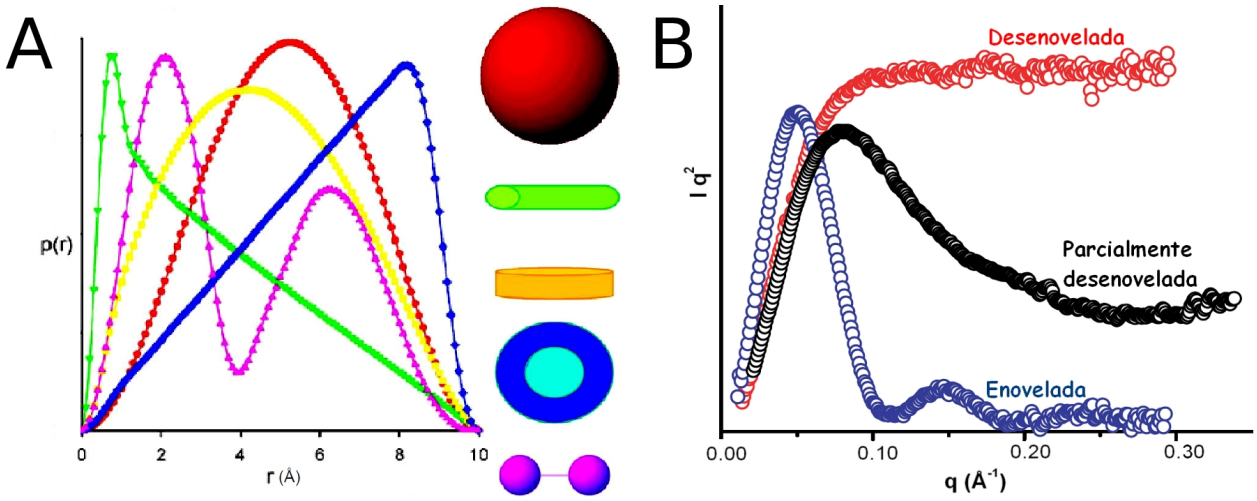


Figura 2.5: Critérios qualitativos auxiliares na caracterização de proteínas por SAXS. (A) Função de distribuição de distâncias $p(r)$. (B) Gráfico de Kratky $I(q) \times q^2$ vs q. Figuras adaptadas das referências [9, 10].

Um outro critério qualitativo é o gráfico de Kratky (Figura 2.5-B) que é preparado diretamente dos dados da curva de intensidade em função do módulo do vetor espalhamento $I(q)$, pelo gráfico $I(q)\times q^2$ vs q. Espera-se que uma proteína devidamente enovelada exiba uma "forma de sino" com um máximo bem definido no gráfico de Kratky. Diferentes estágios de desenovelamento podem ser verificados pelo grau de extinção do máximo da curva de Kratky sendo que a forma característica de "sino" é completamente extinta no caso de proteínas completamente desenoveladas.

O gráfico de Porod ($I(q)\times q^4$ vs q) também pode ser usado com uma terceira análise qualitativa. Idealmente, partículas globulares com densidade eletrônica uniforme e com interface bem definida com a densidade eletrônica do solvente exibem um decaimento proporcional a q^{-4} em "altos"

[2]De fato, a distribuição de distâncias para uma esfera perfeita pode ser calculada analiticamente tendo por resultado $p(r) = 12x^2(2 - 3x + x^3)$, com $x \equiv r/D_{max}$ cujo máximo se localiza em $x \approx 0.525$.

ângulos. Porod mostrou que o comportamento assintótico no final da curva é dado por

$$I(q) \rightarrow (\Delta\rho)^2 \frac{2\pi}{q^4} S, \tag{2.4}$$

onde $\Delta\rho$ é o contraste da densidade eletrônica entre a proteína e o solvente e S a superfície da partícula.

Portanto, espera-se que o final da curva no gráfico de Porod aproxime-se assintoticamente a um valor constante que é proporcional à superfície da partícula. No caso de proteínas, a densidade eletrônica é considerada aproximadamente uniforme assim como a densidade da camada de hidratação ao redor do seu volume. Entretanto, devido à diferença da densidade eletrônica na interface da proteína e do solvente ser muito suave, é difícil observar a região de Porod de maneira inequívoca. Quando possível, o que pode ser observado experimentalmente são padrões oscilatórios no gráfico de Porod na região de "altos" ângulos em torno de um valor constante. Casos onde a densidade eletrônica é heterogênea, uma região linear na região de "altos" ângulos no gráfico de Porod surge na forma $I(q) \cdot q^4 \approx k_1 I(q) \cdot q^4 + k_2$, com k_1 e k_2 constantes [8].

Estimativa de massa molecular e estado oligomérico

Uma estimativa para a massa molecular da proteína foi obtida usado Lisozima como proteína padrão a 27 mg mL^{-1}, coletada nas mesmas condições experimentais de XfSurE.

A massa molecular de XfSurE (M_{XfSurE}) foi estimada pela relação:

$$M_{XfSurE} \approx \left(\frac{I_{XfSurE}(0)}{I_{Liso}(0)}\right)\left(\frac{c_{Liso}}{c_{XfSurE}}\right) M_{Liso}, \tag{2.5}$$

onde c_{Liso} e c_{XfSurE} são as concentrações das proteínas e $M_{Liso} \approx 14kDa$. Este método é capaz de fornecer estimativas de massa com um erro de aproximadamente 10% [48]. O número de subunidades n de XfSurE foi obtido a partir da estimativa da massa da proteína em solução M_{XfSurE} e da massa de um monômero de proteína M_{mono}, predito pela soma das massa individuais dos aminoácidos constituintes. A simples relação n = M_{XfSurE} / M_{mono} portanto, define o estado oligomérico em solução. Por exemplo, dizemos que há um dímero em solução se $n \approx 2$ ou um tetrâmero, se $n \approx 4$.

Ajustes teóricos

O ajuste de curvas teóricas com a $I(q)$ experimental assim como a obtenção da curva $p(r)$ foram realizadas pelos programas CRYSOL [49], GNOM [47] e SAXSTER[3] [50] (Capítulo 4). Em particular, CRYSOL necessita de um modelo de proteína com todos os átomos enquanto que SAXSTER usa apenas os carbonos alfa dos resíduos de aminoácidos das estruturas. Assim, é bastante conveniente dispormos de ambos os programas CRYSOL e SAXSTER, embora apenas o segundo possa também calcular a $p(r)$ teórica diretamente da estrutura proteica além da $I(q)$.

A qualidade do ajuste entre duas curvas $I(q)$ é dada pela métrica χ, defina como:

$$\chi = \sqrt{\frac{1}{n}\sum_{k=1}^{n}\left\{\frac{I_{exp}(q_k) - cI_{modelo}(q_k)}{\sigma(q_k)}\right\}^2}, \tag{2.6}$$

onde I_{exp} e I_{modelo} representam a intensidade de espalhamento experimental e a calculada para um modelo estrutural, respectivamente. A constante c é um fator de escala analiticamente calculado de forma a minimizar a expressão χ^2 como um todo, impondo $\partial\chi^2/\partial c = 0$. A função $\sigma(q)$ representa o erro experimental de $I_{exp}(q)$. O erro para uma curva teórica é estimado pela função $\sigma(q) = I(q) \times (q + 0,15) \times 0,30$ [51][4].

Modelos estruturais de baixa resolução

Modelos *ab initio* para os envelopes moleculares foram gerados a partir de curvas experimentais usando DAMMIF, DAMAVER e DAMFILT [52, 53, 54]. DAMMIF parte de uma distribuição de átomos fictícios dispostos homogeneamente no volume de uma esfera com diâmetro D_{max} obtido por SAXS, conforme ilustração da Figura 2.6. A partir de uma simulação de Monte Carlo, a distribuição tende a adquirir uma configuração espacial cuja respectiva curva teórica ajusta a curva $I(q)$ experimental. Como o procedimento é estocástico, vários modelos devem ser gerados

[3]O programa SAXSTER desenvolvido neste trabalho de doutorado possui diversos módulos sendo que em um deles é realizada a simulação de dados de SAXS, seja para a obtenção da $I(q)$ ou da $p(r)$. Neste capítulo, nos referimos a este módulo específico simplesmente pela denominação "SAXSTER".

[4]Nesta referência, os autores desenvolveram um programa de simulação de curvas de SAXS a partir de modelos simplificados. Confrontando os resultados produzidos com esse programa com os resultados obtidos através do clássico programa da área, CRYSOL [49], observou-se que $\sigma(q)$ é uma função erro tal que ambos os programas produzem resultados similares.

para que posteriormente, o programa DAMAVER sobreponha todos os modelos dois a dois com o objetivo de estabelecer uma validação cruzada que permita selecionar os modelos que comporão a média dessas configurações. Diferentes simetrias também podem ser inseridas como vínculos no processo de modelagem. Em nossas simulações, um total de 20 modelos e a imposição das simetrias P1, P2 e P222 mostraram-se suficientes à convergência dos resultados. Para uma visualização mais clara do modelo produzido por DAMMIN, o envelope molecular também é representado por uma superfície que engloba os átomos fictícios, cujo cálculo é realizado pelo programa NCSMASK [55].

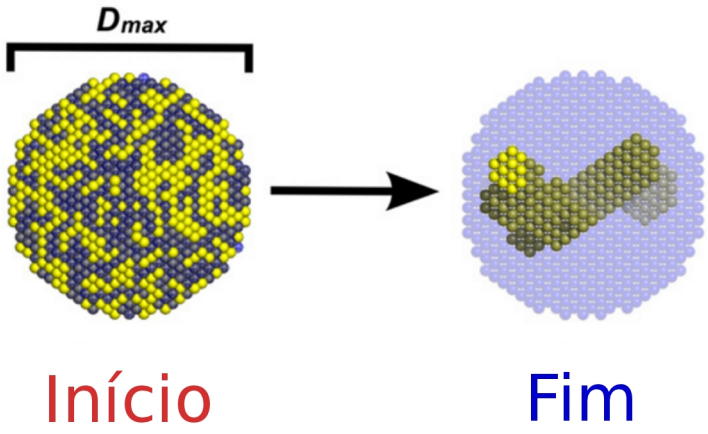


Figura 2.6: Ilustração do método usado pelo programa *ab initio* DAMMIN. A partir de um volume inicial, as configurações dos átomos fictícios associados à proteína (esferas amarelas) e associados ao solvente (esferas azuis), são alteradas até que esse procedimento produza um modelo final cuja intensidade de espalhamento teórico ajuste os dados experimentais e seja formado por uma configuração de átomos fictícios compactados. Imagem adaptada da figura encontrada no artigo [11].

Em paralelo aos métodos *ab initio* utilizados para reconstruir a forma da proteína em solução, também realizamos simulações de corpo rígido, manipulando as subunidades de XfSurE por translações e rotações na busca de um melhor ajuste χ com os dados experimentais. Para isto, escrevemos diversos *scripts* em linguagem Perl e Python (PyMOL) para automatizar todo o processo de análise.

Todas as figuras foram preparadas com os programas GIMP (`http://www.gimp.org/`),

GNUPLOT (`http://gnuplot.sourceforge.net/`) e PyMOL (`http://www.pymol.org`).

2.2.3 – Procedimento experimental para cristalização e coleta de dados

Há vasta literatura disponível sobre as técnicas experimentais mais utilizadas em cristalografia [13, 56, 57]. Portanto, apenas os princípios necessários para a compreensão dos resultados serão apresentados.

Diferentemente da técnica de SAXS, dados de difração de raios X por cristais podem resultar em modelos estruturais de proteína com resolução da ordem de 1 $\mathrm{\AA}$. Entretanto, o grande gargalo na aplicação desta técnica é justamente a obtenção de cristais pertinentes que, por muitas vezes, é uma atividade empírica e sem garantias de sucesso.

Quando o projeto em colaboração se iniciou, já eram conhecidas algumas condições de cristalização da proteína XfSurE. Essas condições foram inicialmente investigadas com um robô pipetador Honeybee 963 (Genomic Solutions) presente no LNLS, usando o método de difusão por vapor em gota sentada em uma placa de 96 poços. Estas condições foram encontradas a 293 K a partir de centenas de tentativas usando *kits* comerciais como Crystal Screen, Crystal Screen 2 e SaltRx (Hampton Research), Wizard I e II Precipitant Synergy (Emerald BioSystems), PACT e JCSG + (NeXtal/Qiagen).

Nossa participação se iniciou com a preparação de experimentos de otimização das condições promissoras onde foram identificados indícios de cristalização. O método baseado na matriz esparsa [57] foi empregado variando os valores de pH, a concentração do agente precipitante e sal e adicionando aditivos e substratos. A proteína foi preparada com concentrações de 10 mg/mL no mesmo tampão descrito anteriormente para as coletas de SAXS: 25 mM de Tris-HCl, pH 7,5, 50 mM de NaCl e 1 mM de ditiotreitol (DTT). Realizamos os experimentos de cristalização com placas TPP de 24 poços usando o método de difusão de vapor com a gota suspensa (Figura 2.7). O reservatório foi preparado com volume de 500 μL de solução enquanto que gotas de 2 μL da solução de proteína foram diluídas em 2 μL de solução precipitante (solução do poço). As lamínulas de vidro foram previamente silanizadas com clorotrimetilsilano (Sigma-Aldrich). O tempo médio de crescimento dos cristais foi de aproximadamente duas semanas em ambiente com temperatura controlada a 18 ^{0} C.

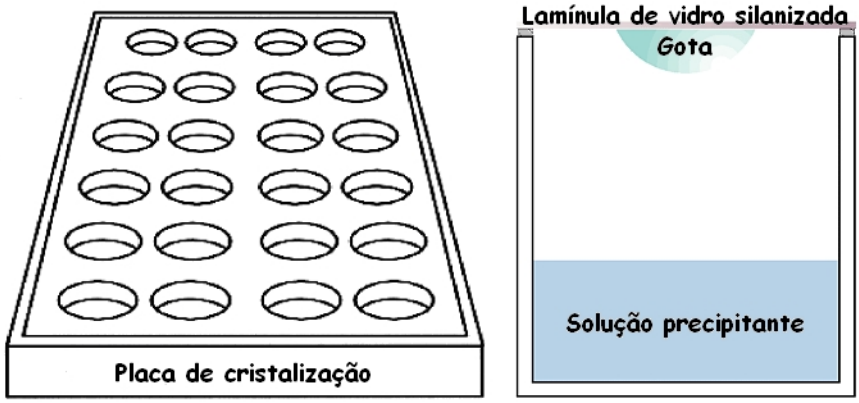


Figura 2.7: Esquema de uma placa com 24 poços utilizada no refinamento das condições de cristalização para XfSurE. A figura também mostra os detalhes de um dos 24 poços usados com o método da gota suspensa.

Coleta de dados e modelos iniciais

Experimentos de difração de raios X foram realizados nas linhas D03B-MX1 e W01B-MX2 do LNLS usando um detector MARCCD 225 pelo método de rotação do cristal. Para minimizar os danos causados pela exposição à radiação [58], um soprador de nitrogênio a 100 K (Oxford Cryosystems) sobre os cristais foi utilizado durante a coleta. Para evitar a formação de gelo cristalino devido à baixa temperatura, os cristais foram embebidos em vários tipos de soluções crioprotetoras (glicerol, etileno glicol, PEG, MPD, etc.) antes de serem levados ao feixe de raios X.

A redução e análise dos dados de coleta foram realizadas com o auxílio dos programas MOSFLM e SCALA da suíte CCP4 [55].

No processo de coleta das intensidades refletidas, a informação da fase dos fatores de estrutura é perdida e várias técnicas experimentais e computacionais são empregadas para reconstruir tridimensionalmente o agente espalhador. A técnica de substituição molecular é largamente empregada devido à praticidade que as estruturas tridimensionais já depositadas no PDB [7] oferecem por poderem ser pontos de partida para a proteína de interesse, sem a necessidade da realização de experimentos mais sofisticados para obtenção das fases iniciais. Conforme discutido, a proteína SurE de *Xylella fastidiosa* possui ao menos cinco estruturas homólogas depositadas

(identidade sequencial $< 54\%$) favorecendo relativamente as chances de sucesso da substituição molecular conduzida pelo algoritmo implementado no pacote BALBES [59].

Refinamento das estruturas

A etapa subsequente à cristalização e coleta de dados de difração é o refinamento da estrutura inicial que consiste no uso de mapas de densidade eletrônica como guia para o ajuste do modelo estrutural inicial. Nesta etapa, correção da orientação de cadeias laterais, adição de moléculas de água com simetria cristalina e outros detalhes dessa natureza são efetivamente realizados para a construção de um modelo estrutural completo.

Entretanto, a minha participação no projeto de cristalografia da XfSurE foi interrompida nesta etapa devido ao início de um estágio de doutorado no exterior cujo trabalho será discutido mais adiante no Capítulo 4. Considerando o interesse de todos os colaboradores envolvidos no projeto, decidiu-se que os refinamentos das estruturas cristalográficas, a partir de sete conjuntos completos coletados, seriam finalizados pelo grupo do Prof. Dr. Jorge Iulek da Universidade Estadual de Ponta Grossa.

2.2.4 – Análise de modos normais de vibração

O modelo de rede elástica utilizado neste trabalho foi descrito por Tirion [41]. A partir de apenas um potencial do tipo “massa-mola”, Tirion conseguiu explicar os “fatores B” de Debye-Waller associados à vibração devido à energia térmica de proteínas dentro de um cristal. Neste modelo de rede elástica, a estrutura proteica é representada apenas pelas coordenadas dos N carbonos-α (C_α). Um simples potencial harmônico é usado para levar em conta todas as interações entre pares de resíduos dentro de uma distância de corte R_C (Figura 2.8).

Assim, a energia potencial na representação de rede elástica é dada por

$$E_{elástica} = \frac{1}{2} \sum_{d_{ij}^0 < R_C} k_{ij}(d_{ij} - d_{ij}^0)^2, \tag{2.7}$$

onde k_{ij} são constantes de mola associadas aos resíduos i e j as quais são usualmente consideradas resíduo-independente, isto é, $k_{ij} = k$ para todos os contatos, d_{ij} é a distância entre os C_α e d_{ij}^0 é a correspondente distância na estrutura cristalográfica. Nota-se a partir da Equação 2.7,

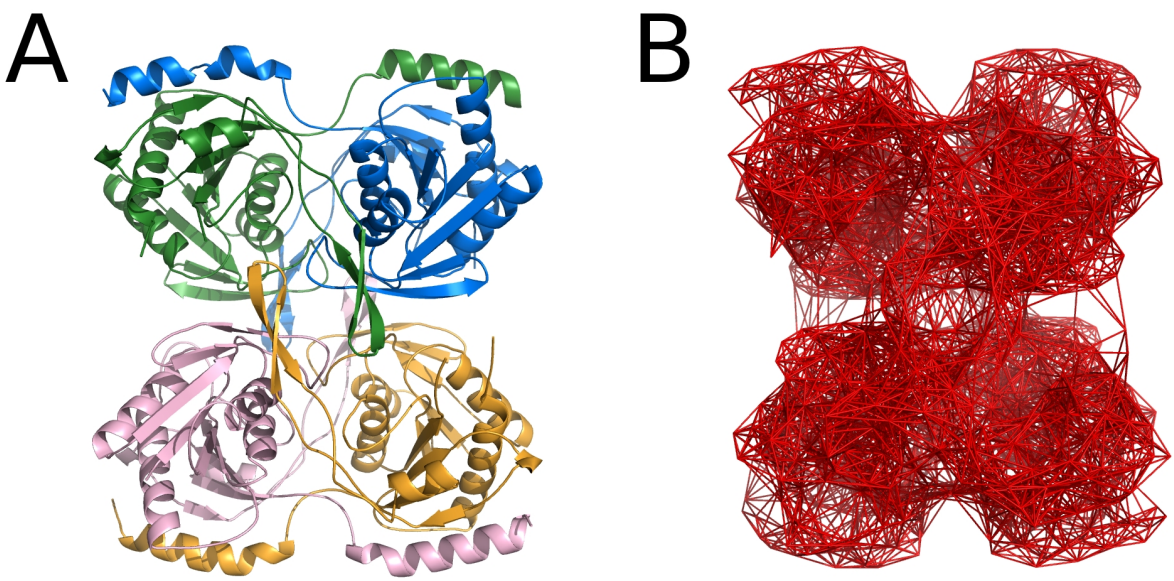


Figura 2.8: Construção de uma rede elástica para a XfSurE. (A) Estrutura cristalográfica. (B) Rede elástica com $R_C = 10\text{Å}$ a partir da qual é realizada a análise de modos normais de vibração.

que a configuração de mínima energia nesta formulação é a própria estrutura cristalográfica. Para realizar AMN a partir do modelo da Equação 2.7, não é necessário portanto, relaxar previamente a estrutura cristalográfica por outros métodos computacionais de minimização de energia.

Uma expansão até segunda ordem em torno do mínimo da Equação 2.7 leva a:

$$E_{elástica} \approx \frac{1}{2}\delta\mathbf{R}^T\mathbf{H}\delta\mathbf{R} = \frac{1}{2}\sum_{d_{ij}^0<R_C} k\delta\mathbf{R}^T H_{ij}\delta\mathbf{R}, \tag{2.8}$$

onde $\delta\mathbf{R} = \mathbf{R} - \mathbf{R_0}$ é um vetor de dimensão 3N, onde $\mathbf{R_0} = \{\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_N\}$ representa o conjunto das coordenadas cartesianas dos C_α da estrutura cristalográfica. H_{ij} são os elementos da matriz Hessiana $\mathbf{H}$, onde $H_{ij} = \frac{1}{2}\nabla^2(d_{ij} - d_{ij}^0)^2$.

Os autovetores da matriz Hessiana são chamados de modos normais. Os primeiros 6 modos estão relacionados aos 3 graus de liberdade de translação e aos 3 de rotação. Assim, o primeiro modo normal que promove mudança das distâncias relativas entre os C_α, é o modo 7. Disto decorre que uma vibração qualquer da estrutura de uma proteína pode ser descrita por uma combinação linear de todos os modos normais que formam uma base vetorial.

A amplitude de movimento no modelo de modos normais é arbitrária, embora todos os resí-

duos oscilem na mesma frequência em um dado modo. Entretanto, sabe-se [60] que as grandes amplitudes vibracionais estão relacionadas às mais baixas frequências e portanto, aos primeiros modos normais.

Projeção das diferenças estruturais nos modos normais

Estruturas diferentes de uma mesma proteína podem ser estudadas por AMN. A ideia consiste em medir o grau de correlação que as diferenças estruturais observadas possuem na direção de deslocamento promovida por um dado modo normal.

Para isso, define-se a métrica Projeção (P_j) em um modo normal j por

$$P_j = \frac{\left|\sum_i^{3N} a_{ij}\Delta r_i\right|}{\left(\sum_i^{3N} a_{ij}^2 \sum_i^{3N} \Delta r_i^2\right)^{1/2}}, \tag{2.9}$$

onde a_{ij} são as coordenadas do modo normal j associados ao aminoácido i, cujas coordenadas cartesianas são x, y e z. $\{\Delta\mathbf{r}_i\}$ é o conjunto de vetores que ligam os C_α correspondentes nas duas estruturas cristalográficas conhecidas.

Logo, se os vetores que representam a diferença estrutural entre dois estados conhecidos coincidem com as mesmas direções de um determinado modo normal, $P_j = 1$. Neste caso, diz-se que este modo normal isolado descreve completamente a transição estrutural observada.

Programas utilizados

Os códigos em FORTRAN que constroem a matriz Hessiana do sistema oscilante acoplado e sua respectiva diagonalização são disponibilizados no *website* do servidor ELNEMO [61]. A partir desses programas obtivemos os modos normais para uma proteína e fizemos todas as análises estruturais através de *scripts* escritos em linguagem Perl que nós implementamos.

2.3 – Resultados e discussão

2.3.1 – Caracterização estrutural de XfSurE por SAXS

Medidas de SAXS foram realizadas em várias concentrações e mostraram que o estado oligomérico de XfSurE não depende da concentração da proteína. Inicialmente, será discutido apenas os dados para a amostra com uma concentração de 12,1 mg/mL cuja curva experimental é mostrada na Figura 2.9-A. O comportamento linear[5] na região de Guinier (Figura 2.9-D) indica a presença de um sistema monodisperso em solução, isto é, apenas uma forma oligomérica está presente afastando a possibilidade da presença indesejada de mistura de oligômeros ou agregação não-específica. Pela linearização da lei de Guinier (Equação 2.2), obtivemos o raio de giro de $32,7 \pm 0,2$ Å e a intensidade extrapolada para ângulo zero, $I_{XfSurE}(0) = 1,9564 \cdot 10^{-3}$ [u. a.].

O gráfico de Kratky (Figura 2.9-C) apresenta um máximo bem definido, sendo um forte indicativo qualitativo que a proteína está bem enovelada em solução.

A função de distribuição de distâncias $p(r)$ (Figura 2.9-B) indica uma máxima distância intramolecular de 100 Å. Pela característica do perfil da curva $p(r)$, qualitativamente podemos dizer que a forma de XfSurE aproxima-se de um elipsoide devido ao pico estar um pouco afastado para esquerda em relação ao centro, que é a posição assumida para uma esfera perfeita.

Para estimar a massa de XfSurE e consequentemente o seu estado oligomérico, lisozima ($M_{Liso} = 14,4$ kDa) foi utilizada como proteína padrão, na concentração de 27 mg/mL. A partir da análise de Guinier para a proteína padrão, obtivemos $I_{Liso}(0) = 5,8127 \cdot 10^{-4}$ [u. a.], o que nos leva a concluir pela Equação 2.5 que o valor da massa de XfSurE estimada é de $\approx 108 \pm 10\%$ kDa. Uma vez que a massa do monômero de XfSurE é de $28,3$ kDa, o número de subunidades que estão agrupadas é correspondente a $n = 108/28,3 \approx 4$, isto é, a proteína é tetramérica em solução. Este resultado foi posteriormente corroborado pela técnica de filtração em gel, cujo valor para a massa da XfSurE foi estimado em $\approx 117 \pm 1$ kDa [16].

Os principais parâmetros de XfSurE obtidos por SAXS estão resumidos na Tabela 2.1.

[5] É importante ressaltar que monodispersidade e idealidade são condições tais que garantem a linearidade na região de Guinier embora a recíproca nem sempre seja verdadeira.

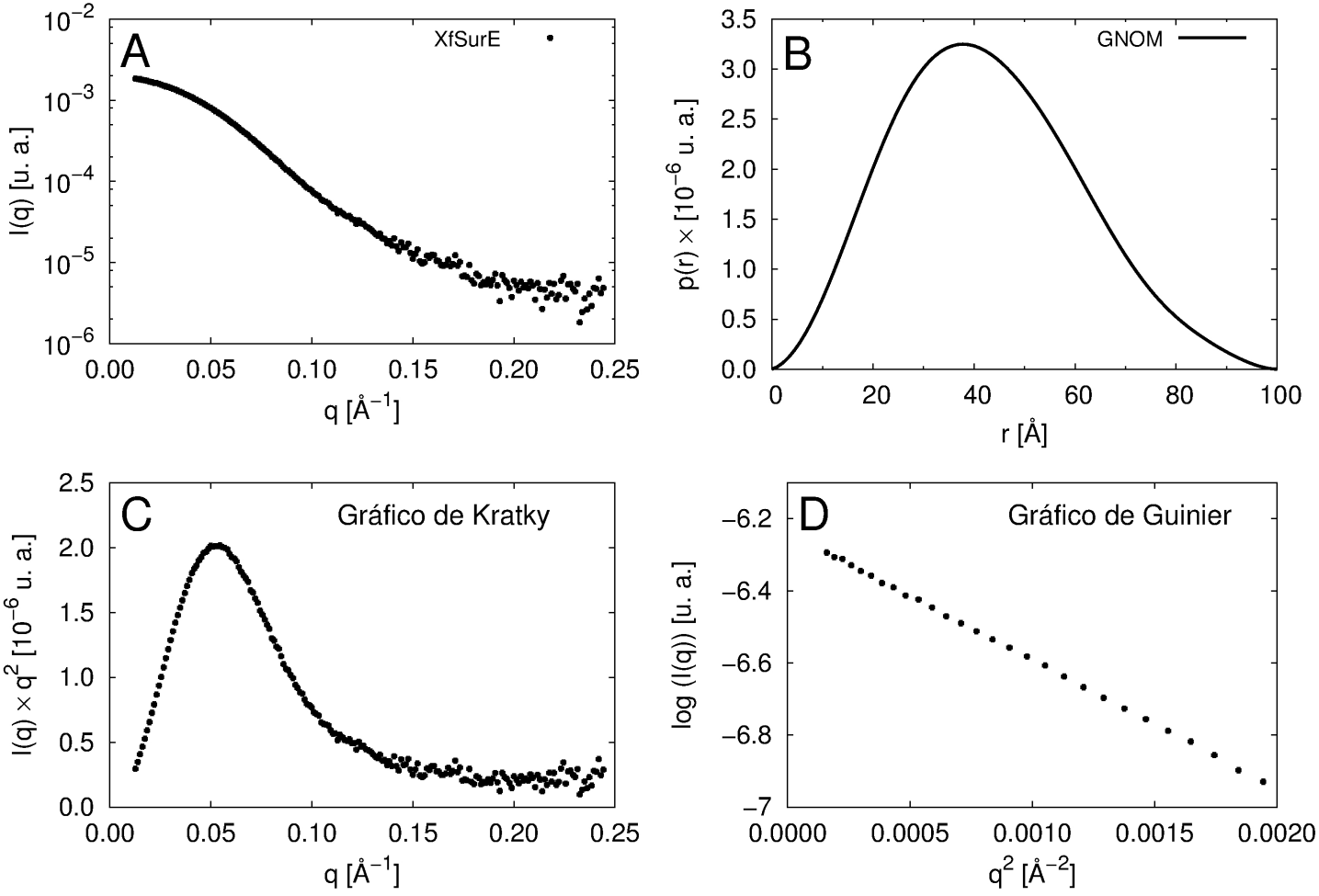


Figura 2.9: Dados de SAXS da XfSurE na condição nativa. (A) Intensidade de raios X em função do módulo do vetor de espalhamento - $I(q)$. (B) Transformada de Fourier indireta da curva $I(q)$ obtida por GNOM: distribuição de distâncias $p(r)$. (C) Gráfico de Kratky da curva $I(q)$. (D) Gráfico de Guinier para o início da curva $I(q)$.

Tabela 2.1: Parâmetros de XfSurE obtidos por SAXS.

R_G (Å) por Guinier[a]	R_G (Å) pela $p(r)$[b]	D_{max} (Å)	M (kDa) por SAXS	M (kDa) teórica[c]
32,7 $\pm$ 0,2	32,67 $\pm$ 0,04	100	108	28,3

[a] Raio de giro calculado pela linearização do início da curva $I(q)$; o erro é relativo ao ajuste linear.

[b] Raio de giro calculado indiretamente por GNOM.

[c] Massa de um monômero de XfSurE composto por 263 aminoácidos.

Determinação da estrutura de baixa resolução

Após a análise preliminar das curvas de SAXS, as quais fornecem parâmetros globais acerca da estrutura da proteína, o envelope molecular foi modelado pelos métodos descritos na Seção 2.2.2 com a intenção de se determinar o arranjo entre as cadeias da estrutura quaternária tetramérica.

Antes porém, levou-se em conta a forma do monômero característico das proteínas da família SurE e os possíveis oligômeros que porventura pudessem ser formados. Suspeitou-se que o tetrâmero de XfSurE pudesse ser reconstruído por duas operações de simetria independentes. Como um exemplo, considere o tetrâmero cristalográfico da SurE de *Salmonella typhimurium*, com 54% de identidade sequencial com XfSurE, formado pelas cadeias A, B, C e D (Figura 2.3, à direita). A primeira operação de simetria possível para reconstruir o tetrâmero é uma operação P2 sobre o dímero formado pelas cadeias A e B. Esta operação consiste na rotação de 180^0 do dímero em torno do eixo perpendicular à página que atravessa o centro do tetrâmero. Com isto, promove-se a sobreposição do dímero formado pelas cadeias A e B sobre o dímero formado pelas cadeias D e C, respectivamente, isto é, a cadeia A é projetada sobre D e a cadeia B é projetada sobre C. Em uma análise mais cuidadosa, percebe-se que o tetrâmero também pode ser reconstruído a partir de operações de simetria P222 aplicada a apenas um monômero. Esta operação requer três rotações de 180^0 em relação a três eixos perpendiculares, cuja origem se localiza no centro do tetrâmero. Assim, a cadeia A pode ser sobreposta à cadeia B por uma rotação em torno do eixo vertical no plano da página, à cadeia C em relação ao eixo horizontal e à cadeia D em relação ao eixo perpendicular à página.

2.3 Resultados e discussão

Na Figura 2.10, três modelos para o envelope da XfSurE obtidos com DAMMIN são apresentados. Na construção destes modelos, também exploramos a possibilidade do envelope da XfSurE possuir simetrias, uma vez que simetria e função proteica estão intimamente relacionadas [62]. É importante enfatizar que o algoritmo utilizando por DAMMIN se vale apenas da informação experimental da curva de SAXS como um vínculo, e portanto, as operações de simetria impostas pelo programa não são baseadas nas informações cristalográficas das homólogas de XfSurE. É por isso que este tipo de modelagem é conhecida como *ab initio* na área de SAXS.

Como esperado, todos os envelopes da Figura 2.10 apresentam grosseiramente a forma de um elipsoide como foi previsto pela forma da $p(r)$ de XfSurE. Além disso, todos possuem a máxima distância intramolecular $D_{max} = 100\text{Å}$ que está associada ao comprimento longitudinal dos envelopes. Os envelopes em que as simetrias P2 e P222 (Figura 2.10-B e Figura 2.10-C) foram impostas possuem os melhores acordos com dados de SAXS, com $\chi = 1.89$ e $\chi = 1.65$, respectivamente. O envelope obtido sem imposição de simetria embora qualitativamente semelhante aos demais, fornece um ajuste de pior qualidade ($\chi = 2.89$) para curva $I(q)$ experimental.

Esta caracterização em baixa resolução foi suficiente para distinguir o envelope P222 (Figura 2.10-C) como o melhor, com o respectivo valor do ajuste χ cerca de 13% menor em relação ao χ do segundo melhor envelope (P2) e cerca de 43% em relação ao envelope sem simetria (P1). Diante destes resultados e da análise de simetria cristalográfica dos tetrâmeros de SurE, o modelo estrutural P222 (Figura 2.10-C) foi selecionado para as análises subsequentes.

Modelagem de corpo rígido

No período onde os as primeiras análises foram realizadas, não dispúnhamos da estrutura cristalográfica de XfSurE, e por esta razão a discussão a seguir baseia-se na estrutura da SurE de *Salmonella typhimurium* (StSurE) com 54% de identidade sequencial.

Uma vez escolhido o melhor modelo estrutural de baixa resolução para a XfSurE, realizamos uma sobreposição do tetrâmetro cristalográfico de StSurE no envelope modelado por SAXS usando o programa SUPCOMB [63]. A Figura 2.11-A mostra o excelente acordo entre as duas estruturas. Particularmente, o acordo entre o envelope de XfSurE e a estrutura homóloga StSurE evidencia as características da simetria P222 presente na estrutura quaternária de XfSurE e confirmam, de maneira independente, que os tetrâmeros cristalográficos propostos para as diferentes

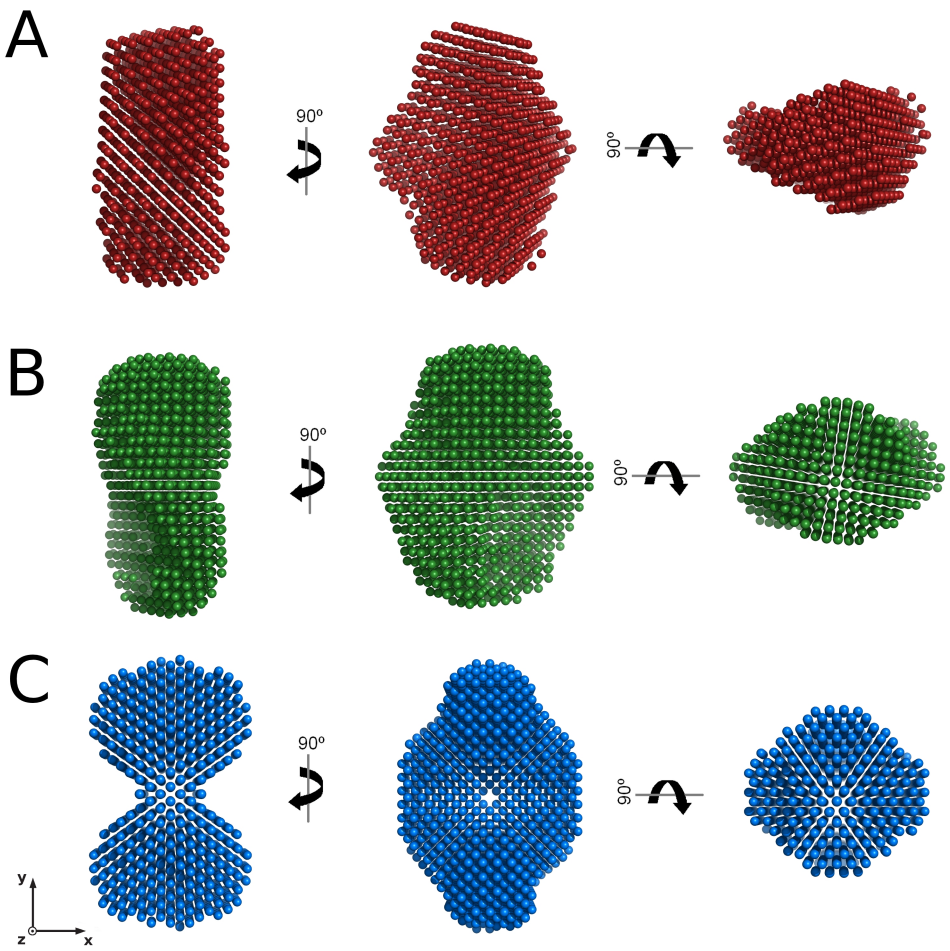


Figura 2.10: Representação da estrutura de baixa resolução da proteína XfSurE produzida pelo programa DAMMIN pela distribuição de átomos fictícios que ajusta os dados de SAXS. (A) Sem imposição de simetria molecular (P1), com ajuste $\chi = 2.89$. (B) Com imposição da simetria P2 e (C) P222, com ajustes $\chi = 1.89$ e $\chi = 1.65$, respectivamente. As estruturas ao centro e à direita estão rotacionadas de 90° em relação ao eixo y e ao x sucessivamente, conforme indica a figura.

proteínas SurE de outros organismos têm sentido biológico para a proteína SurE de *Xylella fastidiosa*. A partir da estrutura tetramérica da SurE de *S. typhimurium*, confrontamos o respectivo perfil teórico da $I(q)$ com a curva experimental de XfSurE. A qualidade do ajuste foi de $\chi = 2.04$, indicando que ao menos a disposição espacial das subunidades do tetrâmero está correta. Portanto, do nosso conhecimento, essa foi a primeira caracterização estrutural em solução de uma proteína SurE.

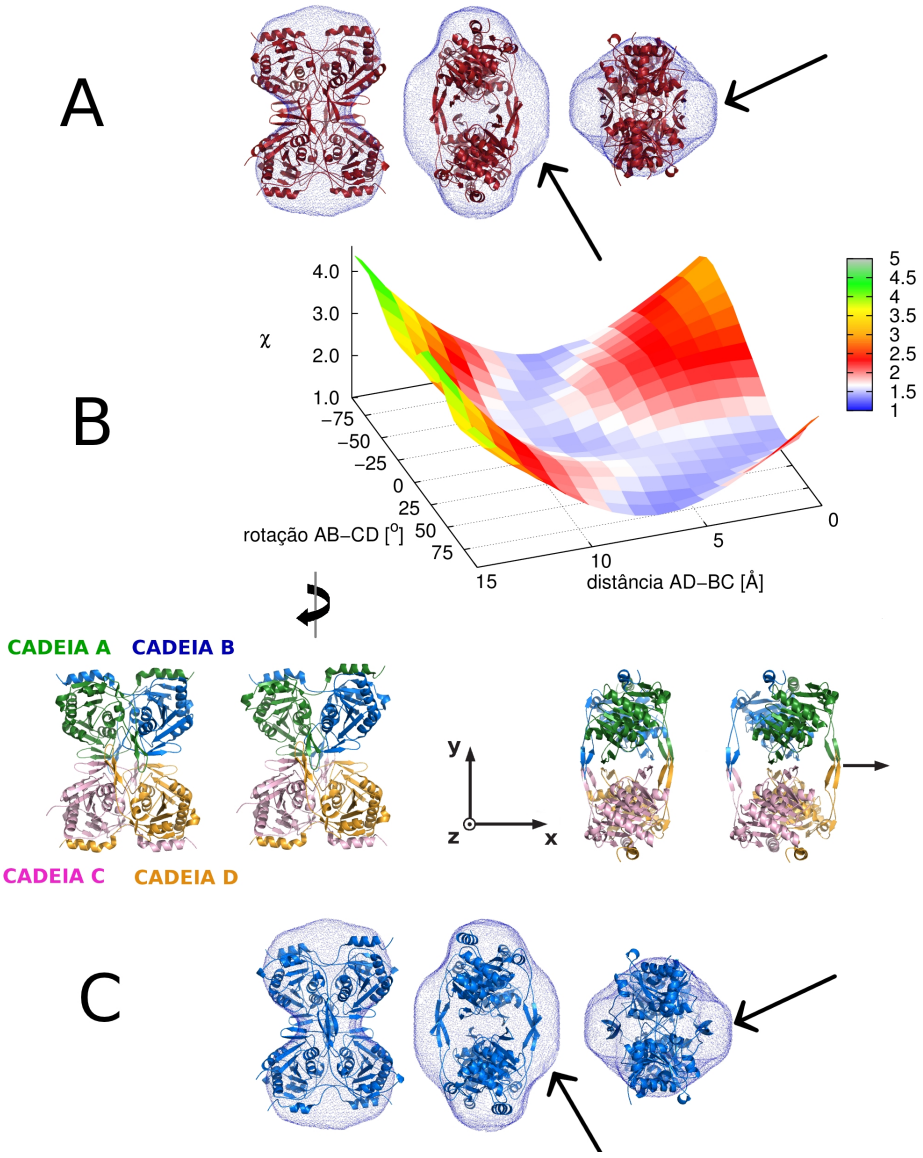


Figura 2.11: Modelagem de corpo rígido a partir de uma homóloga de XfSurE. (A) Envelope de baixa resolução de XfSurE sobreposto ao tetrâmero cristalográfico da SurE de *Salmonella typhimurium* (StSurE, 54% de identidade). As setas indicam volumes "vazios" em regiões do envelope que não são ocupadas pela estrutura de StSurE. (B) Modelagem de corpo rígido a partir de movimentos de rotação em torno do eixo y do dímero superior e de translação das subunidades diagonais do tetrâmero de StSurE na direção x. O ajuste da curva teórica de SAXS contra a curva $I(q)$ experimental de XfSurE é mostrado para cada modelo. (C) Melhor modelo obtido sobreposto ao envelope de XfSurE. As setas indicam o preenchimento do "vazio" pela estrutura modelada.

2.3 Resultados e discussão

Examinando minuciosamente a Figura 2.11-A, reparamos que as regiões indicadas pelas setas apresentam “vazios” correspondentes às regiões onde as alças de tetramerização se encontram, duas a duas. De alguma forma o envelope derivado a partir dos dados de XfSurE sinaliza que a estrutura procurada deve apresentar diferenças nessas regiões em relação a estrutura de StSurE. Esta é uma suposição razoável, já que StSurE tem apenas cerca de 54% de identidade sequencial com XfSurE [16].

Modelagem de corpo rígido foi a estratégia computacional inicial escolhida para que conformações alternativas pudessem ser criadas de forma a melhorar o ajuste aos dados de SAXS. A maneira mais direita, seria produzir um tetrâmero que tenha as alças de tetramerização mais afastadas, de forma a preencher os “vazios” percebidos na Figura 2.11-A. A Figura 2.11-B mostra um movimento de corpo rígido onde isso é possível. O afastamento translacional das cadeias diagonais A e D em relação à diagonal formada pelas subunidades B e C, promove diretamente o efeito desejado do distanciamento das alças de tetramerização. Por outro lado, este movimento distorce parcialmente a simetria P222 do envelope. De forma a preservar, ao menos aproximadamente, a simetria global, outro movimento de corpo rígido mostra-se conveniente. Esse segundo movimento cria a rotação do dímero AB em relação ao dímero CD.

A Figura 2.11-B mostra a combinação dos movimentos de corpo rígido entre as subunidades e o respectivo ajuste χ entre a $I(q)$ teórica de uma conformação específica e a curva experimental $I(q)$ para XfSurE. A estrutura original de StSurE corresponde à origem do sistema de coordenadas. A superfície χ em função dos movimentos combinados revela que diversos arranjos entre as subunidades são possíveis, todos com $\chi \approx 1.0$. Para nossa surpresa, com exceção da região onde a distância entre as subunidades diagonais AD-BC é menor 5Å, há uma região de distâncias no intervalo [5Å-10Å] onde uma população de conformações com grandes torções entre os dímeros AB-CD e com inesperado acordo da ordem de $\chi \approx 1.0$ são encontradas. A princípio, configurações com grandes torções entre os dímeros AB-CD, da ordem de 75 graus, podem ser meramente falsos positivos, isto é, configurações que apresentam um baixo valor de χ, mas não correspondem à estrutura procurada. SAXS é uma técnica de baixa resolução e não é impossível que duas estruturas que são grosseiramente parecidas exibam $I(q)$ e $p(r)$ muito próximas[6]. Além dessa possibilidade, grandes torções desta natureza desestabilizariam o tetrâmero por afastar se-

[6]Os Capítulos 3 e 4 tratam desse assunto mais adequadamente abordando o problema do reconhecimento de formas tridimensionais a partir dos perfis de SAXS.

veramente as alças de tetramerização e por isso a superfície de χ apenas faria sentido biológico na região central da Figura 2.11-B e na região próxima à origem.

A Figura 2.11-C mostra um exemplo de conformação onde o distanciamento das diagonais é de 7.5Å e a rotação entre dímeros é de 25^o em relação à posição original dessas subunidade na estrutura de StSurE. Como pode ser notado, a estrutura modelada se ajusta melhor ($\chi = 1.502$) no envelope de SAXS que a estrutura original ($\chi = 2.04$).

Novas perguntas sobre o comportamento da enzima XfSurE

Já era de conhecimento que o sítio ativo das SurEs estaria localizado em uma região do monômero que se encontra no centro do tetrâmero, uma região diretamente afetada por mudanças na área de interface entre os dímeros AB-CD. Curiosamente, a modelagem de corpo rígido, embora uma abordagem simples mas viável, forneceu resultados de bons ajustes aos dados de SAXS a partir de conformações "torcidas" justamente na interface dos dímeros AB-CD. Alterações na estrutura quaternária é uma característica muito presente em enzimas alostéricas que usam deste recurso para regular a entrada do substrato no sítio catalítico. Portanto, estes resultados possibilitaram levantar uma questão que ainda não havia sido feita: Seria a enzima XfSurE alostérica?

Posteriormente, novos estudos de cinética enzimática foram conduzidos por colaboradores especialmente com a intenção de explorar essa possibilidade. Interessantemente, os resultados revelaram novas propriedades catalíticas de XfSurE na presença de substratos naturais, especialmente na presença de 3'-AMP pelo qual XfSurE possui maior afinidade. Foi claramente demonstrado [16] que na presença de 3'-AMP, XfSurE exibe um comportamento alostérico com comportamento cooperativo positivo, cujo coeficiente de Hill[7] é igual a $2,67 \pm 0,09$.

A princípio, as diversas estruturas modeladas que se ajustam bem aos dados de SAXS $\chi \approx 1,0$ representariam diferentes estados conformacionais que juntos, contribuiriam para a conformação média assumida por XfSurE em solução, já que esta é o único tipo de medida acessível pela técnica experimental utilizada. Em caráter puramente especulativo e do ponto de vista do modelo clássico de Monod-Wyman-Changeux [33] para proteínas alostéricas, o mecanismo alostérico de

[7] O coeficiente de Hill quantifica o comportamento alostérico cooperativo de uma enzima. No caso da XfSurE, como há quatro sítios possíveis de ligação, o valor máximo para este coeficiente é 4 — no caso de uma ligação do tipo cooperativa positiva, onde a presença de um ligante favorece a ligação de outra molécula de substrato.

XfSurE incluiria apenas duas configurações: a relaxada (R) e a tensionada (T), estando a primeira relacionada com a estrutura semelhante à StSurE e a segunda com a estrutura que apresenta a torção entre os dímeros AB-CD. Assim, mudanças na área de contato entre estes dímeros poderiam eventualmente controlar o acesso do substrato nos sítios ativos do tetrâmero.

Estariam esses movimentos estruturais hipotéticos realmente relacionados de algum modo à cooperatividade e à regulação dos substratos? Quais dados adicionais são necessários? As seções seguintes apresentam esforços na busca de possíveis respostas para essas perguntas.

Novas medidas de XfSurE na presença de ligantes

É concebível que possíveis conformações tetraméricas discutidas na seção anterior possam estar relacionadas aos graus de liberdade da proteína nativa em solução, o que constitui em si uma hipótese para o mecanismo de regulação de entrada e saída do substrato.

Medidas de SAXS adicionais foram realizadas para a proteína XfSurE na presença de ligantes e substrato com o objetivo de observar indícios de mudança da estrutura quaternária de XfSuE com o intuito de correlacioná-las aos estudos estruturais.

Experimentos ideais para este tipo de estudo requerem uma instrumentação de SAXS adequada, com dispositivos do tipo *stopped-flow* que permitem o controle dos componentes da mistura proteína + ligante dentro do porta-amostra assim como uma instrumentação de dados resolvidos no tempo. Na linha SAXS2 do LNLS, o que pôde ser feito à época foi uma medida da ordem de minutos de exposição e uma mistura preparada antecipadamente para ser injetada diretamente no porta-amostra. Em outras palavras, estas medidas realizadas perdem qualquer relação com a cinética enzimática devido às escalas de tempo envolvidas diferirem em várias ordens de grandeza. Assim, se mudanças conformacionais são observadas, elas devem ser permanentes ou resultado da média espacial e temporal das conformações em solução.

Para desempenhar sua função, a enzima XfSurE necessita de um cofator metálico (Mn^{2+}) para ocorrer a catálise, clivando o fosfato da molécula 3'-AMP. Portanto, na ausência de resolução temporal das medidas, decidimos coletar curvas com 3'-AMP e Mn^{2+} separados, já que rapidamente o substrato seria consumido pela reação enzimática, caso estivessem juntos em solução[8].

[8]Ainda que o cuidado na preparação da mistura proteína+substrato na ausência do cofator tenha sido criterioso, quantidades residuais de Mn^{2+} poderiam ser suficientes para desencadear a catálise. Uma

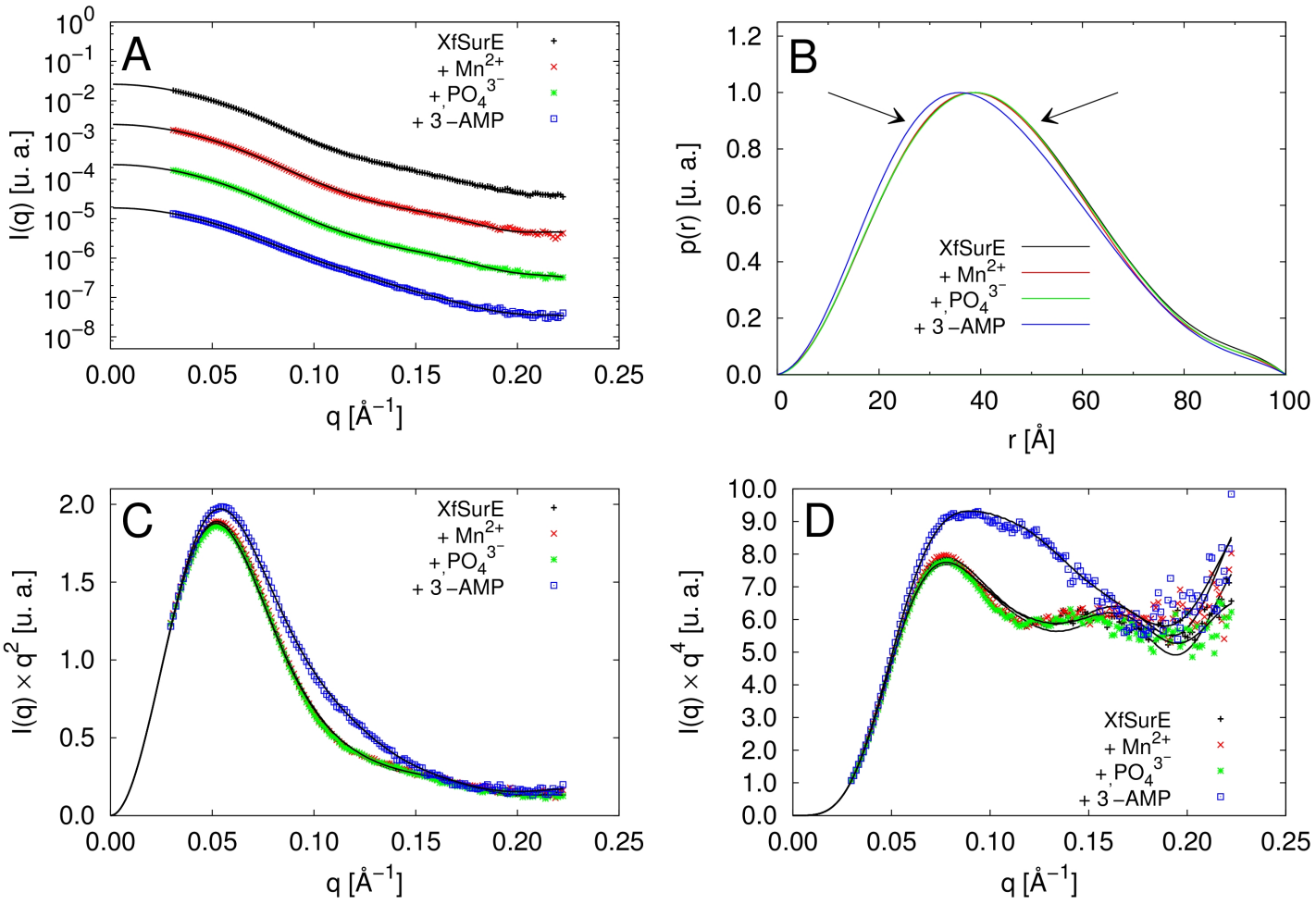


Figura 2.12: Curvas experimentais de SAXS para XfSurE na forma nativa e com ligantes: XfSurE + 0,5 mM Mn^{2+}, XfSurE + 1 mM PO_4^{3-} e XfSurE + 100 mM 3'-AMP. (A) Curvas $I(q)$ e (B) respectivas $p(r)$, cujas setas destacam os desvios entre as distribuições de distâncias de XfSurE na forma nativa e com o substrato 3'-AMP. Os gráficos de (C) Kratky e (D) Porod acentuam as diferenças observadas nas curvas $I(q)$ da proteína na presença do substrato. As linhas sólidas são os ajustes teóricos obtidos pelo método da transformada de Fourier indireta da curva $p(r)$ obtidos por GNOM.

2.3 Resultados e discussão

A Figura 2.12-A mostra as novas medidas realizadas da curva de espalhamento $I(q)$ para XfSurE na forma nativa, XfSuE na presença de seu cofator metálico Mn^{2+}, XfSuE na presença de um dos produtos da reação de catálise, o PO_4^{3-} e XfSurE na presença de seu substrato de maior afinidade, o 3'-AMP. Não é evidente, na escala da figura, as diferenças significativas entre as curvas $I(q)$. No entanto, as diferenças dos dados, observados pela $p(r)$ (Figura 2.12-B), mesmo que sutis, tornam-se mais claras sugerindo que um tetrâmero semelhante ao determinado anteriormente persiste na presença do substrato. Pelos gráficos de Kratky (Figura 2.12-C) e Porod (Figura 2.12-D), observa-se que não há diferenças significativas entre as curvas da proteína nativa e da proteína na presença dos íons. Já nesses mesmos gráficos, a curva de XfSurE na presença do substrato 3'-AMP definitivamente é diferenciada das demais.

O gráfico de Porod para XfSurE nativa e com íons apresenta um comportamento assintótico em torno de um valor constante $I(q)\times q^4 \approx 6.0$ em unidades arbitrárias, que é o comportamento esperado para uma proteína globular de densidade eletrônica homogênea e interface com o solvente bem definida. A curva para XfSurE + 3'-AMP aparentemente converge para o mesmo valor, inclusive com comportamento semelhante ao da amostra nativa na região $0,17\text{Å}^{-1} < q < 0,22\text{Å}^{-1}$, embora não se possa afirmar o comportamento destas curvas para $q >> 0,20\text{Å}^{-1}$, dificultando, neste caso, uma análise mais acurada. Por exemplo, se a curva apresentasse um comportamento linear crescente para $q \to \infty$, poderíamos dizer que existem flutuações da densidade eletrônica. Uma explicação plausível para o comportamento observado na Figura 2.12-D, de acordo com a Equação 2.4 de Porod, é considerar que o contraste da densidade eletrônica entre a proteína e o solvente ($\Delta\rho$) é mais suave para a amostra com 3'-AMP que o contraste das amostras com íons, devido ao decaimento da curva ser menos acentuado para o primeiro caso.

Análises adicionais indicam que não há diferenças significativas entre os raios de giro das amostras coletadas e para as amostras com ligantes. Os envelopes obtidos com DAMMIN também são semelhantes aos envelopes da proteína nativa. Diante disto, proceder com a modelagem de corpo rígido como a realizada para a proteína nativa (Figura 2.11), pode resultar em análises equivocadas. As diferenças entre as $p(r)$ (Figura 2.12-B) são sutis e a escolha de um modelo

maneira inequívoca para garantir a ausência da reação seria a utilização de: quelantes para eliminar a ação do íon metálico; análogos de AMP, evitando a especificidade da reação; mutagênese, para que a complexação no sítio ativo não seja realizada. Independente se há ou não a catálise, os resultados mostram inquestionáveis alterações conformacionais.

estrutural em uma superfície com vários mínimos locais é certamente uma tarefa delicada.

Outros estudos de proteínas alostéricas por SAXS [64, 65, 66] também apresentam sutilezas nas diferenças encontradas nas curvas $I(q)$ e $p(r)$ de amostras com e sem ligante. Nestes casos, informações adicionais, experimentais ou teóricas, favorecem a interpretação do envelope de SAXS tornando-a mais confiável.

Análises mais significativas para o comportamento de XfSurE em solução juntamente com a análise de modos normais de vibração foram realizadas posteriormente (Seção 2.3.3) a partir das estruturas cristalográficas determinadas para XfSurE.

2.3.2 – Experimentos de cristalização e coleta de dados

Concomitantemente aos experimentos de SAXS, foram realizados diversos ensaios de cristalização. As amostras foram preparadas com a XfSurE na forma nativa e com ligantes, principalmente os cofatores metálicos e substratos artificiais, substratos naturais e inibidores, na tentativa de se obter cristais pela técnica de co-cristalização. Alternativamente, alguns cristais formados apenas com a enzima na forma nativa foram submetidos a um banho em uma solução preparada com algum dos ligantes (*soaking*), para que houvesse a oportunidade extra, falhando a co-cristalização, da obtenção de complexos.

Quando na presença de 3'-AMP, o cofator de maior afinidade com a enzina é o manganês, seguido de magnésio e cobalto. Estabelecer uma relação estequiométrica adequada entre um monômero de XfSurE e o íon manganês sempre se mostrou um desafio adicional aos ensaios de cristalização. Soluções da enzima na presença de soluções preparadas com $MnCl_2$ acima de 1,0 mM, geralmente precipitavam. Considerando uma solução típica de 10,0 mg/ml da XfSurE cujo monômero possui massa de 28,3 kDa, a relação estequiométrica equivalente é de aproximadamente 1 (uma) cadeia de XfSurE ($\approx$ 0,35 mM) para cada 3 (três) íons Mn^{2+} ($\approx$ 1,0 mM), tomando uma conotação de limite estequiométrico a ser atingido mas com certa margem para que se possa observar o íon aderido aos sítios da estrutura cristalográfica.

As fotos na Figura 2.13 mostram algumas gotas de cristalização em diferentes etapas do processo de otimização das condições de cristalização da enzima. A primeira foto à esquerda mostra uma gota que apresenta apenas um indício de cristalização com um cristal mal formado. As fotos sucessivas mostram o refinamento dessas condições até chegar em uma gota límpida onde

etapas adicionais de otimização foram realizadas para que melhores cristais possíveis fossem obtidos, como os dois apresentados no final da figura.

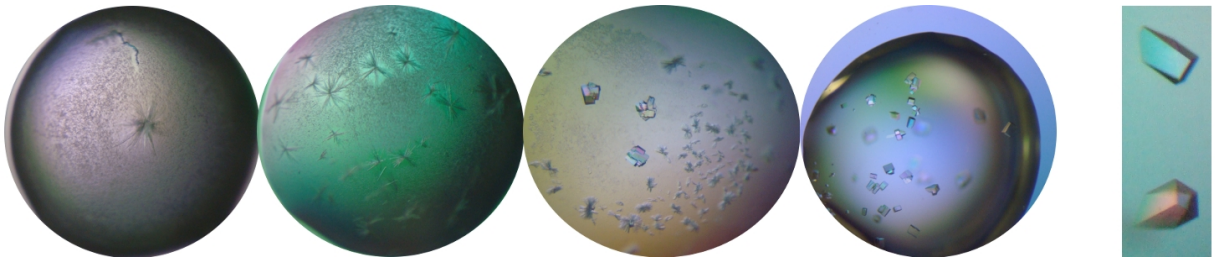


Figura 2.13: Etapas da otimização das condições de cristalização da XfSurE pelo método da gota suspensa. Típicos cristais de XfSurE com dimensões máximas de $140\times70\times70\mu$m são mostrados como resultado final do refinamento e prontos para serem submetidos aos experimentos de difração de raios X.

Em várias oportunidades, tentativas de co-cristalização de XfSurE junto a soluções crioprotetoras foram feitas sem muito sucesso. Além disso, alguns cristais apresentaram problemas com vários tipos de soluções crioprotetoras pela técnica *soaking* pois trincavam muito facilmente quando manuseados em contato com as mesmas. Contudo, apesar da fragilidade de alguns cristais, conseguimos coletar conjuntos de duas formas cristalinas, com um dímero e um tetrâmero na unidade assimétrica (ASU, *ASymmetric Unit*), reportando a variabilidade do conteúdo da ASU para a SurE de *Xylella fastidiosa* [17] e evidenciando o valor de estudos em solução por SAXS na determinação da unidade biológica funcional da enzima.

Ao todo, sete conjuntos de dados foram coletados. Em geral, os cristais que se mostraram mais robustos à manipulação, após os processamentos e análises, revelaram uma unidade assimétrica tetramérica enquanto que os mais frágeis, apresentaram um dímero na ASU. O padrão de difração típico para um cristal cuja ASU é tetramérica está mostrado na Figura 2.14-A. A tamanha fragilidade de alguns cristais em contato com soluções crioprotetoras fizeram com que alguns conjuntos fossem coletados na ausência de crioproteção, efeito caracterizado pelas imagens de difração com gelo cristalino na Figura 2.14-B produzido por um cristal com um dímero na ASU, embora os anéis de gelo não impossibilitaram o sucesso da determinação da estrutura de XfSurE.

O resumo das condições de cristalização dos melhores cristais é mostrado na Tabela 2.2. Os termos “maior” e “menor” para as unidades assimétricas diméricas (Dim.) ou tetramérica (Tetra.)

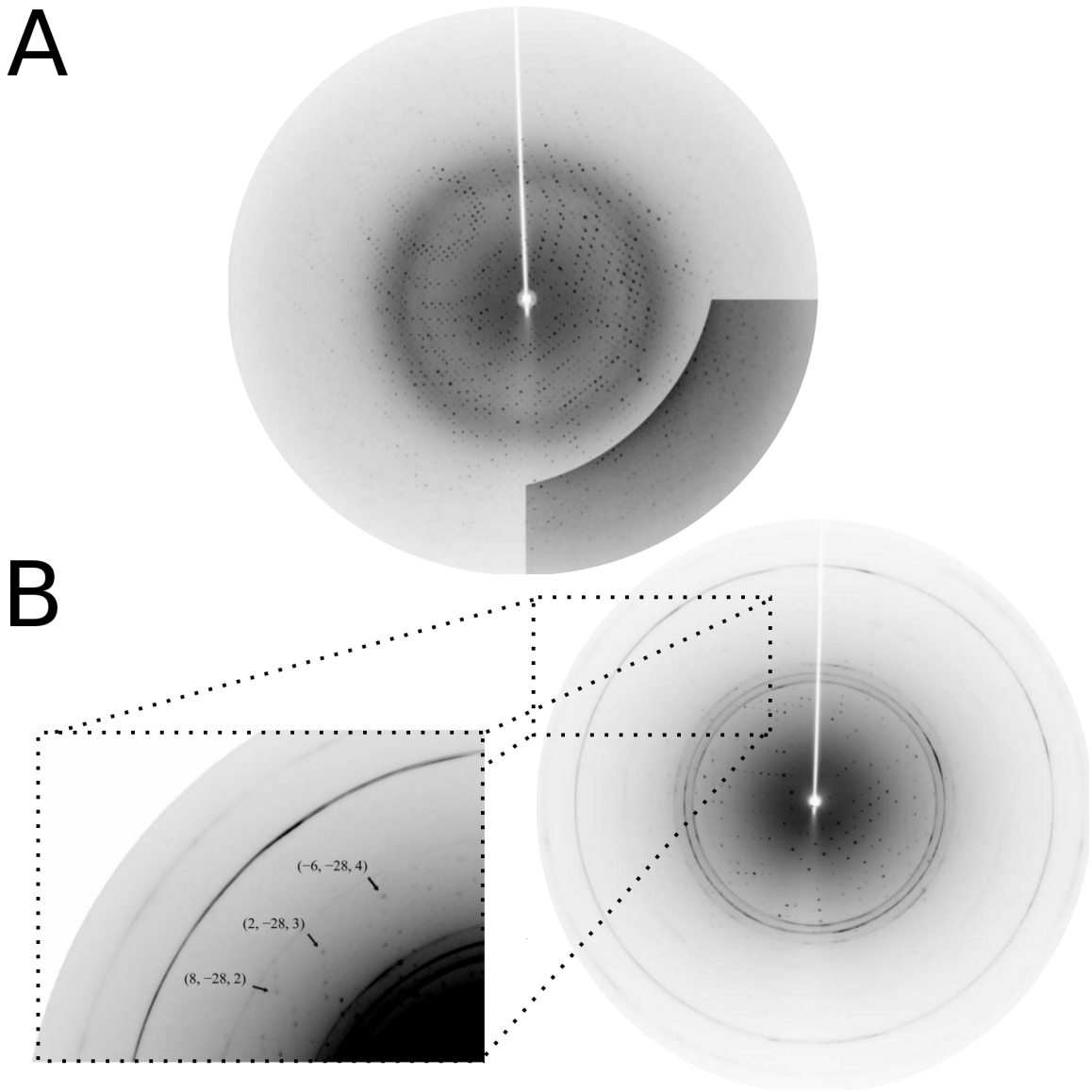


Figura 2.14: Padrão de difração dos cristais de XfSurE. (A) Padrão de difração com resolução de 1,93 Å de um cristal que possui um tetrâmero na unidade assimétrica. O contraste na borda da figura acentua as reflexões de mais alta resolução. (B) Padrão de difração com resolução de 2,90 Å representativo de um conjunto de imagens coletado para um cristal que possui um dímero na unidade assimétrica. Uma expansão da região de maior resolução é apresentada junto às reflexões *hkl*. Os anéis de gelo observados na figura não evitaram o sucesso da determinação da estrutura de XfSurE.

fazem referência ao tamanho do tetrâmetro cristalográfico obtido por operações de simetria em cada caso, como será especificado mais adiante.

Tabela 2.2: Condições de cristalização de quatro cristais de XfSurE.

	Conteúdo da unidade assimétrica			
	Dim. ASU menor	Dim. ASU maior	Tetra. ASU menor	Tetra. ASU maior
Condição de cristalização				
XfSurE (mg/ml)	7,0	6,5	6,5	6,1
PEG 3350 (%)	16	20	20	20
pH	7,6	7,6	7,6	7,5
BisTris (M)	0,1	0,1	0,1	0,1
NaI (mM)	130	140	140	140
DTT (mM)	5	-	-	5
co-cristalização	0,5 mM $MnCl_2$	2,5mM 3'-AMP e 0,1mM $MnCl_2$	-	0,1 mM $MnCl_2$
Coleta				
Crioproteção	-	glicerol 30%	glicerol 20%	glicerol 20%
Soaking	5mM $MnCl_2$	-	2,5mM 3'-AMP e 10 mM $MnCl_2$	-
Resolução	2,90	2,62	2,76	1,93

Pode ser visto que as condições são muito próximas para os quatro cristais reportados. O fator discriminante na explicação do surgimento de dímeros ou tetrâmeros na ASU poderia vir das diferentes estratégias de co-cristalização e/ou *soaking* empregadas no trato de cada cristal ou até mesmo vir de características inerentes de mobilidade da proteína que por nuances do ambiente cristalino, teve uma conformação selecionada em detrimento de outra. Detalhes desta ordem só podem ser esclarecidos com a análise das estruturas refinadas, onde seria possível verificar os contatos cristalinos entre as moléculas e possíveis alterações conformacionais entre elas.

Refinamento e estruturas de alta resolução de XfSurE

Nesta época, estávamos centrados no desenvolvimento de outro projeto da tese descrito no Capítulo 4. Dado o grande volume de dados a serem refinados e o interesse de todos os colaboradores, esta etapa foi realizada externamente pelo grupo do Prof. Jorge Iulek, que disponibilizou as estruturas cristalográficas finais a partir das quais realizamos as análises aqui descritas. Os estudos estruturais em alta resolução da proteína XfSurE culminaram em uma dissertação de mestrado. Por esta razão, as estatísticas do refinamento, os parâmetros de qualidade do modelo assim como outros detalhes adicionais sobre as estruturas obtidas podem ser encontradas na referência [67].

Os quatro cristais coletados nas condições descritas na Tabela 2.2 resultaram nas estruturas da Figura 2.15. Todos os quatro tetrâmeros foram sobrepostos estruturalmente uns aos outros pelo programa TM-score [68] e portanto, todas as figuras foram feitas na mesma escala, em todas as orientações.

À primeira vista, todos os tetrâmeros cristalográficos de XfSurE são muito parecidos com as estruturas apresentadas anteriormente na Figura 2.3 de proteínas homólogas. Especialmente, o monômero possui o mesmo enovelamento Rossmann característico entre todas as SurEs.

No entanto, as estruturas “Dim. ASU menor” (Figura 2.15-A) e “Dim. ASU maior” (Figura 2.15-B) são visualmente diferentes entre si, e os termos “menor” e “maior” tornam-se evidentes. O mesmo efeito de escala ocorre com as estruturas “Tetra. ASU menor” (Figura 2.15-C) e “Tetra. ASU maior” (Figura 2.15-D). Quantitativamente, o valor do RMSD entre as estruturas menores (“Dim. ASU menor” e “Tetra. ASU menor”) é de $1,661\text{Å}$, e entre os tetrâmeros maiores (“Dim. ASU maior” e “Tetra. ASU maior”) de $0,993\text{Å}$. Já para os tetrâmeros formados por dois dímeros (“Dim. ASU menor” e “Dim. ASU maior”) o RMSD é $2,175\text{Å}$, e entre os tetrâmeros da ASU (“Tetra. ASU menor” e “Tetra. ASU maior”) é de $2,430\text{Å}$.

O que chama mais atenção nestas pequenas distorções é que há indícios de um efeito global de mudança da estrutura quaternária por um fator de escala na direção longitudinal do tetrâmero, isto é, as estruturas “maiores” aparentam serem mais alongadas que as “menores” na direção do eixo y que em relação ao eixo x.

A Tabela 2.3 apresenta as áreas de contato na interface entre as quatro cadeias dos tetrâmeros cristalográficos de XfSurE. Há uma interação mais contundente entre as subunidades forma-

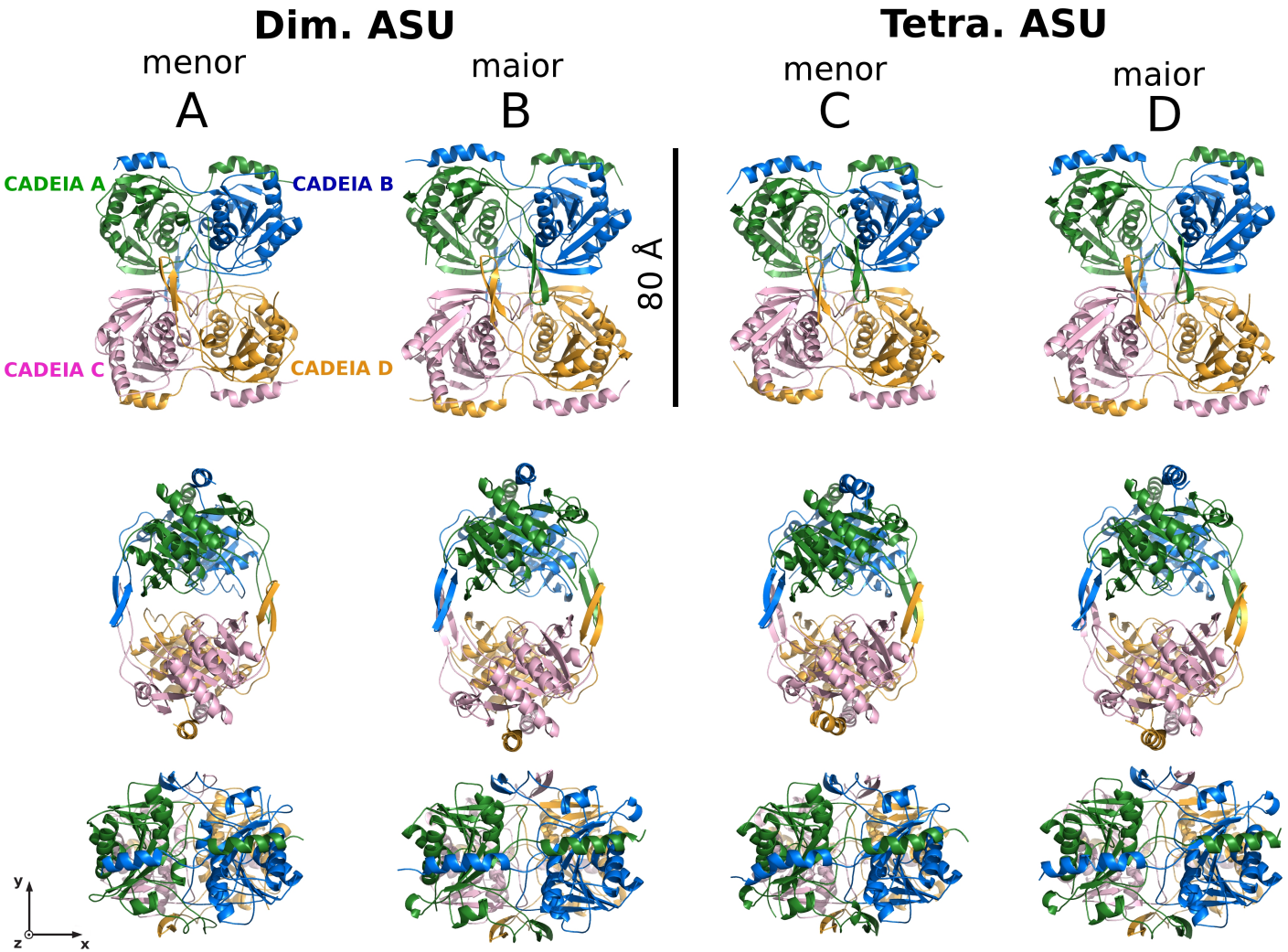


Figura 2.15: Estruturas tetraméricas cristalográficas da XfSurE obtidas a partir de quatro cristais distintos. Tetrâmeros formados por dímeros da unidade assimétrica, denominados: (A) “Dim. ASU menor” e (B) “Dim. ASU maior”. Tetrâmeros presentes nas unidades assimétricas: (C) “Tetra. ASU menor” e (D) “Tetra ASU maior”. As orientações moleculares das segunda e terceira linhas foram obtidas a partir das estruturas acima da figura por uma rotação de 90^0 em relação ao eixo y e ao eixo x, respectivamente.

das pelas cadeias A-B e pelas cadeias C-D. Como exemplo, a cadeia A da estrutura “Tetra. ASU maior” possui maior área de contato com a cadeia B ($3802,9\text{Å}^2$) seguida da cadeia D ($575,3\text{Å}^2$), associada à interação mediada pelas alças de tetramerização entre A-D. A menor área de contato da cadeia A ocorre com a cadeia C ($345,8\text{Å}^2$), o que é compreensível, já que a distância entre estas duas subunidades é maior. Este padrão de área de contato permeia todas as estruturas em todas as subunidades, ou seja, as subunidades lado a lado na horizontal (ao longo do eixo x) possuem maior área de contato, seguida pelas subunidades diagonais do tetrâmero e em menor contato estão as subunidades verticais vizinhas. Comparando as formas cristalinas, os tetrâmeros formados por dímeros na ASU apresentam uma área de contato cerca de 30$\%$ menor entre as subunidades A-B e C-D em relação aos tetrâmeros da ASU. A interação entre as alças de tetramerização (A-D e B-C) possuem área equivalentes nos em todas as estruturas.

Também é atestado que existem estruturas que são mais assimétricas que outras. O “Tetra. ASU menor” por exemplo, possui a cadeia A mais próxima de C (Área$_{AC} = 449,9\text{Å}^2$) em relação à proximidade da cadeia B com a cadeia D (Área$_{BD} = 387,5\text{Å}^2$). O mesmo ocorre para o “Dim. ASU menor” que possui Área$_{AC} = 282,9\text{Å}^2$ e Área$_{BD} = 539,2\text{Å}^2$. Nos outros dois casos, um arranjo mais simétrico dos monômeros é observado. Usando o critério da área de contato entre A-C e entre B-D para definir simetria e assimetria, as estruturas menores são denominadas assimétricas enquanto que as maiores, simétricas.

Tabela 2.3: Áreas de contatos entre as subunidades do tetrâmetro, calculadas com PISA [69].

Estrutura	A-B (Å^2)	C-D (Å^2)	A-D (Å^2)	B-C (Å^2)	A-C (Å^2)	B-D (Å^2)
Dim. ASU menor	2833,2	2429,3	554,9	412,7	282,9	539,2
Dim. ASU maior	2749,2	2569,8	427,2	507,2	262,3	283,6
Tetra. ASU menor	3856,4	3777,5	542,5	673,9	449,9	387,5
Tetra. ASU maior	3802,9	3937,9	601,8	575,3	345,8	335,3

Condições de cristalização e tetrâmeros cristalográficos

Diante da premissa que existem tetrâmeros simétricos e assimétricos de acordo com a definição da área de contato Área$_{AC}$ e Área$_{BD}$, o procedimento natural é o de conferir novamente as condições de cristalização para que algum indício oriundo das condições experimentais possa ajudar a esclarecer a razão das diferenças dos tetrâmeros ou possa ajudar a descartar algumas possibilidades.

A partir dos dados da Tabela 2.2, as estruturas maiores — consideradas simétricas — foram cristalizadas com e sem substrato 3'-AMP e ambas na presença do cofator metálico Mn^{2+}. Todos os demais componentes são praticamente os mesmos, com exceção do ditiotreitol (DTT: $C_4H_{10}O_2S_2$) que foi adicioando à condição do "Tetra. ASU maior". Antes das estruturas serem refinadas, se especulava sobre a possibilidade de dois resíduos de cisteína estabilizarem o tetrâmero pois, cada monômero possui dois aminoácidos desse tipo. Mais que isto, os resíduos CYS-71 e CYS-192 de acordo com alinhamentos sequenciais realizados, estariam próximos à alça de tetramerização, que é bem conservada entre as SurEs [16]. Poderia se especular que a presença do DTT — um agente redutor — reduzisse possíveis pontes dissulfeto entre esses resíduos de cisteína alterando a forma global da estrutura. No entanto, os dados do refinamento [67] confirmaram que apesar dos resíduos CYS-71 e CYS-192 estarem localizados na alça de tetramerização, possuem uma distância maior que 10 $\mathrm{\AA}$ em ambas as estruturas com e sem DTT. Logo, como a ligação covalente S-S é da ordem de 2 $\mathrm{\AA}$, descartamos a influência que o DTT teria neste tipo ligação e o consecutivo efeito global observado é completamente descartado para a estrutura.

O 3'-AMP também parece não ter sido o causador das assimetrias pois, tanto o cristal "Tetra. ASU menor" (assimétrico) quanto o "Dim. ASU maior" (simétrico) tiveram contato com ele, o primeiro por *soaking* e o segundo co-cristalizado com este substrato. Além disso, "Dim. ASU menor" e "Tetra. ASU maior" estiveram na presença de DTT, na ausência de 3'-AMP e ainda assim são considerados tetrâmeros diferentes.

Devido à fragilidade de alguns cristais de XfSurE quando em contato com soluções crioprotetoras, a formação de gelo cristalino torna-se inevitável na ausência de crioprotetor como exemplifica o padrão de difração (Figura 2.14-B), para o cristal do "Dim. ASU menor". Logo, a última possibilidade aparente é considerar esse efeito indesejado como indutor de mudança estrutural. A

princípio, esta possibilidade é real mas não esclarece o que poderia ter ocorrido quando a situação é comparada com outra estrutura de tamanho "pequeno" como o "Tetra. ASU menor" que foi banhado em glicerol 20$\%$ e não apresentou padrões de difração característicos do gelo.

Contudo, existe uma indeterminação sobre os contatos cristalinos serem ou não os agentes causadores das mudanças conformacionais detectadas. Talvez, a influência da condição de cristalização seja mais branda e ao invés de induzir uma mudança conformacional, a condição tenha apenas selecionado uma conformação estrutural particular inerente aos movimentos coletivos da proteína em solução.

Perfis teóricos de SAXS a partir das estruturas cristalográficas de XfSurE

Antes do avanço em direção à análise de modos normais de vibração de XfSurE, seria interessante verificar o ajuste teórico produzido a partir das estruturas cristalográficas determinadas, sobre a curva de espalhamento experimental $I(q)$ obtida com uma amostra de XfSurE nativa. A Figura 2.16 mostra os ajustes obtidos com as estruturas cristalográficas formadas por dois dímeros (Figura 2.16-A) e com os tetrâmeros na ASU (Figura 2.16-B). Além disso, as respectivas curvas nas Figuras 2.16-A' e 2.16-B', mostram a $p(r)$ para cada caso.

É surpreendente constatar que os perfis teóricos de SAXS das estruturas cristalográficas não se ajustam tão bem à curva experimental da própria proteína. Existem desvios sistemáticos na região $0,10 < q < 0,15$ Å^{-1} da curva $I(q)$ para todos os casos. As distribuições de distâncias $p(r)$ mostram que a XfSurE em solução é maior ($D_{max} = 100$ Å) que as estruturas cristalográficas cuja distância máxima é de ≈ 80 Å, conforme comparação com a escala na Figura 2.15. Essa característica não decorre do efeito esperado de tamanho aparente da proteína em solução devido à camada de hidratação, pois nos cálculos dos perfis teóricos foi levado em conta uma camada de hidratação de ≈ 3 Å o que levaria a um tamanho máximo aparente de ≈ 86 Å ($80 + 2 \cdot 3$), fato também observado no final das curvas $p(r)$ das Figuras 2.16-A' e 2.16-B', cujo $D_{max} \approx 90$ Å.

Estas diferenças entre os perfis teóricos e experimentais podem estar associadas a existência de uma outra forma da proteína em solução, muito provavelmente simétrica — de acordo com o melhor modelo do envelope de SAXS (P222) para a proteína nativa. Os ajustes teóricos obtidos (Figuras 2.16-A e 2.16-B) refletem positivamente esta proposição. Como pode ser visto, as estruturas maioreis ("simétricas") possibilitam melhores ajustes ($\chi = 2,09$ e $\chi = 2,17$) que as estruturas

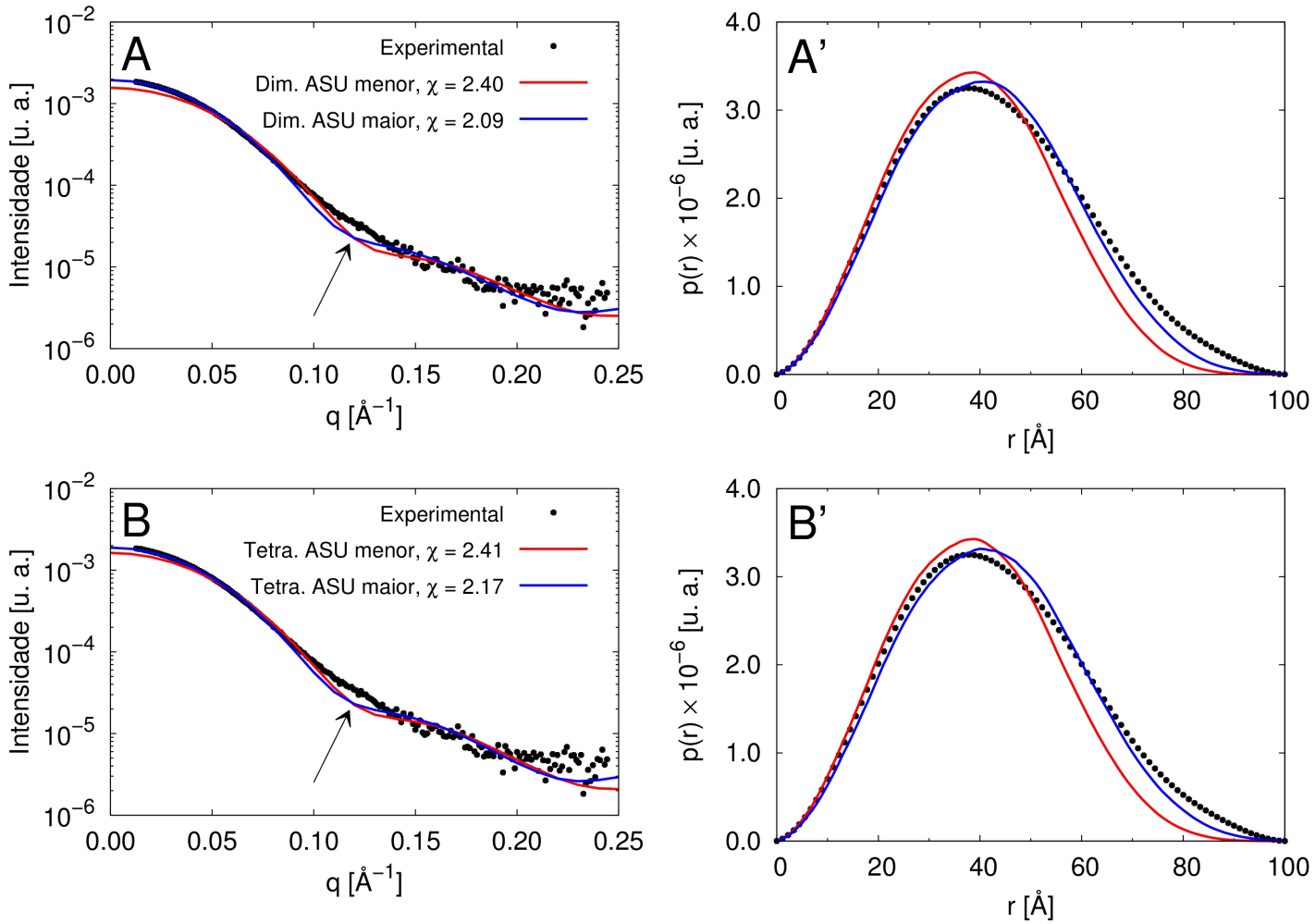


Figura 2.16: Ajustes das curvas teóricas de SAXS calculadas a partir das estruturas tetraméricas cristalográficas de XfSurE construídas pelo conteúdo da unidade assimétrica (ASU). (A e A') Intensidade de espalhamento e (A', B') distribuição de distâncias para tetrâmeros formados por dois dímeros e um tetrâmero na ASU, respectivamente. As setas indicam os desvios sistemáticos entre as curvas $I(q)$ na região $0.10 < q < 0.15 \text{Å}^{-1}$

menores (“assimétricas”) com $\chi = 2,40$ e $\chi = 2,41$.

No entanto, uma ponderação deve ser feita em relação a baixa resolução da técnica experimental. A diferença nos ajustes é da ordem de $\chi \approx 0,3$ que dependendo do caso, é suficiente para distinguir dois modelos estruturais mas em outros, essa pequena diferença em χ inclusive pode identificar um modelo equivocado em detrimento do correto. No caso da XfSurE, os quatro modelos possuem as mesmas características e perfis similares de $I(q)$, tornando o julgamento ainda mais delicado. Pela $p(r)$, é verificado que modelos maiores se ajustam melhor à curva de distribuição de distâncias da XfSurE em solução e portanto, a proposição sobre SAXS “preferir” as estruturas simétricas pode ser, na verdade, apenas o efeito de tamanho dos modelos, já que definitivamente a estrutura da enzima deve ser maior em solução que a cristalográfica.

2.3.3 – Análise de modos normais

Nesta seção, procuramos encaminhar possíveis explicações para os fenômenos observados acerca das características da estrutura quaternária de XfSurE à luz da análise dos modos normais de vibração (AMN)

Modos normais de baixa frequência

A partir da estrutura cristalográfica de XfSurE, aplicamos o modelo de rede elástica (Figura 2.8) para todos os quatro modelos estruturais refinados. É importante reiterar que o modelo utilizado independe de cadeias laterais, bastando apenas as coordenadas dos carbonos alfa de cada resíduo. Por isso, esse tipo de modelagem é denominada *coarse-grained* ou grosseira, pela simplificação imposta.

Tivemos o cuidado de verificar se resíduos chaves na região do sítio ativo e outras área de interface entre as cadeias do tetrâmero deixaram de ser modelados por algum motivo, como por exemplo, ruídos no mapa de densidade eletrônica ou regiões desordenadas pela flexibilidade de *loops*. Para resíduos modelados em dupla conformação de cadeias laterais cujos carbonos alfa estivessem um pouco deslocados, foi considerado o C_α de maior ocupância. Não foi encontrada nenhuma irregularidade ao longo da cadeia principal que pudesse resultar em um modelo elástico tendencioso.

No caso de um tetrâmero de XfSurE — cujo monômero possui 263 aminoácidos — existem

1052 coordenadas cartesianas (x, y, z) associadas aos carbonos alfa, totalizando 3156 graus de liberdade, ou ainda, 3150 modos normais de vibração de interesse, uma vez que os 6 primeiros são modos translacionais e rotacionais da estrutura como um todo. Os primeiros modos [7, 8, ...] são de baixa frequência, onde verifica-se movimentos coletivos de domínios estruturais formados por vários resíduos e os últimos modos [..., 3155, 3156] são de alta frequência, geralmente associados a movimentos não correlacionados de domínios.

Um parâmetro do modelo de rede elástica é o raio de corte R_C cuja definição permite controlar o alcance máximo das interações entre pares de resíduos covalentemente ligados ou não. Por exemplo, um valor $R_C < 3,8$ Å é geralmente um limite inferior pois este alcance é apenas compatível com a típica distância entre C_α-C_α, e apenas interações dessa magnitude não seriam suficientes para estabilizar a estrutura como um todo. Por outro lado, um valor de $R_C > 15$ Å poderia compactar artificialmente a estrutura global da proteína a ser analisada. O valor de $R_C = 10$ Å mostrou-se adequado para as simulações pois, os mesmos movimentos coletivos foram obtidos a partir dos modos normais de baixa frequência calculados com um raio de corte na faixa de $8 < R_C < 12$ Å. Portanto, mantivemos o valor de $R_C = 10$ Å em todas as simulações.

A Figura 2.17 mostra o resultado do cálculo dos primeiros modos normais da rede elástica de um tetrâmero cristalográfico (“Tetra. ASU maior”). O modo representado na Figura 2.17-A, por exemplo, é o sétimo modo normal da estrutura, mas é primeiro modo de baixa frequência que possui movimentos relativos entre os C_α. Em geral, todos os padrões de vibração também são encontrados nas demais estruturas cristalográficas da XfSurE. Entretanto, a numeração do modo pode variar um pouco. Por exemplo, o modo 10 do “Tetra. ASU maior” é equivalente ao padrão de vibração encontrado no modo 11 da “Tetra. ASU menor” e assim por diante.

Modo normal 7

O modo número 7 (Figura 2.17-A) promove uma assimetria em relação ao eixo imaginário vertical que passa pelo centro do tetrâmero. O efeito resultante é permitir o afastamento das cadeias A e C e aproximação das cadeias B e D e vice e versa, caso seja considerado a sentido da vibração inversa.

Este tipo de assimetria foi encontrada anteriormente na estrutura da proteína SurE de *Thermus thermophilus* (TtSurE) com 36% de identidade sequencial com a XfSurE, representada na Figura

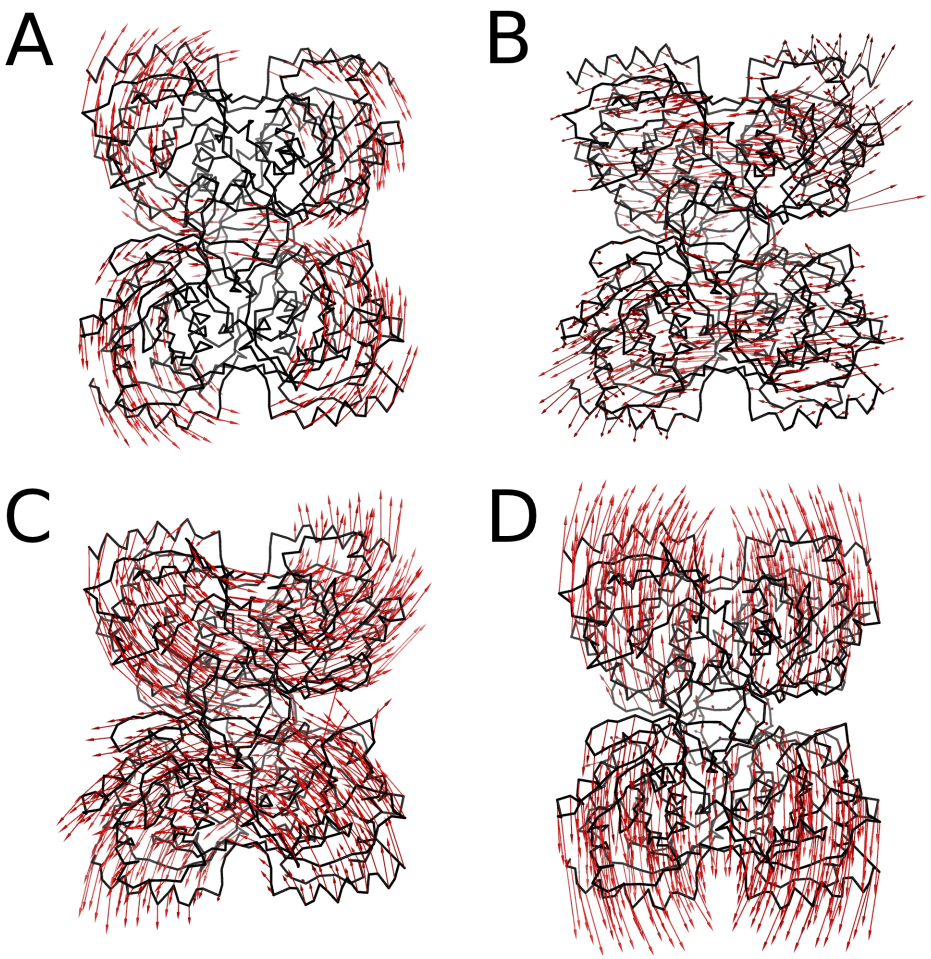


Figura 2.17: Primeiros modos normais de baixa frequência para o maior tetrâmetro de XfSurE (Tetra. ASU maior). A estrutura representa o tetrâmero cristalográfico (linhas pretas). As setas representam os vetores deslocamentos para cada aminoácido da rede elástica segundo a direção especificada por um dado modo normal: (A) 7, (B) 8, (C) 9 e (D) 10. As amplitudes de deslocamento estão exageradamente fora de escala para uma melhor visualização.

2.3 ao centro.

Iwasaki e Miki [27] propuseram um comportamento cooperativo para TtSurE que foi correlacionado com este tipo de assimetria estrutural inerente à enzima apo. Esta correlação foi inferida com base no fato que a estrutura cristalográfica da enzima holo (TtSurE + Mn + AMP) também exibe a mesma assimetria. Assim, o efeito não pôde ser atribuído à presença dos ligantes. No entanto, outra estrutura determinada por Iwasaki e Miki é altamente simétrica com apenas a f presença do

íon metálico. Portanto, a princípio, a estrutura de TtSurE pode exibir ambas as propriedades de simetria e assimetria na presença de ligantes.

Paralelamente, realizamos análise de modos normais para a estrutura desta homóloga de TtSurE e encontramos que a estrutura simétrica se converte na assimétrica a partir de transformações lineares a partir de deslocamentos relacionados às componentes do sétimo modo de vibração nas coordenadas dos C_α. As direções do modo normal e as conhecidas distâncias C_α-C_α entre as estruturas cristalográficas simétrica e assimétrica estão correlacionadas em maior que 50%. De nosso conhecimento, esta foi a primeira vez que os modos normais foram usados para explicar as mudanças conformacionais nas estruturas depositadas no PDB por Iwasaki e Miki.

Modo normal 8

O modo 8 (Figura 2.17-B) permite que os dímeros superiores e inferiores do tetrâmero sofram uma rotação um em relação ao outro, através do eixo vertical que atravessa o centro da estrutura, do mesmo modo que os resultados previstos na ocasião da modelagem de corpo rígido da Seção 2.3.1. Naquele momento, partimos apenas da intuição sobre os possíveis movimentos entre as subunidades que pudessem estabelecer um arranjo que melhorasse o acordo com a curva de SAXS. Agora, partindo apenas da estrutura cristalográfica e da correspondente rede elástica, sem qualquer outro vínculo, o modo normal número 8 surge naturalmente como uma possibilidade vibracional.

Modo normal 9

O modo 9 (Figura 2.17-C) também está associado à rotação entre os dímeros A-B e C-D do tetrâmero. Porém, neste modo, existem dois eixos de rotação, um para cada dímero, que são perpendiculares à página e posicionados nas extremidades longitudinais do tetrâmero, conforme pode ser visto na figura. Não foram encontradas evidências experimentais correspondentes deste tipo de vibração em outras estruturas de SurE, nem mesmo entre as estruturas cristalográficas de SurE de *Xylella fastidiosa*. Se este modo é importante para a função de XfSurE, acreditamos que ele haja similarmente ao modo 7, promovendo o surgimento de assimetrias entre os dímeros superiores (A-B) e os inferiores (C-D).

Modo normal 10

O modo 10 (Figura 2.17-D) tem a propriedade de alongar ou "achatar" longitudinalmente o tetrâmero. Na figura, é bastante claro o estiramento entre os dímeros formados pelas cadeias A-B e cadeias C-D. Por inspeção, este tipo de movimento pode estar correlacionado com as diferenças estruturais observadas entre as estruturas "menores" e "maiores" de XfSurE.

Projeção das diferenças estruturais cristalográficas nos modos normais de XfSurE

As diferenças encontradas na estrutura quaternária entre os tetrâmeros cristalográficos de XfSurE podem ser quantificadas com respeito aos modos normais de vibração da molécula. A estratégia considera pares de tetrâmeros: o inicial — onde é aplicado o modelo de rede elástica na qual é feito o cálculo dos modos normais; e o final — cuja estrutura também serve como referência de ponto de chegada para a estrutura inicial além de ser usada no cálculo das diferenças cristalográficas. O primeiro passo consiste na sobreposição estrutural dos dois tetrâmeros com o uso do programa TM-score [68]. Em seguida, todos os vetores posição entre dois resíduos de aminoácidos correspondentes nas duas estruturas são projetados nas respectivas componentes de um modo normal específico. Para cada modo, as diferenças estruturais entre as estruturas inicial e final dos tetrâmeros são quantificadas pela Equação 2.9. A interpretação deste cálculo é simplesmente equivalente ao produto interno entre dois vetores no espaço e portanto, se dois vetores são idênticos, a projeção é máxima e igual a 1 (um).

Para maior clareza, a Figura 2.18-A mostra o resultado da projeção da diferença estrutural entre os tetrâmeros "Tetra. ASU menor" e "Tetra. ASU maior" de XfSurE, cujo RMSD é 2,430 $\mathrm{\AA}$, nos 50 primeiros modos normais do tetrâmero menor. Pelo gráfico, o modo normal 11 estaria relacionado com as diferenças cristalográficas devido ao pico de projeção (P_{11}) que se sobressai em relação aos demais. Analisando a Figura 2.18-B, nota-se que a diferença estrutural dos tetrâmeros cristalográficos considerados é global, isto é, as diferenças não estão pontualmente localizadas, nem seguem um padrão aleatório. O tetrâmero maior é praticamente uma expansão do menor por um fator de escala.

Partindo do tetrâmero menor e do seu modo normal 11, o padrão do deslocamento que este modo fornece para cada resíduo é realmente semelhante ao das diferenças cristalográficas (Figura 2.18-C). Isto fortalece a hipótese do tetrâmero maior poder ser obtido a partir da

estrutura do menor. Neste caso, a transformação do tetrâmero menor para o maior ocorre através de apenas um modo vibracional, correlacionando os padrões cuja sobreposição é de $P_{11} = 0,56$. Distorcendo o tetrâmero menor por deslocamentos guiados pelas componentes do modo normal 11, com o objetivo de minimizar o RMSD entre a estrutura distorcida e a final, o RMSD diminui de 2,430 $\mathrm{\AA}$ para 2,015 $\mathrm{\AA}$. Este resultado reforça a evidência de não aleatoriedade entre os padrões de deslocamentos observados entre as Figuras 2.18-B e 2.18-C ilustrando o poder desta ferramenta de análise.

O próximo passo na análise consiste em projetar as diferenças entre todas as quatro estruturas cristalográficas de XfSurE nos modos normais de uma estrutura inicial qualquer do conjunto dos quatro tetrâmeros. Isso é importante pois, uma transformação de $A \rightarrow B$ não é necessariamente equivalente a transformação $B \rightarrow A$, já que os modos normais da estrutura inicial (A ou B) podem diferir um pouco uns dos outros, justamente pela conformação ser diferente gerando redes elásticas distintas. Portanto, a combinação de todas as possibilidades levam a 12 histogramas análogos ao da Figura 2.18-A.

A Figura 2.19 mostra todos os resultados das 12 combinações entre as 4 estruturas de XfSurE. As setas indicam os picos de projeção para a transição $A \rightarrow B$ em um dado modo normal da estrutura A. Observa-se que os modos normais mais envolvidos com $P_j \gtrsim 0.40$ são os modos 7, 10, 11 e 12 que ocorrem nas diferentes estruturas. O modo 7 de todas as quatro estruturas possui um padrão de movimento do modo de mesma denominação já mostrado na Figura 2.17-A. Os modos 10, 11, e 12 que surgem na Figura 2.19 em diferentes transições são equivalentes ao modo 10 de “Tetra. ASU maior” que também está mostrado na Figura 2.17-D.

Em geral, $A \rightarrow B \neq B \rightarrow A$ mas observa-se que uma característica predominante nas conversões dos tetrâmeros cristalográficos de XfSurE é que a operação $A \rightarrow B$ corresponde a operação $B \rightarrow A$, tornando a matriz praticamente simétrica em relação a diagonal.

Uma exceção é com a transformação do tetrâmero “Dim. ASU menor” (A) para “Tetra. ASU maior” (B) que está correlacionada com o décimo modo normal da primeira estrutura enquanto que transformação na ordem inversa “Tetra. ASU maior”(A) para “Dim. ASU menor” (B) está correlacionada não só com o décimo mas com o sétimo modo também. A Tabela 2.3 fornece pistas para este caso. O tetrâmero “Dim. ASU menor” tem uma área de interface entre a cadeia A e cadeia C de Área$_{AC} = 282,9\mathrm{\AA}^2$ enquanto que o contato das cadeias B e D é de Área$_{BD} = 539,2\mathrm{\AA}^2$. Isso faz com que surja uma pequena assimetria entre os dímeros superiores (cadeias

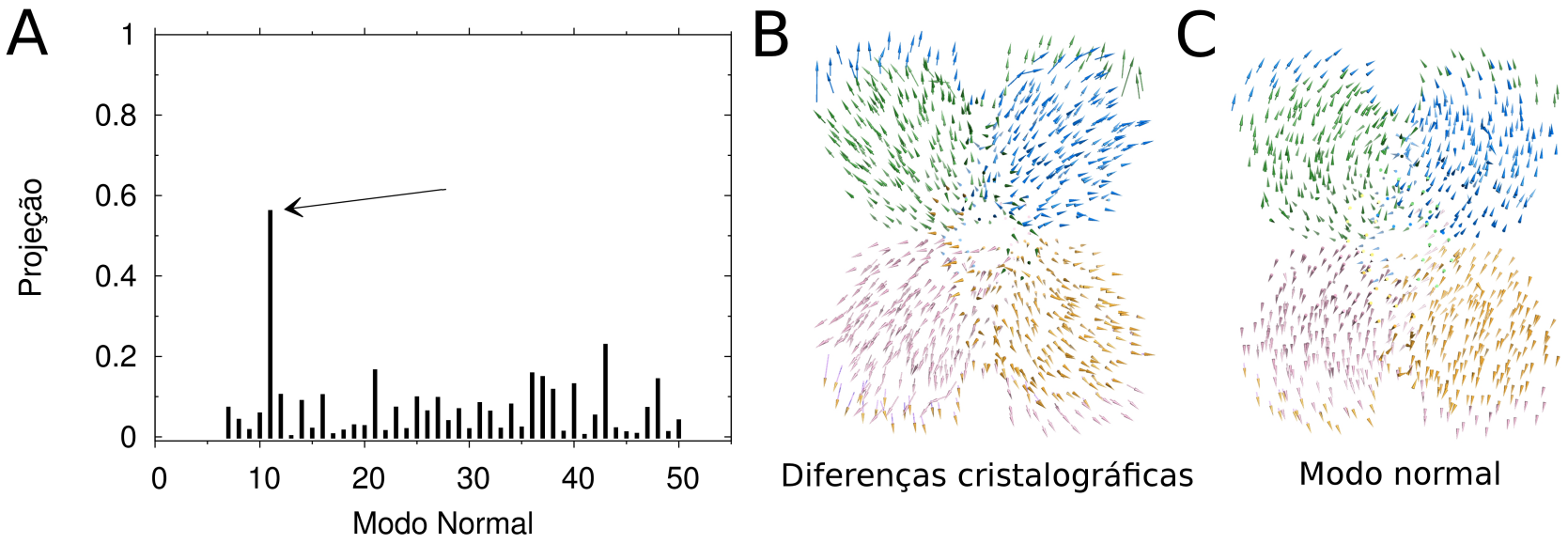


Figura 2.18: Exemplo de projeção das diferenças cristalográficas de XfSurE nos modos normais. (A) Projeção das diferenças cristalográficas entre os tetrâmeros “Tetra. ASU menor” e “Tetra. ASU maior” nos modos normais da rede elástica calculadas para “Tetra. ASU menor”. (B) As setas indicam as distâncias entre os resíduos correspondentes nas duas estruturas cristalográficas. (C) As setas indicam as direções dos deslocamentos dos resíduos da estrutura “Tetra. ASU menor” promovidas pelo modo normal 11.

A-B) e inferiores (cadeias C-D). Interessantemente, esse tipo de assimetria é explicada pelo modo normal 7 que, conforme discussão acima, também foi detectada na SurE de *Thermus thermophilus* [27]. Sendo o "Tetra. ASU maior" uma estrutura mais simétrica (Área$_{AC} = 345,8\text{Å}^2$ e Área$_{BD} = 335,3\text{Å}^2$), a conversão direta $A \rightarrow B$ é mais correlacionada com o modo 10 cujo efeito é de alongar o tetrâmero menor assimétrico convertendo-o em uma estrutura maior e simétrica. Na situação inversa, além do modo 10 — que agora agiria como o responsável pelo "achatamento" do tetrâmero maior simétrico — o modo 7 surgiria como o agente causador da assimetria.

Em síntese, todas as combinações $A \rightarrow B$ envolvendo apenas estruturas "maiores" ou "menores" possuem apenas o modo normal 7 correlacionado com a transição. Neste caso, este modo pode surgir pela necessidade de uma pequena correção ou distorção da simetria molecular.

Quando há transições do tipo "maior" $\leftrightarrow$ "menor" o modo 10 inevitavelmente está envolvido pois este modo normal é responsável pela alteração do tamanho longitudinal da estrutura. O modo 7 pode surgir de maneira associada nesse tipo de transição para preservar ou não a simetria.

Correlação dos modos normais de baixa frequência com os dados de SAXS

Foram identificados dois prováveis modos normais (modos 7 e 10) que estariam relacionados a movimentos de grande amplitude e supostamente relacionados aos movimentos intrínsecos de XfSurE em solução. Assim, a intenção final neste estudo é propor um modelo de baixa resolução que se ajuste melhor aos dados de SAXS da XfSurE, para a sua forma nativa e na presença de seu substrato 3'-AMP.

Uma busca computacional extensiva foi empregada utilizando todas as quatro estruturas cristalográficas e seus respectivos modos normais número 7 e 10. Existem padrões de movimentos equivalentes produzidos pelos modos 10, 11 e 12 entre as estruturas. Logo, o termo "modo 10" refere-se ao padrão de movimento ilustrado na Figura 2.17-D.

A partir da estrutura tetramérica cristalográfica e respectivos modelos sucessivamente distorcidos por imposições prescritas por um dado modo normal, a curva de espalhamento teórica $I(q)$ foi sendo ajustada sobre a curva experimental de SAXS da XfSurE na forma nativa. Houve um excelente ajuste usando o modo normal 10 e a estrutura "Tetra. ASU maior" como ponto de partida. Os resultados podem ser conferidos na Figura 2.20-A. Nesta figura, é feita uma comparação direta entre as curvas $I(q)$ da estrutura de partida ($\chi = 2,17$) e do modelo criado ($\chi = 1,72$). A caracte-

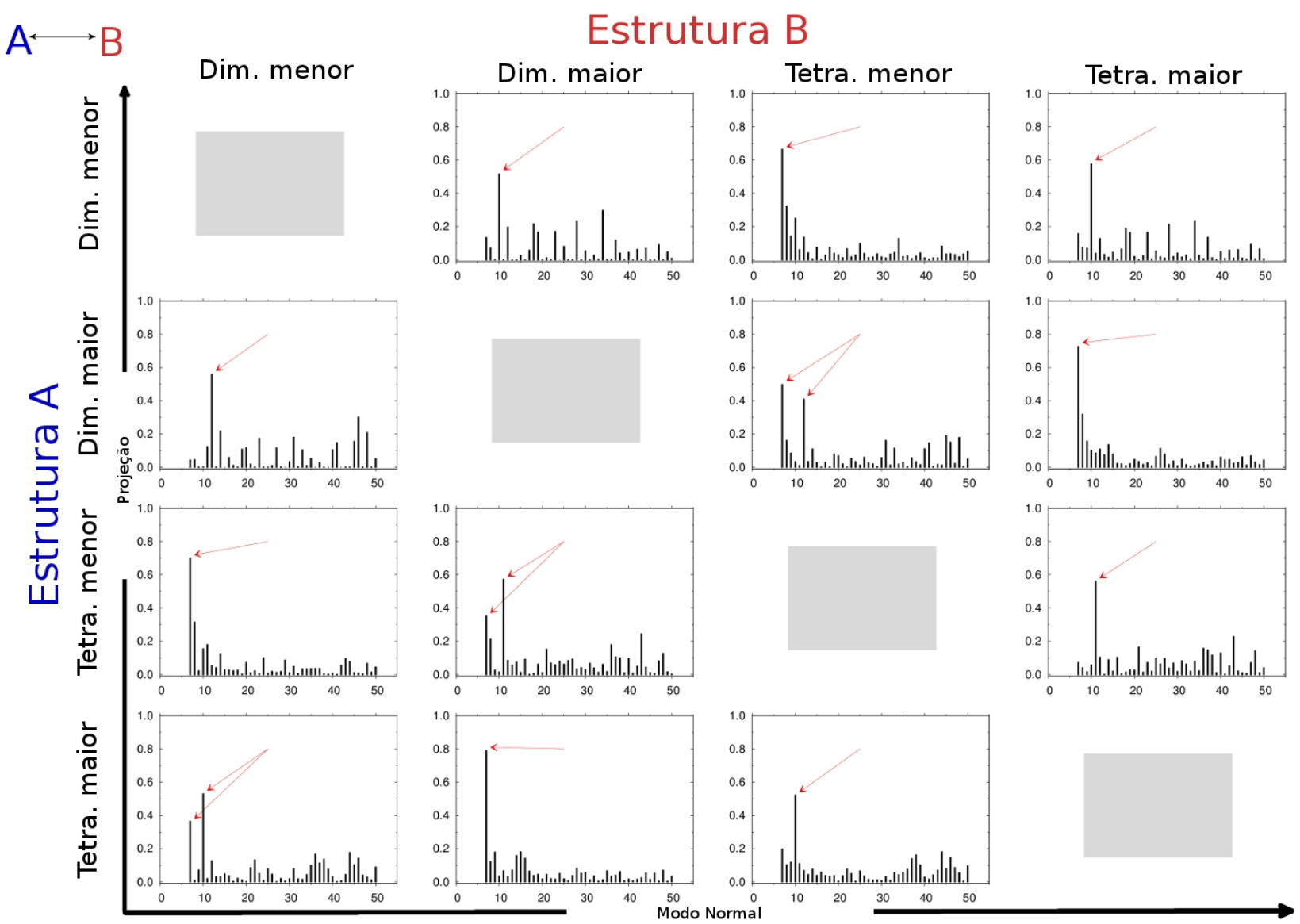


Figura 2.19: Projeção das diferenças estruturais dos tetrâmeros cristalográficos de XfSurE nos modos normais de vibração.

rística marcante da qualidade do ajuste é observada na região $0,10 < q < 0,15\text{Å}^{-1}$ onde todas as estruturas cristalográficas apresentam um desvio sistemático. Nesta região, o modelo apresenta um excelente acordo com a $I(q)$ experimental, mostrando que um tetrâmero ainda maior estaria presente em solução. A figura interna de 2.20-A, mostra a curva de minimização a partir da estrutura inicial (modelo = 0). Após 13 distorções sucessivas, um mínimo global é encontrado. Nesta situação, a Figura 2.20-C mostra que o modelo já está alongado suficientemente para dar conta de um melhor encaixe no envelope de SAXS em relação ao tetrâmero "Tetra. ASU maior". Independentemente, o mesmo procedimento foi realizado minimizando a curva $p(r)$. Neste caso, um modelo um pouco diferente é encontrado (modelo 9) mas que também possui um grande acordo com a $p(r)$ calculada diretamente da $I(q)$ experimental conforme mostra a Figura 2.20-B.

O modelo encontrado ainda preserva a simetria do envelope e é aproximadamente semelhante à estrutura original usada como ponto de partida, embora tenha sido demonstrado que as distorções introduzidas por apenas um modo normal (modo 10) são capazes de minimizar as diferenças dos perfis de SAXS. Em primeira instância, este resultado tem o sentido de evidenciar que a estrutura da XfSurE em solução é mais alongada que suas formas cristalográficas. Em uma segunda análise e considerando a aproximação do modelo de modos normais, pode-se dizer que é surpreendente que apenas um modo normal esteja correlacionado com estruturas em condições experimentais distintas: o ambiente cristalino e em solução. Por isto, infere-se que as transformações da estrutura quaternária observadas cristalograficamente sejam inerentes à proteína e não produzidas artificialmente pelo empacotamento cristalino.

Um segundo modelo foi produzido, mas desta vez, considerando a curva de SAXS para a XfSurE na presença do seu substrato de maior afinidade, o 3'-AMP.

O primeiro método de construção do modelo é análogo ao empregado anteriormente para XfSurE na forma nativa, ou seja, uma busca extensiva foi feita usando todos os tetrâmeros cristalográficos e as respectivas distorções produzidas pelos modos normais. Desta vez, um único modo normal não foi suficiente para produzir um modelo que tivesse um bom acordo com curva $I(q)$ experimental e não criasse uma interpretação ambígua. Diversos mínimos locais foram encontrados semelhantemente aos dados de modelagem de corpo rígido inicialmente realizados para a XfSurE. Coincidentemente, os modos 7 e 10, de maneira independente, forneceram alguns indicativos de boa qualidade no ajuste teórico com a curva $I(q)$ e também com a $p(r)$. Em todos esses casos, a estrutura cristalográfica "Tetra. ASU maior" foi identificada novamente como a

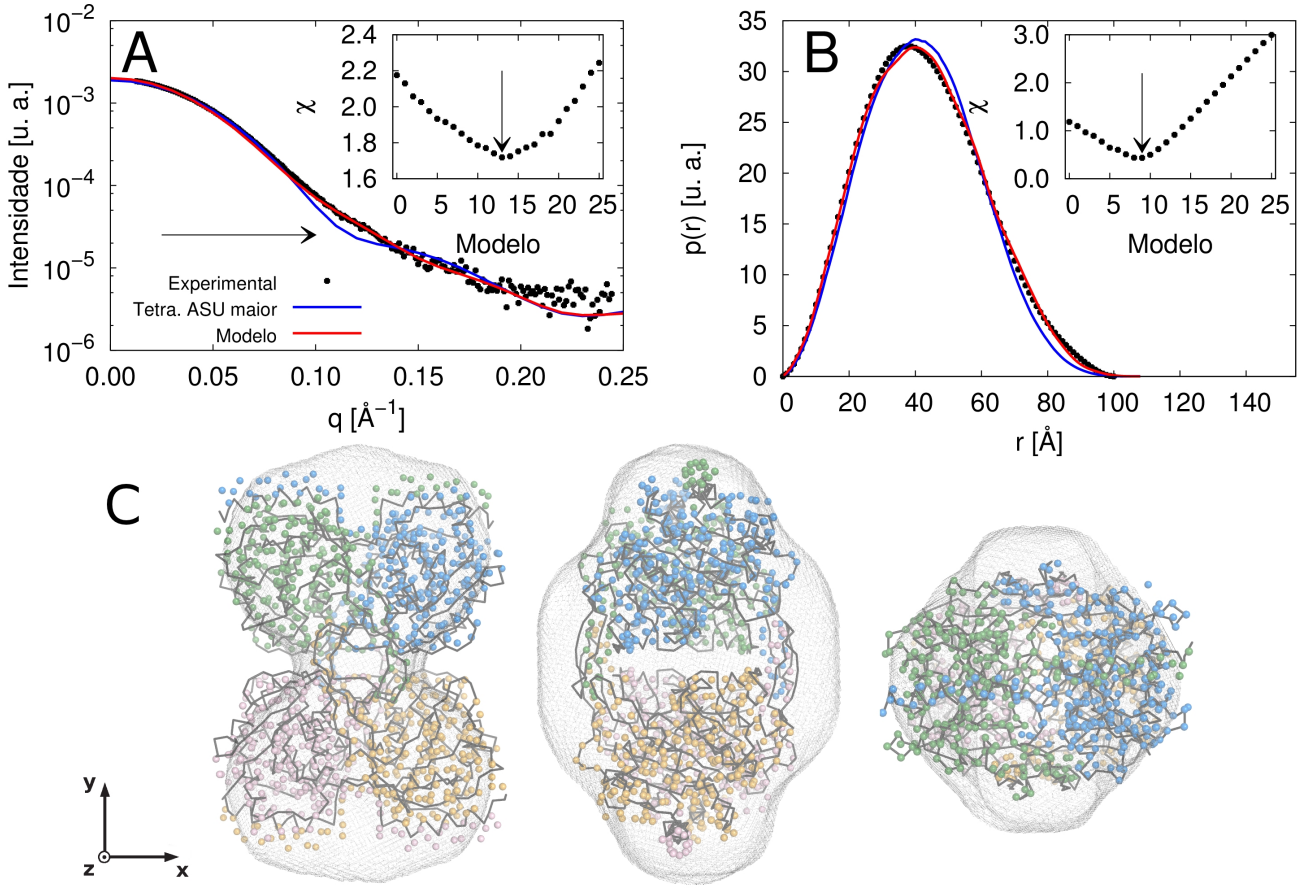


Figura 2.20: Ajustes teóricos das curvas (A) $I(q)$ e (B) $p(r)$ sobre os dados experimentais de SAXS de XfSurE na forma nativa através do estiramento longitudinal da estrutura quaternária cristalográfica denominada “Tetra. ASU maior” pelo seu décimo modo normal de vibração. Os gráficos internos de ambas as figuras A e B mostram a presença de um mínimo global da qualidade do ajuste χ em função do grau de estiramento no modelo. O modelo inicial (índice zero) é referente à estrutura cristalográfica usada como ponto de partida. (C) O envelope molecular obtido por SAXS (cinza), a estrutura cristalográfica (linhas contínuas pretas) e o modelo resultante do ajuste pelo modo normal (esferas) estão sobrepostos em três orientações espaciais. As orientações ao centro e à direita são obtidas por rotações de 90^0 em relação ao eixo y e ao eixo x, respectivamente.

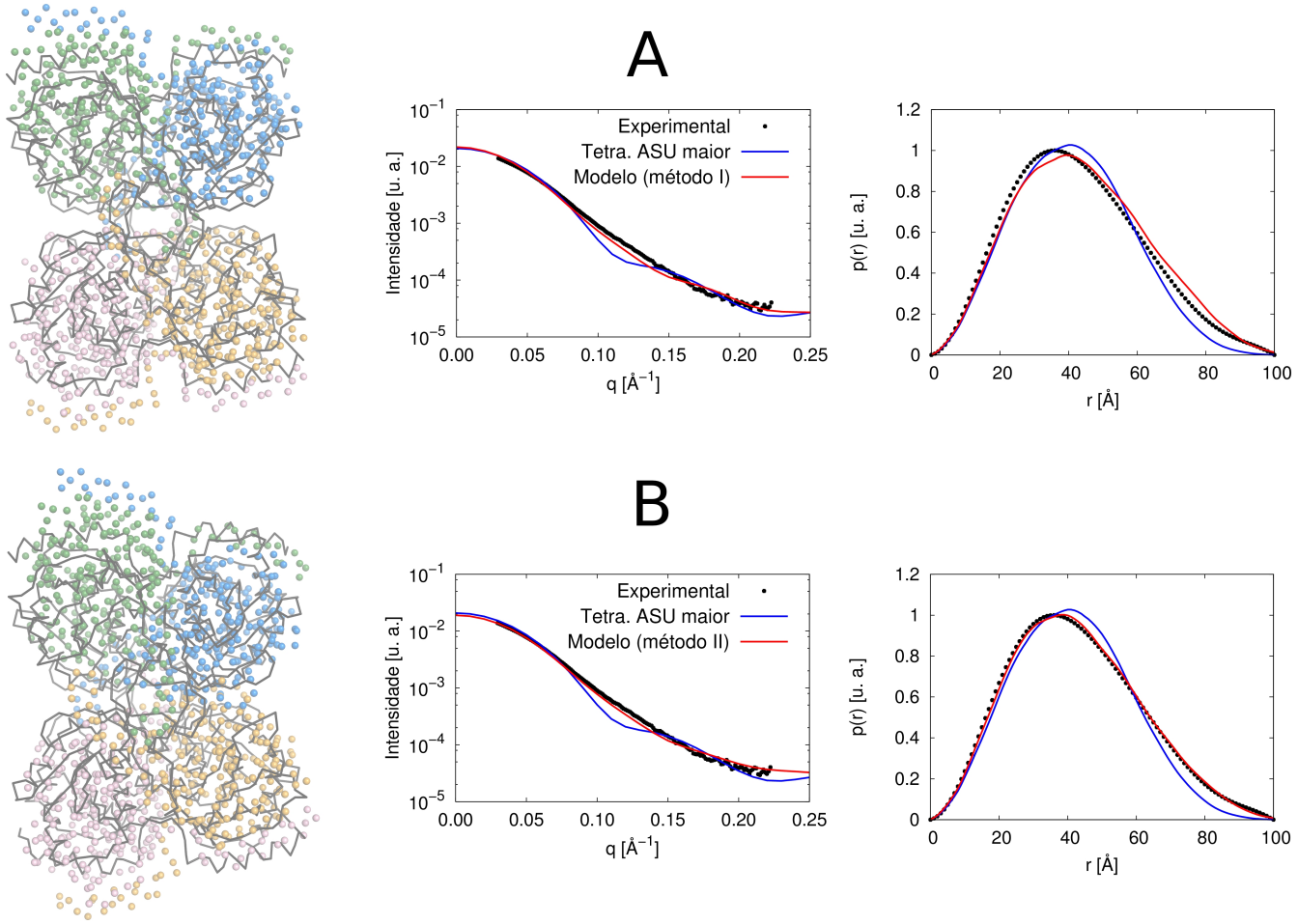


Figura 2.21: Ajustes teóricos das curvas $I(q)$ e $p(r)$ sobre os dados experimentais de SAXS de XfSurE na presença de 3'-AMP através de dois métodos. (A) Método I: Combinação dos modos normais 7 e 10 da estrutura cristalográfica "Tetra. ASU maior". (B) Método II: Combinação do modo normal 10 com rotação de corpo rígido do dímero formado pelas subunidades A e B do tetrâmero de XfSurE em relação ao eixo ortogonal ao plano da página. A estrutura cristalográfica (linhas contínuas pretas) e os respectivos modelos (esferas) estão sobrepostos para cada caso.

melhor estrutura de partida.

Considerando os vários mínimos locais localizados a partir de resultados preliminares, decidiu-se limitar a região de procura usando apenas a combinação dos modos normais 7 e 10 do tetrâmero “Tetra. ASU maior”. Os resultados do melhor modelo distorcido por AMN são mostrados na Figura 2.21-A. Este modelo possui um melhor acordo com a curva experimental ($\chi = 1,59$) da XfSurE + 3’-AMP em relação ao ajuste para a estrutura cristalográfica de partida ($\chi = 2,75$), principalmente na região $0,10 < q < 0,15\text{Å}^{-1}$ que neste caso, exibe um desvio ainda mais acentuado que aquele observado anteriormente na curva de XfSurE na forma nativa. O estiramento longitudinal do modelo de partida auxilia a minimizar esse desvio indicando que XfSurE em solução na presença de 3’-AMP também é mais alongada que suas formas cristalinas. O efeito produzido pelo modo normal 7 teria um papel secundário de corrigir as discrepâncias com a respectiva $p(r)$ já que o modo normal 10 produziria um irregularidade maior. Entretanto, o modo 7 produz uma assimetria na estrutura final fazendo com que as cadeias A e C sejam mais afastadas uma da outra e que as cadeias B e D sejam aproximadas. Pela análise da $p(r)$, os dados sugerem que correções ainda precisam ser feitas no modelo, devido aos desvios observados na região de seu máximo, em torno de 40 Å e também na região 60 < r < 90 Å em relação à curva “experimental”.

Independentemente, outro método de construção do modelo também foi executado. Desta vez, partindo da mesma estrutura cristalográfica inicial e de seu modo normal 10, um movimento adicional de corpo rígido entre os dímeros formados pelas cadeias A-B e C-D também foi permitido. Esse movimento consistiu em uma rotação do dímero A-B em relação ao eixo imaginário que atravessa o centro do tetrâmero, perpendicularmente ao plano da página. Esse movimento de rotação arbitrário tende a imitar o movimento equivalente produzido pelo modo normal 7. Uma busca entre todas as combinações de estiramento e ângulos de rotação foi feita. O melhor ajuste aos dados de SAXS é conseguido com a distorção longitudinal produzida pelo modo 11 associada a uma rotação de 45^0 do dímero A-B em relação a sua posição original na estrutura cristalográfica. Os resultados são apresentados na Figura 2.21-B. O resultado deste segundo método produz um novo modelo que reduz a discrepância entre as curvas de intensidade $I(q)$ em relação ao modelo anterior, de $\chi = 1,59$ para $\chi = 1,34$. Além disto, a correspondente $p(r)$ possui um excelente acordo com a curva obtida diretamente da curva $I(q)$ experimental.

Vale ressaltar que a enzima XfSurE é uma nucleotidase e tem preferência por nucleosídios

monofosfatados, especialmente o 3'-AMP [16]. Sua função catalítica é desfosforilar o substrato, mas para isso, depende de um cofator metálico (Mn^{2+}) que deve estar ligado ao sítio ativo localizado na região central do tetrâmero. A amostra de XfSurE submetida ao SAXS foi preparada apenas com a enzima e seu substrato e portanto, nenhuma reação catalítica foi possibilitada devido a ausência do cofator metálico. As amostras foram preparadas com 100 mM de 3'-AMP que leva a uma relação estequiométrica de aproximadamente 300 moléculas de substrato para cada monômero de XfSurE.

Aparentemente, estes resultados em conjunto sugerem que as moléculas do substrato tenham induzido uma mudança conformacional na estrutura quaternária de XfSurE. Os dois modelos produzidos por dois métodos — modos normais apenas e modos normais combinados a movimentos de corpo rígido — concordam entre si e realçam a característica estrutural assimétrica presente entre os dímeros superiores (A-B) e inferiores (C-D).

É preciso levar em consideração que as estruturas cristalográficas também apresentam esse tipo de assimetria. Pelas condições de cristalização, não é claro que 3'-AMP tenha sido o causador desse efeito pois, conforme discutido, tanto o cristal "Tetra. ASU menor" (assimétrico) quanto o "Dim. ASU maior" (simétrico) tiveram contato com ele: o primeiro por *soaking* e o segundo por co-cristalização.

Ao que parece, ambas as estruturas simétrica e assimétrica estariam presentes em solução, mesmo porque, uma pode ser convertida na outra por um modo normal de vibração de baixa frequência. Dessa forma, as condições do ambiente microscópico induziriam o aumento populacional de uma configuração em detrimento da outra.

Esta característica simetria$\leftrightarrow$assimetria é análoga aos estados relaxado e tensionado do modelo clássico de Monod-Wyman-Changeux (MWC) [33] para proteínas alostéricas. Em outro sentido, acreditamos que os resultados alcançados aqui possam servir como ponto de partida para investigações estruturais mais acuradas sobre os mecanismos alostéricos da enzima SurE de *Xylella fastidiosa* tomando como referência os estados simétrico e assimétrico da XfSurE.

2.4 – Conclusão

O presente estudo consistiu na caracterização estrutural da proteína SurE de *Xylella fastidiosa* (XfSurE) por técnicas experimentais e computacionais. Estudos estruturais da XfSurE realizados com a técnica de espalhamento de raios X a baixos ângulos (SAXS) apontaram para um arranjo tetramérico da enzima apo e, do nosso conhecimento, foi a primeira estrutura em solução descrita na literatura para esta família de proteínas [16]. Modelagem de corpo rígido foi utilizada com o objetivo de melhorar o acordo entre a curva teórica de SAXS predita para o tetrâmero cristalográfico de uma proteína homóloga com os dados experimentais de XfSurE, rearranjando os domínios estruturais do tetrâmero em um espaço de configurações que permitisse um melhor ajuste no envelope determinado *ab initio*. Como resultado, esta abordagem além de fornecer um modelo que se ajusta melhor aos dados de SAXS da XfSurE na forma nativa, abriu novos caminhos em direção ao estudo das propriedades alostéricas que seriam confirmadas posteriormente [16].

A caracterização preliminar por cristalografia [17] mostrou a variabilidade da unidade assimétrica (ASU) que pode ser identificada por dímeros ou tetrâmeros nos cristais da SurE de *Xylella fastidiosa* evidenciando o valor de estudos em solução por SAXS na determinação da unidade biológica funcional da enzima que, a princípio, não possuem nenhum compromisso com o conteúdo da ASU.

Recentemente, algumas abordagens do tipo “ajuste flexível” para SAXS tem surgido na literatura [70]. Entretanto, este procedimento relativamente comum na área de microscopia eletrônica [71], pode ser adaptado para a técnica de espalhamento de raios X a baixos ângulos [72]. Porém, destacamos que um servidor que correlacione os modos normais aos dados de SAXS, de maneira automática, ainda não se encontra disponível. Diante da escassez de ferramentas computacionais que atenderiam nossas necessidades, automatizamos todo o processo de construção e análise dos modelos a partir de *scripts* que escrevemos possibilitando uma abordagem em larga escala.

As estruturas cristalográficas de XfSurE obtidas em colaboração, apresentam diferenças sutis na estrutura quaternária que podem ser explicadas pelos primeiros modos normais de vibração de baixas frequências associados aos movimentos coletivos com grandes amplitudes. Esta foi a primeira descrição dos modos normais para uma proteína da família SurE. No caso da SurE de *Xylella fastidiosa*, os resultados afastam a hipótese que as diferenças conformacionais observa-

das seriam frutos de uma distorção artificial promovida pelo empacotamento cristalino. É mais provável que uma condição de cristalização tenha selecionado um particular modo vibracional. Neste caso, as diferenças estruturais estariam associadas a uma genuína transição da estrutura quaternária da XfSurE.

Os dados cristalográficos e os modos normais também se correlacionam com os dados de SAXS, indicando que um tetrâmero simétrico mais alongado estaria presente em solução em amostras da enzima apo. Já em amostras com o substrato 3'-AMP, um tetrâmero assimétrico modelado a partir de uma combinação de dois modos normais estaria presente. Esses resultados tornam-se importantes à medida em que se sabe que XfSurE é uma enzima alostérica e movimentos de subunidades podem estar relacionados ao seu mecanismo de regulação.

Assim, uma extensão lógica deste trabalho na direção da caracterização computacional dos mecanismos alostéricos, seria o estudo por simulações de dinâmica molecular (DM). Entretanto, é praticamente certo que simulações de DM atomística de uma proteína com milhares de átomos — como é o caso do tetrâmero da XfSurE — esbarrem em problemas técnicos. Tradicionalmente, DM é aplicada em uma escala temporal tipicamente compreendida no intervalo de dezenas a centenas de nanosegundos enquanto que movimentos de grande amplitude — como os descritos para XfSurE — relacionados a transições alostéricas de outras proteínas estão associados a uma escala temporal da ordem de nanossegundos - milissegundos [39, 40]. Alternativamente, diversas simulações de DM em paralelo poderiam ser inicializadas a partir dos tetrâmeros cristalográficos e as respectivas estruturas distorcidas pelos modos normais relaxadas com os campos de forças usuais. A afinidade do ligante no sítio poderia ser calculada e eventuais caminhos de acesso poderiam ser mapeados em função de cada estágio de deformação do tetrâmero segundo um modo normal.

Ainda que técnicas experimentais de baixa resolução e técnicas computacioanais do tipo *coarse-grainded* tenham sido predominantes neste trabalho, os resultados apontam que os modos normais 7 e 10 de XfSurE identificados por SAXS e cristalografia têm como característica fundamental a alteração das áreas da interface via uma transição do tipo simetria$\leftrightarrow$assimetria que poderiam controlar o acesso do substrato aos quatro sítios ativos do tetrâmero. Do ponto de vista do modelo clássico de Monod-Wyman-Changeux [33] para proteínas alostéricas, XfSurE incluiria apenas duas configurações: a relaxada (R) e a tensionada (T), correspondentes às estruturas tetramérica simétrica e assimétrica, respectivamente.

2.4 Conclusão

Neste sentido, os resultados alcançados podem servir como ponto de partida para investigações mais acuradas sobre os mecanismos alostéricos da XfSurE. Estes achados devem ser incluídos em um artigo que encontra-se em preparação.

Capítulo 3

Classificação de objetos tridimensionais por métricas inspiradas em SAXS

O estudo descrito neste capítulo contempla a análise do reconhecimento de formas de objetos tridimensionais das mais variadas categorias com o propósito de ilustrar uma estratégia computacional de busca por formas semelhantes, dado um objeto-alvo.

O experimento computacional aqui descrito foi realizado com objetos diversos, como por exemplo, xícaras, carros e plantas e portanto, completamente distintos das estruturas tridimensionais assumidas por proteínas, que é a temática deste trabalho de doutorado. Estes objetos de uso cotidiano são facilmente distinguíveis aos olhos humanos e, por esta razão, nossa abordagem têm um caráter puramente pedagógico no tocante às idéias que serão introduzidas mais adiante no Capítulo 4, quando teremos a oportunidade de aplicá-las no reconhecimento do enovelamento proteico.

3.1 – Introdução

As técnicas de classificação e reconhecimento de estruturas tridimensionais constituem um tópico de uma área mais abrangente de pesquisa denominada reconhecimento de padrões e podem ser aplicadas aos mais variados assuntos e objetivos. Na computação gráfica [73], essas técnicas têm sido utilizadas como ferramenta de busca de modelos tridimensionais disponíveis na *web* semelhantemente como palavras-chave são procuradas em *sites* tal como o *google*. As aplicações desses modelos se estendem desde à indústria de jogos eletrônicos ao desenvolvimento de próteses humanas. Na quimiometria [74], técnicas de reconhecimento de padrão são largamente utilizadas em bancos de dados moleculares na procura de modelos estruturais ou *scaffolds* de compostos candidatos a fármacos. Já no campo da bioinformática estrutural [75], a técnica de modelagem comparativa se vale, entre outras abordagens, do reconhecimento de formas estruturais conhecidas na predição da estrutura molecular da proteína de interesse.

Um dos desafios da técnica de reconhecimento de formas tridimensionais reside na proposição de métodos de comparação entre dois objetos. Para que o grau de similaridade entre duas estruturas seja medido, diversas métricas podem ser empregadas, todas devendo satisfazer o requisito mínimo de responder, por exemplo, as seguintes perguntas: Os objetos pertencem a uma mesma classe? Até que ponto conseguimos, a partir da métrica X, distinguir uma xícara de um avião? E uma laranja de uma maçã?

Na tentativa de ilustrar e endereçar estas questões, um experimento computacional foi elaborado com objetos tridimensionais de variadas classes (xícaras, carros, plantas, corpo humano, móveis, etc.) que são facilmente distinguíveis aos olhos humanos e por isso esta escolha tem caráter puramente pedagógico. O experimento visa mensurar estatisticamente o reconhecimento de objetos pertencentes a uma mesma classe dentro de um banco de dados heterogêneo a partir de métricas inspiradas em SAXS baseadas nas propriedades do objeto como o raio de giro R_G e a função de distribuição de pares $p(r)$.

3.1.1 – Objetivos

O principal objetivo deste capítulo é o de explorar, por meio de analogias, o conceito no qual os pesquisadores da área de proteínas e usuários da técnica de SAXS se baseiam para inferirem

sobre o conteúdo informacional dos dados de SAXS ser suficiente para discriminar estruturas tridimensionais em solução, mesmo sendo SAXS uma técnica inerentemente de baixa resolução.

O estudo proposto neste capítulo, embora realizado com objetos tridimensionais das mais diversas categorias, fornece subsídios para a melhor compreensão da aplicação destas técnicas no reconhecimento do enovelamento de estruturas de proteínas que será abordado posteriormente.

Diante destas considerações, acredita-se que, ao menos do ponto de vista pedagógico, a analogia entre uma estrutura proteica com a de um objeto 3D genérico pode ser conceitualmente percebida preparando o leitor para a discussão que se seguirá no Capítulo 4.

3.2 – Metodologia

3.2.1 – Proteínas e objetos 3D

De uma maneira bastante aproximada, proteínas podem ser representadas por uma estrutura de pontos discretos distribuídos no espaço tridimensional através das respectivas coordenadas atômicas. Analogamente, objetos tridimensionais tais como xícaras, aviões ou mesas podem ser representados computacionalmente através de um conjunto de N pontos, onde cada um é representado pela tripla (x_i, y_i, z_i). Entretanto, uma diferença deve ser mencionada. A diferença consiste na falta de informação volumétrica dos objetos 3D utilizados aqui, isto é, um objeto é descrito apenas pela sua superfície. A criação artificial de um volume associado à superfície de um objeto poderia trazer problemas topológicos intrínsecos. Seria difícil ou até mesmo impossível, por exemplo, discriminar duas xícaras idênticas se uma estivesse cheia de líquido e a outra vazia ou então comparar uma esfera sólida a uma casca esférica. De fato, o reconhecimento de padrão entre estes objetos exemplificados poderia falhar classificando-os como objetos de diferentes “classes”, embora aparentemente não sejam para os propósitos aqui apresentados. Por essa razão, considera-se aqui apenas os pontos contidos ao longo da superfície dos objetos 3D.

3.2.2 – Banco de dados de estruturas tridimensionais

Objetos 3D (xícaras, carros, plantas, corpo humano, etc) utilizados neste trabalho foram retirados do *Princeton Shape Benchmark* [76] muito utilizado em estudos de reconhecimento de

padrão desta modalidade. Este banco de dados possui um total de 1814 modelos tridimensionais divididos manualmente em 161 classes.

3.2.3 – Amostragem Monte Carlo

A superfície de um objeto pode ser descrita por uma malha formada por polígonos das mais variadas formas onde cada ponto (x_i, y_i, z_i) representa um dos vértices. A malha do tipo triangular é considerada como a mais simples (vide Figura 3.1A).

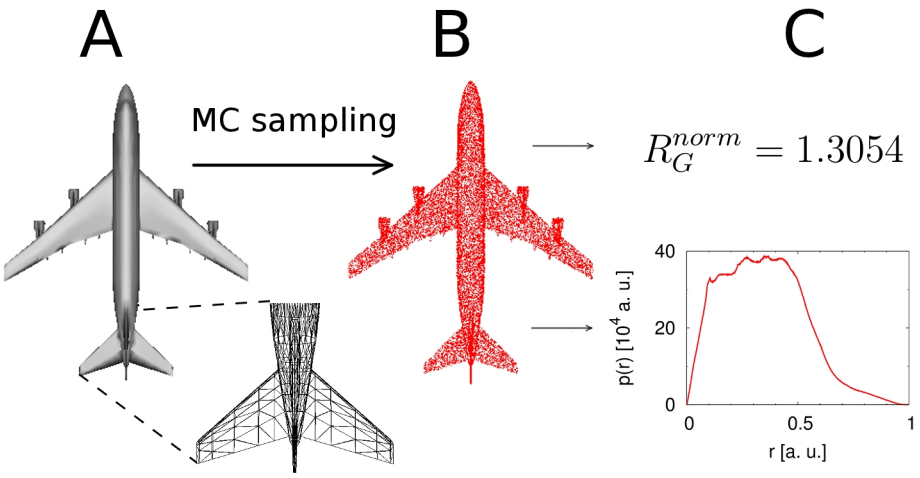


Figura 3.1: (A) Modelo 3D com malha triangular (em destaque). (B) Conjunto de N pontos uniformemente amostrados na superfície do modelo tridimensional. (C) Raio de giro normalizado (R_G^{norm}) e $p(r)$ como característica geométrica do modelo 3D.

Devido a não homogeneidade da disposição espacial dos vértices, cada objeto é reduzido a um conjunto de N pontos com densidade uniforme amostrados pela técnica de Monte Carlo (Figura 3.1B). Para fins de padronização, uma amostragem de 20000 pontos foi realizada ao longo da superfície onde cada triângulo da malha contribui com um certo número de pontos proporcional a sua área.

3.2.4 – Métricas de reconhecimento de padrão baseadas em SAXS

Utilizou-se R_G e $p(r)$ como assinaturas geométricas de um objeto devido a interpretação destas propriedades ser mais direta que a da $I(q)$. A discussão sobre a recuperação da informação

tridimensional contida na $p(r)$ e no R_G é avaliada tendo em vista um experimento computacional por meio da análise de *rankings* descrito na próxima seção. A seguir, as métricas de reconhecimento de forma tridimensional são apresentadas.

Analogamente ao caso de proteínas em solução submetidas ao SAXS, considera-se cada ponto i na superfície de um objeto, como centros espalhadores de raios X com mesmo fator de forma $f(q) = 1$. Logo, para cada objeto, a função PDDF pode ser escrita como:

$$p(r) = \sum_{i=1}^{N} \sum_{j=1}^{N} \delta(r - d_{ij}), \tag{3.1}$$

onde $N = 20000$ são os pontos amostrados por Monte Carlo na superfície do objeto.

Assim, o grau de similaridade entre dois objetos tridimensionais pode ser avaliado através da comparação entre suas funções $p(r)$, através da métrica:

$$\chi^2 = \sum \left[p_{target}(r_i) - p_{template}(r_i) \right]^2 . \tag{3.2}$$

As curvas $p_{target}(r)$ e $p_{template}(r)$ devem ser normalizadas a um fator comum antes desta comparação, pois se assim não o fosse, um objeto poderia ser considerado de classe diferente a de um outro idêntico se este último diferisse do primeiro apenas por um fator de escala. Dessa forma, é arbitrado $\int_0^\infty p(r)dr = p(r_{max}) = 100$. É importante salientar que essa normalização não ocorre com proteínas, onde ainda que formas semelhantes entre duas proteínas sejam verificadas em uma dada escala, serão inevitavelmente consideradas como proteínas diferentes já que o tamanho global é um critério determinante para este tipo de objeto.

O raio de giro também pode ser calculado a partir da distribuição de pontos amostrados, pela expressão:

$$R_G^2 = \frac{1}{N} \sum_{i=1}^{N} (r_i - R_C)^2, \tag{3.3}$$

onde r_i é a coordenada (x_i, y_i, z_i) de um ponto na superfície e R_C é a coordenada do centro geométrico[1] da distribuição.

[1] Se considerássemos um objeto com uma distribuição de massa não homogênea, R_C estaria relacionado ao centro de massa. Esta distinção é importante já que, em geral, o centro geométrico e o centro de massa de um objeto não coincidem.

Da mesma forma que a normalização é imprescindível para o cálculo da métrica envolvendo a $p(r)$, deve-se também normalizar o raio de giro. Isto é feito dividindo o valor calculado pela Equação 3.3 pelo valor do raio de giro de uma casca esférica de mesma área que o objeto 3D em questão, extraindo a raiz quadrada:

$$R_G^{norm} = R_G\sqrt{\frac{4\pi}{A}}, \tag{3.4}$$

com R_G representando o raio de giro do objeto 3D e A, a área de sua superfície.

Consecutivamente, define-se a métrica envolvendo o raio de giro na forma

$$\chi^2 = [R_G^{norm}(target) - R_G^{norm}(template)]^2\,. \tag{3.5}$$

3.2.5 – Experimento computacional

Dos 1814 objetos presentes no banco de dados, 161 foram selecionados aleatoriamente como sendo alvos, um de cada classe. Para um determinado alvo (*target*), sua $p(r)$ assim como seu R_G^{norm} são comparados às respectivas $p(r)$ e R_G^{norm} de todos os elementos restantes do banco de dados (*templates*), utilizando as métricas descritas pela Equação 3.2 e pela Equação 3.5. De posse dos valores de χ^2 de cada métrica, é feito o ranqueamento, do menor para o maior valor, onde espera-se que as primeiras posições do *ranking* sejam ocupadas por *templates* com forma similar ao *target* e que também pertençam à mesma classe dele. Como o número de elementos pode variar de uma classe para outra, a razão (número de elementos da classe do alvo) / (número total de elementos do banco) deve ser mantida constante a fim da avaliação conjunta entre todos os alvos ser a mais justa possível.

Adicionalmente, pode-se aferir qual das métricas é mais efetiva no reconhecimento da forma do *target* dentre as estruturas do banco de dados. Portanto, é de se esperar que a $p(r)$ seja mais eficiente que o raio de giro no sucesso do ranqueamento, já que o primeira representa a distribuição de distâncias do objeto e o segundo representa apenas um número associado a esta distribuição. A Figura 3.2 mostra a ideia geral do experimento computacional.

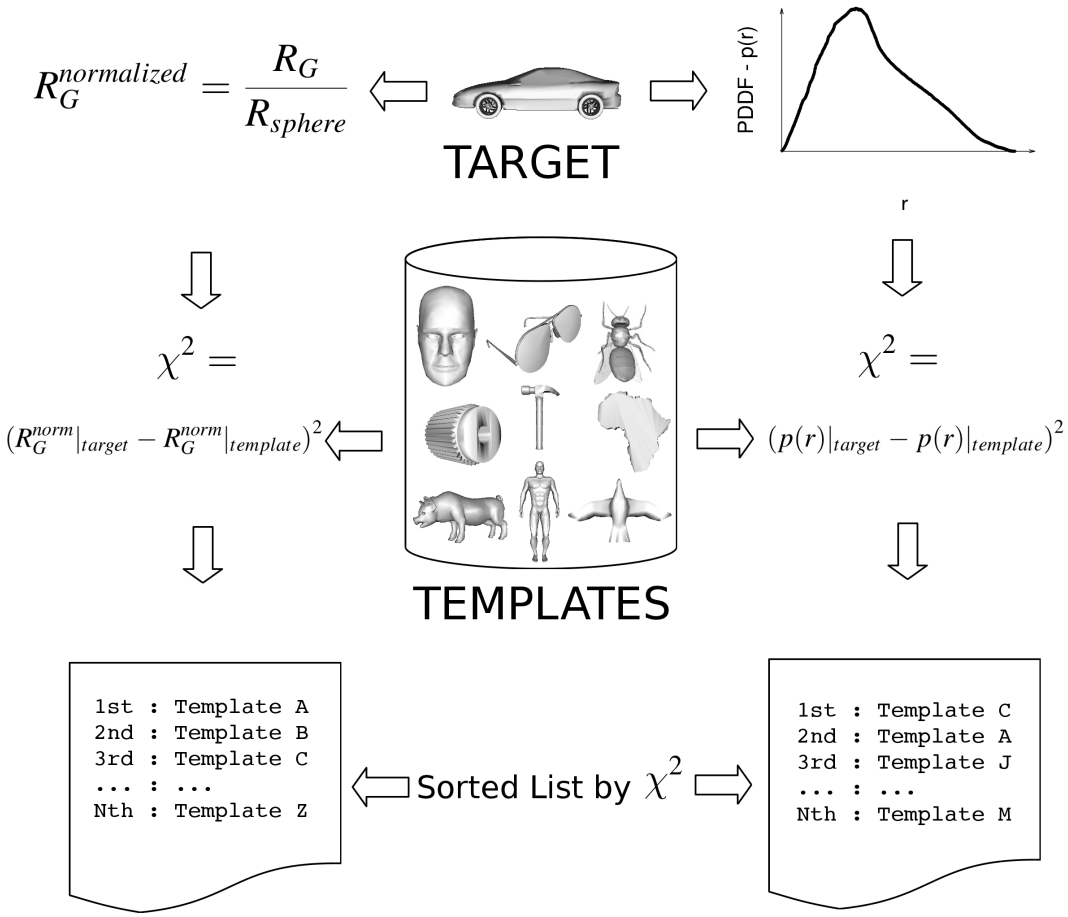


Figura 3.2: Fluxograma do Experimento Computacional para o reconhecimento de formas tridimensionais.

3.2.6 – Avaliação dos resultados de ranqueamento

Após realizado o ranqueamento dos *templates* segundo a métrica baseada na $p(r)$ ou no R_G, dá-se início à quantificação dos resultados levando em conta o posicionamento no *ranking* e a classe a que cada *template* pertence.

O caso ideal seria encontrar todos os *templates* pertencentes a mesma classe do *target* dispostos em sequência nas primeiras posições do *ranking*. Já no outro caso extremo, *templates*

da mesma classe do *target* estariam distribuídos aleatoriamente ao longo do *ranking* devido a imperfeição da métrica utilizada.

Este tipo de avaliação pode ser realizada em um diagrama *Precison vs Recall* [77] que possui a seguinte interpretação: o eixo *Precision* mensura a fração de elementos reconhecidos ao longo do *ranking* e o eixo *Recall* mensura a razão entre os elementos já ordenados pelo total a ser reconhecido.

Para maior entendimento e clareza, considere o seguinte exemplo: suponha que para um determinado alvo, existam 10 elementos de mesma classe dispersos em um banco de dados com mais 90 objetos de classes diferentes, em um total de 100 objetos. Se uma métrica de reconhecimento de padrão chega a ser completamente aleatória, a probabilidade de se obter um objeto de mesma classe do *target* é de $10\% = 10/100$. Logo, o diagrama *Precison vs Recall* exibiria uma reta horizontal com *Precision* igual a 10%. Agora, se os *templates* de mesma classe do alvo estivessem sistematicamente deslocados em direção às posições iniciais do *ranking*, o valor de *Precision* iniciaria próximo a 100% e à medida que um *template* em determinada posição do *ranking* não correspondesse a classe do *target*, a precisão cairia, sem no entanto aumentar o valor de *Recall* que só iria aumentar quando um próximo *template* da mesma classe do *target* fosse encontrado ao longo do *ranking*. Diante disto, dois métodos podem ser diretamente comparados através da área abaixo da curva em um diagrama *Precison vs Recall*. Aquele método que apresentar maior área no diagrama é considerado o mais preciso.

3.3 – Resultados e discussão

A Figura 3.3 ilustra alguns resultados do experimento computacional para 3 casos dos 161 analisados.

O *target* indicado na Figura 3.3A é um típico caso de sucesso, onde claramente observa-se que os quatro primeiros *templates* pertencem à mesma classe do *target*, isto é, um esqueleto humano. Ao lado, também podemos observar o relativo acordo da forma das distribuições $p(r)$ geradas por este tipo de objeto 3D refletindo uma anisometria aproximadamente cilíndrica do *target*, devido a $p(r)$ qualitativamente apresentar um máximo global acentuadamente deslocado para a esquerda, que é o caso esperado para uma distribuição de distâncias de um cilindro.

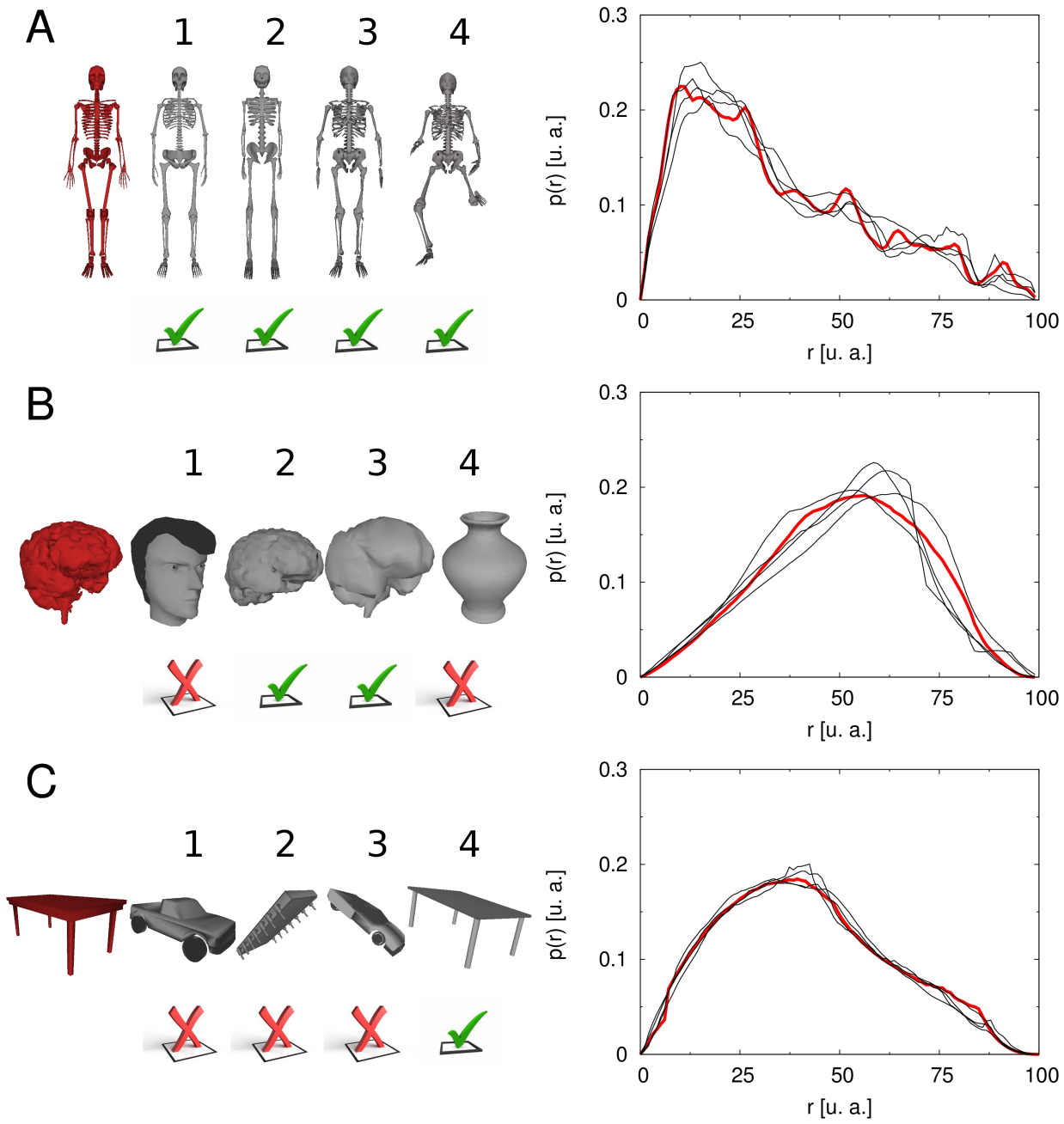


Figura 3.3: Exemplos de reconhecimento de padrão de forma de objetos tridimensionais pela métrica baseada na $p(r)$. Os *targets* são indicados na cor vinho enquanto que o *templates* em cinza. Os números acima dos *templates* indicam a posição ocupada após o ranqueamento. (A) Caso de sucesso. (B) Caso intermediário. (C) Caso de insucesso. As respectivas curvas $p(r)$ também são mostradas para cada caso.

Já o caso apresentado na Figura 3.3B, mostra que embora o método reconheça objetos da mesma classe, alguns falsos positivos também estão presentes, como é o caso do "objeto cabeça" na primeira posição e o "objeto vaso" na quarta. Lembrando que os objetos possuem apenas pontos ao longo de suas superfícies, a explicação para o relativo insucesso é automaticamente explicada pela característica geométrica em comum desses objetos de serem arredondados e ocos por dentro refletindo uma $p(r)$ típica de uma esfera oca com um máximo global deslocado para direita.

Interessantemente, o caso mostrado na Figura 3.3C apresenta o melhor acordo qualitativo entre as $p(r)$ ranqueadas. Entretanto, este caso é o que apresenta um maior número de falsos positivos nas primeiras posições do *ranking* em relação aos outros casos já discutidos. O objeto da classe mesa pode ter sido confundido com os outros por estes também apresentarem características geométricas de baixa resolução em comum, isto é, uma extensa superfície plana conectada às pequenas hastes como é o caso do segundo *template* (*microchip*) e do primeiro e terceiro *templates*, onde as rodas dos objetos do classe carro fariam o papel das hastes, ou mais precisamente, dos pés da mesa.

O resultado médio para os 161 *targets* analisados é mostrado na Figura 3.4. Como era esperado, o método de reconhecimento segundo a métrica da função de distribuição de pares (PDDF) ou simplesmente $p(r)$ é mais preciso que aquele que usa apenas a comparação dos raios de giro normalizados entre dois objetos 3D, pois possui maior área sob a curva. Este diagrama mostra que para a métrica baseada na $p(r)$, cerca de metade dos *templates* (*Recall*=50%) de mesma classe que o *targets* são reconhecidos com um pouco mais de 40% de precisão e que todos os *templates* possíveis de mesma classe (*Recall*=100%) são reconhecidos com 15% de precisão indicando que existem muitos objetos no banco de dados com uma similar distribuição de distâncias porém classificados como objetos de outra classe, contribuindo para a precisão cair. Como um exemplo, um objeto da classe vassoura é facilmente confundida com um objeto da classe taco de baseball devido à sua característica cilíndrica peculiar.

Em geral, ambas as métricas de reconhecimento de padrão estrutural inspiradas em SAXS, baseadas na distribuição radial de pares $p(r)$ ou no raio de giro R_G conseguem superar uma predição puramente aleatória. Quando cerca de 10% dos *templates* são localizados ($Recall = 10\%$) ao longo do *ranking*, essa superação é da ordem de 80% ($Precision$) em relação a uma predição aleatória.

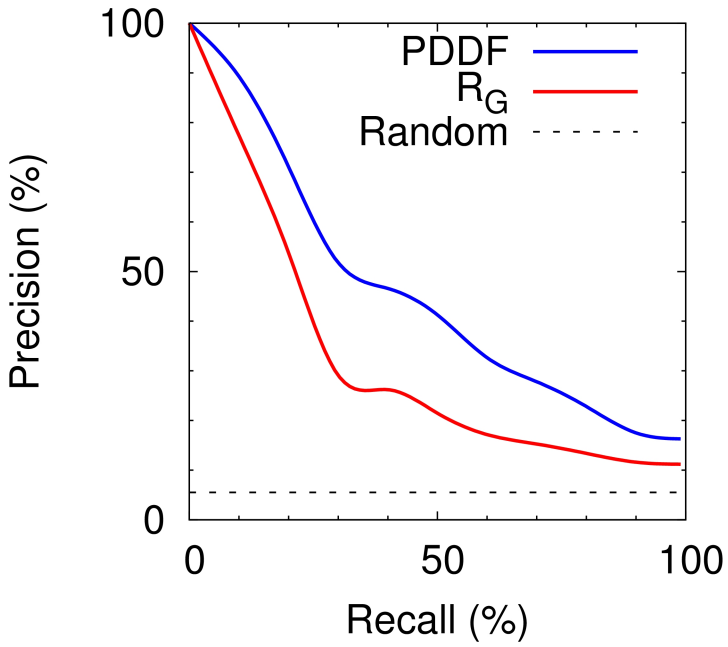


Figura 3.4: Resultado global do experimento computacional de reconhecimento de padrão de 161 objetos 3D de classes distintas. A linha tracejada mostra o valor esperado para um método de reconhecimento aleatório. Os resultados para as métricas baseadas na $p(r)$ (PDDF) e no R_G são mostrados em azul e vermelho, respectivamente.

Estes resultados também confirmam a suspeita inicial de que conforme mais informação estrutural disponível — como o caso da PDDF em relação ao R_G — melhor é o desempenho do reconhecimento da forma de um objeto 3D.

3.4 – Conclusão

A partir dos resultados obtidos, foi possível verificar que, embora conceitualmente o conteúdo informacional de uma curva unidimensional de SAXS pode, em certo grau, discriminar quantitativamente estruturas tridimensionais e ser bastante útil na procura por objetos com formas similares, este procedimento traz uma limitação quando se deseja obter a medida de similaridade da classe de objetos e não apenas da forma geométrica. Isso traz uma forte ponderação indicando que a classe de um objeto não é fruto apenas de sua forma mas também de uma classificação, muitas

vezes subjetiva.

No caso de proteínas, diversas classificações são possíveis. Como um exemplo, o banco de dados SCOP [78] possui uma classificação tradicional baseada no conteúdo da estrutura terciária de uma proteína, isto é, proteínas são agrupadas em quatro classes do tipo: α, β, $\alpha + \beta$ e α/β dependendo do arranjo tridimensional das estruturas secundárias do tipo hélice-α e fita-β. Baseando-se apenas nessas quatro classes e usando uma métrica baseada no raio de giro, Lobanov *et al* [79] e Ivankov *et al* [80] conseguiram correlacionar o grau de compacidade de proteínas frente a essas classes, mostrando que proteínas do tipo puramente α são menos compactas que as do tipo puramente β, sendo os casos intermediários uma combinação, isto é, $\alpha+\beta$ e α/β. Uma outra possível analogia com o termo "classe" empregado neste capítulo, seria a alusão ao termo "homologia", onde se leva em consideração a sequência de aminoácidos e considerações sobre ancestralidade comum a elas.

Certamente, estas questões levantadas são pontos-chave na discussão da aplicação destas técnicas no reconhecimento de enovelamentos de proteínas. É nesta direção que encaminhamos o Capítulo 4 que tratará da combinação da informação relativa a estrutura primária e terciária de uma proteína como instrumentos de classificação do enovelamento proteico.

Capítulo 4

Reconhecimento de enovelamento de proteínas assistido por SAXS

Neste capítulo, são apresentados os principais resultados alcançados durante um período de doutorado sanduíche no *Center for Computational Medicine & Bioinformatics* da *The University of Michigan*, Estados Unidos, e supervisionado pelo Prof. Dr. Yang Zhang. Estes resultados podem ser conferidos no artigo *Improving protein template recognition by using small-angle x-ray scattering profiles* [50] e também na página `http://zhanglab.ccmb.med.umich.edu/SAXSTER/` do servidor SAXSTER, frutos deste projeto.

O objetivo deste trabalho foi o desenvolvimento de um programa, denominado SAXSTER, que tem a habilidade de gerar modelos estruturais mais prováveis para uma proteína-alvo a partir de alinhamentos ótimos obtidos por *threading* e de estruturas similares localizadas no *Protein Data Bank* com o auxílio de dados experimentais obtidos pela técnica de espalhamento de raios X a baixos ângulos.

As diversas etapas necessárias ao desenvolvimento e análise de desempenho do programa estão descritas ao longo deste capítulo.

4.1 – Introdução

Apesar do considerável progresso que tem sido feito na predição do enovelamento de proteínas, a abordagem baseada em proteínas-molde[1] (TBM, *Template-Based Modeling*) é, atualmente, a técnica mais robusta e confiável [81, 82]. Esta abordagem utiliza estruturas conhecidas de proteínas homólogas ou mesmo de baixa identidade sequencial, no intuito de usá-las como guias no processo de modelagem.

A etapa crítica da TBM é a identificação de proteínas depositadas no banco de dados PDB [7] cujas estruturas apresentem alta similaridade com a estrutura da proteína a qual se deseja conhecer. *Threading* [83, 84] é, portanto, a técnica computacional pela qual, apenas a partir da sequência de aminoácidos da proteína-alvo, estruturas tridimensionais candidatas são identificadas dentre uma grande variedade disponível no banco de dados.

Os algoritmos atuais de *threading* funcionam bem na tarefa de reconhecimento de estruturas que têm uma razoável relação evolutiva com a proteína-alvo. Todavia, quando uma proteína é carente de proteínas-molde, os algoritmos apesar de eventualmente identificá-los, muitas vezes falham durante a classificação relativa entre as demais estruturas, consecutivamente fracassam em colocá-los ao longo das primeiras posições do *ranking*, o que degrada significativamente a predição da estrutura proteica desejada.

Para contornar este problema, os pesquisadores têm desenvolvido uma variedade de métodos para explorar os dados experimentais disponíveis, na tentativa de utilizá-los como vínculos no processo de modelagem estrutural. Dentre as técnicas mais difundidas, onde estes métodos são aplicados, encontram-se a Cristalografia por Difração de Raios X [85] e a Ressonância Magnética Nuclear (RMN) [86, 87, 88].

Comparada com a cristalografia de raios X, a técnica de espalhamento de raios X a baixos ângulos (SAXS, *Small-Angle X-ray Scattering*) é vantajosa à medida em que possibilita o estudo das proteínas em ambientes próximos as condições fisiológicas e não requer a obtenção de cristais.

[1] Tradução livre para o termo em inglês *template protein*. Advertimos o leitor no emprego do termo "proteínas-molde" que é adotado aqui para não haver confusão entre os termos "modelo" e "molde" que possuem significados distintos. O termo "modelo" refere-se à estrutura final proposta para a proteína-alvo e "molde" refere-se à estrutura de outra proteína que pode ser utilizada como ponto de partida na modelagem da proteína-alvo.

No entanto, SAXS não tem recebido muita atenção no campo da bioinformática dentro do campo de predição de enovelamentos, principalmente devido à baixa resolução dos dados (10-50 Å) [10]. Ao contrário de cristalografia e RMN, que especificam as coordenadas atômicas, SAXS fornece apenas informação da forma da molécula de proteína sem nenhuma atribuição específica das posições atômicas. A vantagem de SAXS consiste na capacidade da análise estrutural em solução, que pode diferir da correspondente estrutura cristalina, possibilitando a verificação de possíveis relaxamentos de domínios ou de *loops* levando a estrutura à assumir conformações diferenciadas. Apesar da baixa resolução, SAXS tem se mostrado bastante útil como vínculo experimental na detecção de enovelamentos de proteínas consideradas de difícil predição computacional.

4.1.1 – Objetivos

Neste trabalho, a estratégia de TBM é aplicada e combinada aos dados de SAXS e às técnicas de predição de enovelamentos, com o objetivo de melhorar o ranqueamento das proteínas-molde identificadas pela técnica computacional *threading*.

Esta possibilidade foi parcialmente explorada por Zheng e Doniach [89, 90], na qual um algoritmo baseado em *threading* — sem a inserção de *gaps* ao longo dos alinhamentos proteína-alvo e proteína-molde — foi aplicado a um pequeno conjunto de proteínas-teste como uma forma de demonstrar a viabilidade da abordagem. Uma vez que os alinhamentos ótimos quase sempre possuem *gaps*, o resultado do trabalho pioneiro de Zheng e Doniach não pôde ser utilizado para fins práticos.

A partir da viabilidade do conceito proposto por Zheng e Doniach, o principal objetivo neste presente trabalho foi desenvolver um programa, denominado SAXSTER, que tem a habilidade de gerar modelos estruturais mais prováveis para uma proteína-alvo a partir de alinhamentos ótimos obtidos por *threading* e de estruturas similares identificadas no banco de dados PDB com o auxílio de dados de SAXS. Além disto, SAXSTER é disponibilizado gratuitamente através da internet, permitindo que usuários da técnica de SAXS possam ser favorecidos pelo trabalho aqui realizado.

4.2 – Metodologia

4.2.1 – A ideia fundamental de SAXSTER

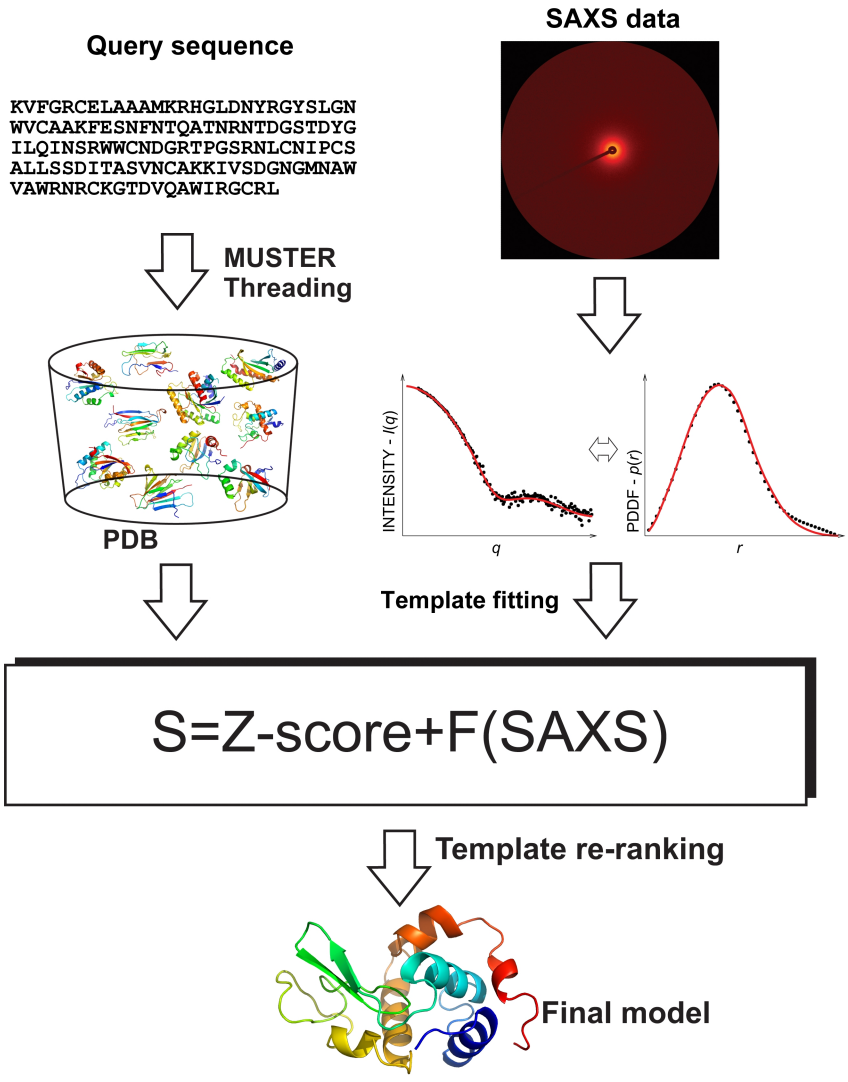


Figura 4.1: Fluxograma do programa SAXSTER. Os dados de entrada são a sequência de aminoácidos e a respectiva curva de SAXS que pode ser tanto a $I(q)$ quanto a $p(r)$. Um modelo estrutural completo é disponibilizado como saída.

Em linhas gerais, o programa SAXSTER (*SAXS-assisted multi-source ThreadER*) tem a habili-

dade de gerar modelos estruturais mais prováveis para uma proteína-alvo a partir de alinhamentos ótimos obtidos por *threading* e de estruturas similares identificadas em um banco de dados com o auxílio de dados de SAXS. A curva experimental de SAXS deve ter sido obtida para uma amostra de proteína monomérica e pode ser fornecida a $I(q)$ ou $p(r)$ como dados de entrada juntamente com a respectiva sequência de aminoácidos. A partir dos dados de entrada, é realizada uma busca no banco de dados PDB [7] utilizando os vínculos experimentais para que a estrutura da proteína-alvo possa ser predita.

Levando-se em conta a qualidade do alinhamento *threading* a partir de uma proteína-molde, assim como pelo respectivo acordo com o dado experimental, um modelo computacional completo para a proteína de interesse é construído e disponibilizado ao usuário. As ideias gerais de funcionamento do programa SAXSTER são mostradas na Figura 4.1.

4.2.2 – *Threading* no reconhecimento de enovelamentos

Um algoritmo *threading* é uma técnica computacional usada em bioinformática capaz de reconhecer proteínas similares não apenas pela identidade sequencial, mas também pelas características estruturais, dada apenas a sequência de aminoácidos da proteína-alvo. O programa *threading* utilizado neste trabalho foi o MUSTER (*Multi-Source Threader*) [91] desenvolvido anteriormente no laboratório do Prof. Dr. Yang Zhang. Nesta etapa, o presente trabalho consistiu em implementá-lo segundo as necessidades da estratégia do SAXSTER possibilitando a construção de modelos tridimensionais a partir dos alinhamentos gerados.

O único dado de entrada do MUSTER é a sequência de aminoácidos da proteína de interesse, tendo como resultado final uma lista de proteínas-molde em ordem de importância. Este ordenamento é realizado por uma função escore (Equação 4.1) que utiliza diversas fontes de informação tais como: perfis sequenciais, predição de estrutura secundária, acessibilidade ao solvente, ângulos de torção da cadeia principal e matriz de hidrofobicidade. A função guia o processo de otimização através da técnica de programação dinâmica, do alinhamento entre os aminoácidos da sequência da proteína-alvo, índice q (*query*) e dos aminoácidos de uma proteína do PDB, índice t (proteína-molde ou *template*). A seguir, cada um dos termos da Equação 4.1 que MUSTER utiliza na produção do alinhamento entre a proteína-alvo e as proteínas-molde são apresentados e discutidos.

$$
\begin{aligned}
S(i,j) = & \sum_{k=1}^{20} \left(P_{c_q}(i,k) + P_{d_q}(i,k) \right) L_t(j,k)/2 \\
& + c_1 \sum_{k=1}^{20} P_{s_t}(j,k) L_q(i,k) \\
& + c_2 \delta \left(s_q(i), s_t(j) \right) \\
& + c_3 \left(1 - 2|SA_q(i) - SA_t(j)| \right) \\
& + c_4 \left(1 - 2|\psi_q(i) - \psi_t(j)| \right) + c_5 \left(1 - 2|\phi_q(i) - \phi_t(j)| \right) \\
& + c_6 M \left(AA_q(i), AA_t(j) \right) \\
& + c_7.
\end{aligned}
\tag{4.1}
$$

Perfis sequenciais obtidos por PSI-BLAST

A primeira linha da Equação 4.1 está relacionada ao perfil obtido para a sequência-alvo a partir do alinhamento multisequencial gerado pelo programa PSI-BLAST [92] utilizando um determinado valor de corte para o parâmetro $E - value$ que mensura a qualidade do alinhamento entre as sequências. O termo $P_{c_q}(i,k)$ refere-se à frequência do k-ésimo aminoácido na posição i no alinhamento multisequencial para todos os alinhamentos obtidos com $E - value < 0.001$, maximizando a possibilidade de identificação de proteínas com grande similaridade sequencial com a proteína-alvo. Já o termo $P_{d_q}(i,k)$ é calculado utilizando-se um corte de $E - value < 1.0$, isto permite que se possa averiguar o grau de similaridade sequencial com as proteínas-molde que estão distantes das primeiras posições indicadas pelo PSI-BLAST, maximizando a possibilidade de detecção de proteínas não homólogas que porventura venham a ter estruturas tridimensionais similares à proteína-alvo. O termo $L_t(j,k)$ é o perfil *log-odd* [92] da proteína-molde que essencialmente compara as probabilidades de alinhamento entre dois aminoácidos relacionados ou não por características bioquímicas e evolutivas.

A segunda parcela de $S(i,j)$, correspondente à segunda linha da equação, é análoga ao termo já descrito e também refere-se ao perfil sequencial obtido por PSI-BLAST. A diferença consiste na análise do perfil da proteína-molde $P_{s_t}(j,k)$ ponderado pelo termo de *log-odd* $L_q(i,k)$ da proteína-alvo.

Estrutura secundária

A terceira linha da Equação 4.1 simplesmente compara o tipo de estrutura secundária (α, β e *coil*) em que se localizam os aminoácidos das proteínas-alvo $s_q(i)$ e de uma determinada proteína-molde $s_t(j)$. Quando as estruturas secundárias são as mesmas, isto é, quando $s_q(i) = s_t(j)$ atribui-se $\delta = 1$, caso contrário $\delta = -1$. Visto que não se conhece *a priori* a estrutura tridimensional da proteína-alvo, a estrutura secundária é predita pelo programa PSIPRED [93] e a estrutura da proteína-molde é determinada pelo programa STRIDE [94].

Acessibilidade ao solvente

O termo que compara a acessibilidade ao solvente entre os aminoácidos alvo e molde é descrito na quarta linha da Equação 4.1. A determinação desta propriedade para a proteína-molde é realizada pelo programa STRIDE, e a predição a partir da sequência de aminoácidos da proteína-alvo é feita por um dos algoritmos do programa MUSTER [91].

Ângulos da cadeia principal

A quinta linha de $S(i, j)$, correspondente ao quarto e quinto termos, mensura a similaridade entre os ângulos ψ e ϕ das cadeias principais das proteínas molde e alvo. Os ângulos da primeira são extraídos pelo programa DSSP [95] e os da segunda são preditos pelo programa ANGLOR [96].

Hidrofobicidade

Analogamente aos termos de estrutura secundária de $S(i, j)$, a penúltima linha da Equação 4.1 mede a correlação entre os aminoácidos das proteínas alvo e molde através de uma matriz de hidrofobicidade de dimensão 20×20 [97]. Isso é feito pois, sabe-se que sequências com padrões de distribuição similares de resíduos hidrofóbicos (V, I, L, F, Y, W, M) são geralmente homólogas estruturais. Assim, se os resíduos comparados pertencem ao grupo hidrofóbico descrito, $M = 1$. Se os resíduos são idênticos mas não são hidrofóbicos então $M = 0.7$, exceto para os resíduos Prolina e Glicina que possuem M também associado à unidade. Atribui-se $M = 0$ para todos os casos omissos.

Pesos entre os termos da função escore

Dado que a função escore $S(i,j)$ possui sete termos mais uma constante (c_7), $S(i,j)$ possui nove parâmetros ao todo considerando mais duas constantes relacionadas a penalidade dos *gaps* no alinhamento relativo a criação de um *gap* (g_O) e a sua extensão (g_E). Esses nove parâmetros foram otimizados [91] usando uma estratégia de *benchmarking* usando 111 proteínas-alvo contra o banco de dados PDB. Os valores dos parâmetros do programa MUSTER utilizados em nossas simulações são: $c_1 = 0,39, c_2 = 0,66, c_3 = 1,60, c_4 = 0,19, c_5 = 0,19, c_6 = 0,31, c_7 = 0,99, g_O = 7,01$ e $g_E = 0,55$.

Exemplo de alinhamento obtido por MUSTER

A Figura 4.2 representa graficamente o impacto que os termos estruturais (estrutura secundária, ângulos da cadeia principal, etc) presentes na função $S(i,j)$ têm na obtenção de um alinhamento quando comparamos a uma função escore que apenas privilegia alinhamentos sequenciais sem nenhum auxílio de informações estruturais da proteína-alvo e da proteína-molde.

Como pode ser percebido na Figura 4.2B à esquerda, quando *threading* é aplicado, o alinhamento gerado reflete um alinhamento estrutural muito semelhante à conformação tridimensional assumida pela proteína-alvo sobre a proteína-molde. Em contrapartida, a Figura 4.2B à direita mostra que a informação proveniente apenas da sequência de aminoácidos não é suficiente para que a estrutura da proteína-alvo seja reconhecida ao longo da estrutura da proteína-molde, fato que é evidenciado por parte do alinhamento construído em um domínio estrutural da proteína-molde que não faz parte da proteína-alvo.

Classificação dos alinhamentos obtidos por MUSTER

MUSTER utiliza a função escore $S(i,j)$ como guia na construção do alinhamento ótimo entre a proteína-alvo e a proteína-molde empregando o algoritmo de Needleman-Wunsch [98]. Esse procedimento é realizado contra todas as 40096 estruturas do banco de dados local do laboratório do Prof. Zhang, que basicamente é uma versão filtrada do *Protein Data Bank* para que proteínas redundantes não sejam contabilizadas mais de uma vez. Na etapa seguinte, as estruturas-molde

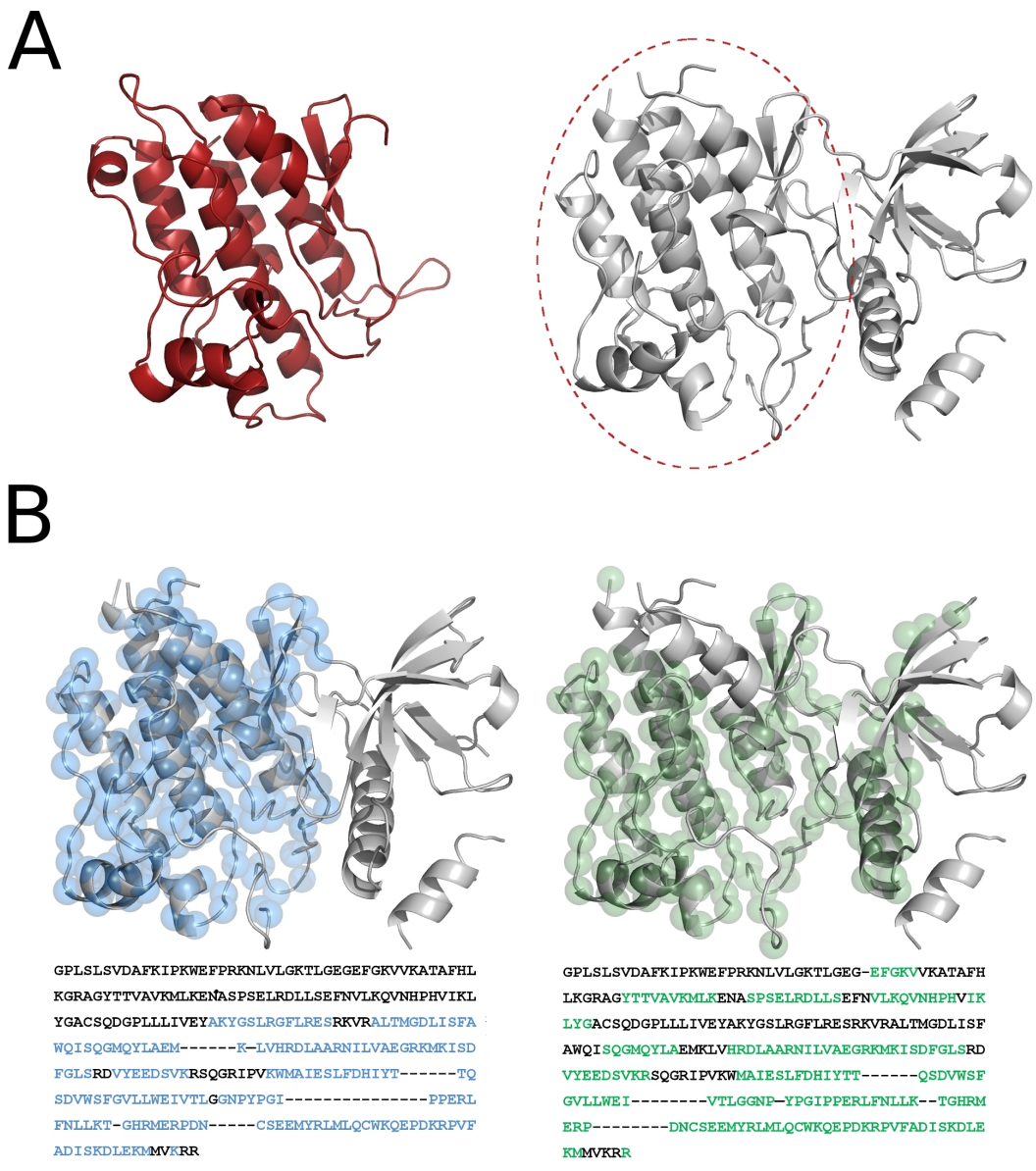


Figura 4.2: Exemplo de alinhamentos dos tipos *threading* e sequencial entre proteína-alvo e proteína-molde. Os alinhamentos estão indicados na base da figura. (A) Estrutura tridimensional da proteína-alvo (vermelho, à esquerda) por hipótese desconhecida e uma proteína-molde (cinza, à direita) obtida no PDB com destaque de um domínio estrutural semelhante à estrutura da proteína-alvo. (B) Resultados dos alinhamentos obtidos por *threading* (à esquerda) e obtidos por alinhamento sequencial (à direita) sobreposto à estrutura da proteína-molde.

são ordenadas e classificadas pela medida estatística *Z-score*, definida por:

$$Z - score = \frac{S - \langle S \rangle}{\sqrt{\langle S^2 \rangle - \langle S \rangle^2}}, \tag{4.2}$$

onde $S = 1/N_{ali} \sum_i^{N_{ali}} S(i,i)$ é a soma dos escores originais ao longo dos N_{ali} pares de resíduos alinhados. De acordo com resultados prévios, quando $Z - score > 7,5$, 98% das proteínas-molde possuem estrutura terciária com o enovelamento muito próximo ao da proteína-alvo enquanto que $Z - score < 7,5$, apenas $5,3\%$ dos moldes terão topologia correta. Assim, categorizamos uma proteína-alvo como de fácil ou difícil predição, de acordo com o *Z-score* da proteína-molde ocupante da primeira posição do *ranking*:

- $Z - score > 7,5$: proteína-alvo fácil;
- $Z - score \leq 7,5$: proteína-alvo difícil.

4.2.3 – Construção de modelos a partir de alinhamentos *threading*

Como visto na seção anterior, modelos *threading* quase sempre contêm *gaps*. Uma vez que perfis de SAXS do tipo $I(q)$ ou $p(r)$ são normalmente extraídos a partir de modelos completos, isto é, de modelos que possuem todos os aminoácidos da proteína de interesse, tentamos três métodos diferentes para rapidamente construir um modelo completo do traçado dos carbonos alfa (C_α) da proteína-alvo a partir dos respectivos alinhamentos *threading* tratando os *gaps* por meio de remoções e inserções de resíduos de aminoácidos.

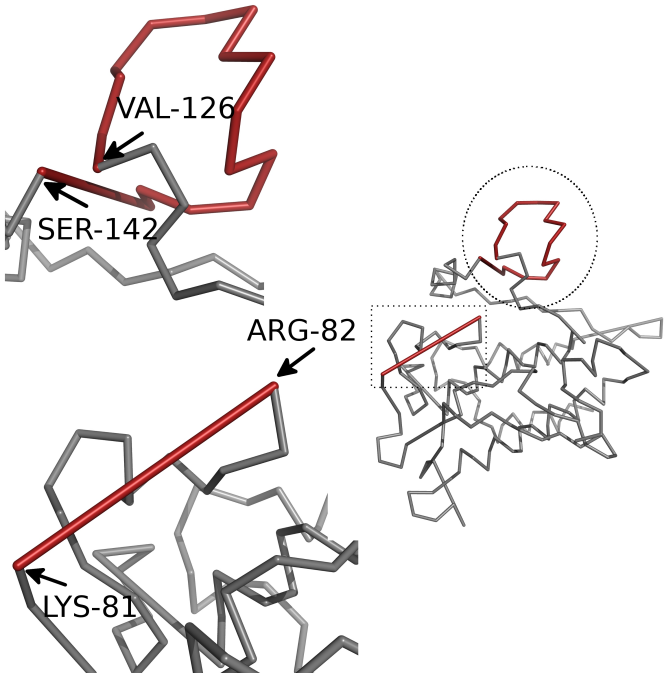
Na primeira abordagem, as coordenadas dos resíduos alinhados são copiadas a partir da estrutura da proteína-molde e são mantidas estáticas no decorrer da simulação. Nesta cópia, realiza-se a substituição da identidade dos resíduos de aminoácidos da proteína-molde para os da proteína-alvo segundo a posição no alinhamento. As regiões não alinhadas são construídas por *random walk* com passo de $3,8\text{Å}$ entre C_α-C_α guiado por restrições de forma a reduzir os choques entre elementos estruturais e garantir a convergência do preenchimento de *gaps* localizados nas regiões entre N-terminal e C-terminal. Durante o *random walk*, qualquer tentativa que produza uma distância $< 3.8\text{Å}$ entre qualquer C_α é descartada. A ligação virtual entre três C_α consecutivos são restritas a ângulos de ligação no intervalo entre 65Å - 165Å para refletir os valores observados no gráfico de Ramachandran. Para orientar o *random walk* em direção ao seu ponto final, só

PROTEÍNA-ALVO:

```
----------------------------------------------------------------------------
-------------------------HEMGSLYDYLQLT----TLDTVSCLRIVLSIASGLAHLHIEIFGTQGKPAIA
HRDLKSKNILVKKNGQCCIADLGLA--VMHRVGTK--------RYMAPEVLDETIQVDCFDSYKRVDIWAFGLVLWEV
ARR-MVSNGIVEDYKPPFYDVVPNDPSFEDMRKVVCVDQQRPNIPNRWFSDPTLTSLAKLMKECWYQNPSARLTALRI
KKTLTKI--D--

PROTEÍNA-MOLDE:
GPLSLSVDAFKIPKWEFPRKNLVLGKTLGEGEFGKVVKATAFHLKGRAGYTTVAVKMLKENASPSELRDLLSEFNVLK
QVNHPHVIKLYGACSQDGPLLLIVEYAKYGSLRGFLRESRKVRALTMGDLISFAWQISQGMQYLAEM------K--LV
HRDLAARNILVAEGRKMKISDFGLSRDVYEEDSVKRSQGRIPVKWMAIESLFDHIYT------TQSDVWSFGVLLWEI
VTLGGNPYPGI---------------PPERLFNLLKT-GHRMERPDN-----CSEEMYRLMLQCWKQEPDKRPVFADI
SKDLEKMMVKRR
```

(a)



(b)

Figura 4.3: Construção de um modelo tridimensional para a proteína-alvo a partir de um alinhamento *threading* e a estrutura de uma proteína-molde utilizando *random walks*. (a) Alinhamento *threading*. (b) Construção de um *loop* de 15 resíduos (destacado em vermelho, à esquerda, acima.) a partir de *random walks* para ligar o resíduo VAL-126 ao resíduo SER-142 pertencentes à sequência da proteína-alvo devido a um *gap* na sequência da proteína-molde nesta região do alinhamento. Também é mostrado a ligação de grande comprimento entre os resíduos LYS-81 e ARG-82 devido a um *gap* na sequência da proteína-alvo. As regiões do alinhamento destacadas na estrutura também estão destacadas por uma elipse azul na Figura (a).

passos com $l < 3,54n$ Å são permitidos, onde l é a distância entre o corrente C_α e o primeiro C_α do próximo fragmento alinhado na proteína-molde. O parâmetro n é o número de C_α restante para o preenchimento total do *gap*. Já os *gaps* localizados nas regiões do N-terminal e C-terminal não possuem restrições de chegada e portanto, representam o tipo mais comum de erros na construção de um modelo. Por possuir uma das extremidades livre, o preenchimento desse tipo de *gap* pode, dependendo da sua extensão, ser indefinidamente afastado da região enovelada que o alinhamento *threading* identificou, ocasionando um modelamento severamente artificial nestas regiões.

O segundo método emprega o programa MODELLER [99] na construção de modelos completos a partir dos alinhamentos *threading* usando a classe *automodel*, no modo padrão. Desta maneira, MODELLER não só constrói modelos completos preenchendo os *gaps*, mas também relaxa a estrutura como um todo a partir de campos de força tradicionais como o CHARMM. Os modelos construídos desta forma satisfazem as restrições dos alinhamentos e geralmente são muito próximos da estrutura da proteína-molde.

A terceira abordagem utiliza o componente de modelagem de cauda e de *loop* da suíte de programas I-TASSER [100, 101] que possui uma abordagem muito similar aos procedimentos de preenchimento de *gaps* descritos nesta seção. A diferença consiste na aplicação de um campo de força empírico próprio baseado em estatísticas de banco de dados e estereoquímica permitindo I-TASSER guiar o processo de preenchimento dos *gaps* pela técnica de Monte Carlo com troca entre réplicas. Deste modo, I-TASSER mantém a estrutura alinhada original e modela mais precisamente os *loops* em relação à abordagem de *random walks*. Rodamos I-TASSER utilizando uma única réplica com 200 passos o que conduz uma simulação menor que 1 min para um alinhamento típico.

A Figura 4.3 exemplifica a abordagem da construção de um modelo por *random walks* mas também serve como discussão para as dificuldades encontradas por MODELLER e I-TASSER no modelamento a partir de um alinhamento, conforme discussão a seguir. Um *gap* na sequência de aminoácidos da proteína-alvo denota que os respectivos resíduos da proteína-molde devem ser simplesmente descartados, como é o caso da grande região do N-terminal da proteína-molde na Figura 4.3 (a). Analogamente, todos os resíduos da proteína-molde indicados em vermelhos devem ser suprimidos do modelo final por estarem relacionados aos *gaps* na sequência da proteína-alvo. É interessante notar que este tipo de remoção de resíduos na região central, inevitavelmente,

faz com que dois resíduos da proteína-molde que são espacialmente distantes sejam conectados, dada a prescrição do alinhamento, gerando efeitos conhecidos como *big bound* (Figura 4.3 (b) (à esquerda, abaixo)) que poderiam ser desfeitos apenas por uma simulação computacional que tivesse o objetivo de relaxar a estrutura e não preservar o alinhamento original para que este possa ser identificado, como os programas MODELLER e I-TASSER original fazem.

Outra dificuldade conhecida refere-se ao preenchimento de resíduos relacionados aos *gaps* ao longo da sequência da proteína-molde. Esse tipo de *gap* é alinhado a resíduos da proteína-alvo que devem ser inseridos em uma região espacial ocupada por dois resíduos covalentemente conectados na estrutura da proteína-molde fornecendo dificuldades para a convergência de qualquer que seja o algoritmo utilizado para a construção do *loop*. A Figura 4.3 (b) (à esquerda, acima) ilustra essa situação.

Eventuais erros de modelagem desta natureza não são perceptíveis do ponto de vista da baixa resolução dos dados de SAXS e portanto, satisfazem o propósito de aplicação que aqui se destina, embora estes conhecidos erros somente são resolvidos por longas simulações com o intuito de modelar a proteína de interesse e não tão somente identificar moldes no PDB. Enfatizamos que o nosso objetivo em construir rapidamente um modelo para a proteína de interesse, através dos métodos descritos, é de sermos capazes de classificá-los de acordo com a similaridade aos correspondentes perfis de SAXS para que proteínas-molde, a serem usadas como ponto de partida, sejam identificadas.

4.2.4 – Simulação de dados de SAXS

Para um determinado modelo estrutural de proteína, simulamos o perfil de intensidade de SAXS de acordo com a equação de Debye [102]

$$I(q) = \sum_{i=1}^{N+W} \sum_{j=1}^{N+W} f_i(q) f_j(q) \frac{sen(qd_{ij})}{qd_{ij}}, \tag{4.3}$$

onde $q = (4\pi sen\theta)/\lambda$ é o vetor de espalhamento cuja unidade é Å^{-1} amostrado até $q_{max} = 0,50\text{Å}^{-1}$ em intervalos igualmente espaçados de $\Delta q = 0,01\text{Å}^{-1}$; λ é o comprimento de onda dos raios X; 2θ é o ângulo de espalhamento; N é o número de átomos do modelo; W é o número de moléculas fictícias de águas inseridas ao modelo e responsáveis em representar a camada de hidratação; d_{ij} é a distância euclidiana entre os átomos i e j; $f(q)$ é o fator de espalhamento

atômico o qual depende do tipo de átomo considerado (H, C, N, O ou S). Todos os hidrogênios são tratados implicitamente e posicionados artificialmente sobre o respectivo átomo pesado no qual está ligado covalentemente. O termo que leva em conta a contribuição do espalhamento do solvente pelo correspondente volume excluído é adicionado ao fator de espalhamento:

$$f_i(q) = f_i^{vacuo}(q) - \rho_{solv} V_i exp\left(-\frac{V_i^{2/3}}{4\pi} q^2\right), \tag{4.4}$$

onde $f_i^{vacuo}(q)$ é a contribuição do espalhamento no vácuo do átomo i e retirado das tabelas internacionais de cristalografia [56] e ρ_{solv} é a densidade eletrônica da água a 20^0C que é mantida fixa em nossas simulações $\rho_{solv} = 0.334e^{-}\text{Å}^{-3}$. O volume excluído V_i do i-ésimo átomo e de um grupo atômico (CH, CH_2, CH_3, NH, NH_2, NH_3 e SH) são retirados de resultados experimentais [103, 49]. Além disto, as simulações incluem um modelo explícito de W moléculas fictícias de água dispostas ao redor da proteína de forma a imitar uma camada de solvatação de aproximadamente 3Å de espessura que é um valor típico determinado experimentalmente [104]. Para este propósito, parte-se de uma rede cúbica de face centrada (CFC), com parâmetro de rede L_{cela} onde cada ponto da rede representa uma molécula de água centrada na coordenada do átomo de oxigênio com dois hidrogênios implícitos. O fator de espalhamento para essa molécula fictícia advém da contribuição no vácuo para os dois átomos O e H. A proteína é inserida no interior da rede CFC e somente moléculas fictícias de água cuja distância a qualquer C_α no intervalo entre $3,5-6,5\text{Å}$ são mantidas na rede e as demais são removidas. Com este procedimento, o único parâmetro livre deste modelo é a densidade da rede que pode ser escrita em termos do tamanho da cela L_{cela} que por sua vez, pode ser representada pelo contraste da camada de hidratação $\delta\rho$, conforme o seguinte desenvolvimento:

A densidade volumétrica de pontos ρ_{CFC} da rede CFC com N_{CFC} pontos e volume V_{CFC} é definida por:

$$\rho_{CFC} = \frac{N_{CFC}}{V_{CFC}} = \frac{4k^3}{L}, \tag{4.5}$$

onde $k = 1, 2, 3, ...$ representa o número de celas unitárias nas direções x, y e z e $L = k \times L_{cela}$ representa a aresta do cubo formado pela rede CFC constituída por k celas. Levando em conta que a rede CFC possui 4 pontos no interior de cada cela convencional (1/8 ponto em cada um dos 8 vértices do cubo + 1/2 ponto em cada uma das 6 faces), cada cela contribui com 4 molé-

culas de água ou em outras palavras, com 40 elétrons por cela. Logo, a densidade de elétrons da rede CFC pode ser ajustada variando o valor do parâmetro L_{cela}: aumentando-o para diminuirmos a densidade e diminuindo-o para aumentarmos a densidade eletrônica. Arbitrariamente, podemos definir a densidade eletrônica da rede CFC como sendo o contraste $\delta\rho$ que a camada de hidratação ρ_{hidra} possui em relação a densidade de elétrons na água líquida ρ_{solv}, pela seguinte expressão:

$$\delta\rho = \rho_{hidra} - \rho_{solv} = \frac{40e^-}{L^3_{cela}}. \tag{4.6}$$

Desta forma, podemos diretamente ajustar o contraste da camada de hidratação alterando o comprimento da aresta da rede CFC.

Devemos atentar para que as moléculas de água dispostas ordenadamente no espaço não contribuam para o surgimento de um padrão de espalhamento característicos em altos ângulos, como picos de difração. Para garantir que estas correlações não surjam no perfil de SAXS, posicionamos aleatoriamente as moléculas de água em torno dos sítios da rede CFC.

Analogamente ao desenvolvimento para a simulação da $I(q)$, a função de distribuição de distância $p(r)$ pode ser escrita na forma da Eq. (4.7):

$$p(r) = \sum_{i=1}^{N+W} \sum_{j=1}^{N+W} f_i(q=0) f_j(q=0) \delta(r - d_{ij}), \tag{4.7}$$

onde δ é uma função que possui valor nulo para todos os casos onde $r \neq d_{ij}$ e 1, caso contrário.

Para o cálculo da $p(r)$, usa-se a discretização na variável r com passo $\Delta r = 1\text{Å}$. É esperado que a função $p(r)$ seja uma função suave [8]. No entanto, em alguns casos é observado um padrão levemente oscilatório que decorre da aproximação de um átomo por um ponto no espaço. Para contornar este problema, implementamos o mesmo algoritmo de suavização da $p(r)$ que é utilizado no programa GASBOR [105] que essencialmente, utiliza uma função com certa largura ao invés da imposição da distribuição $\delta(r - d_{ij})$.

Modelo CG: perfil de SAXS a partir dos C_α

Conforme discutido na Seção 4.2.3, que tratou da construção de modelos a partir de alinhamentos *threading*, os modelos construídos contêm apenas as coordenadas dos C_α dos resíduos

alinhados e dos C_α que foram computacionalmente modelados preenchendo os *gaps* do alinhamento. Isto impossibilita o uso imediato da Equação 4.3 para o cálculo de $I(q)$ e da Equação 4.7 para o cálculo da $p(r)$ a partir desses modelos.

Diante deste obstáculo, o modelo desenvolvido por Yang *et al* [106] foi estendido simplesmente reescrevendo as Equações 4.3 e 4.7 baseando-as na substituição do fator de espalhamento atômico $f(q)$ por um fator de espalhamento efetivo f_{ef} associado a cada resíduo de aminoácido centrado no respectivo C_α:

$$\begin{cases} I(q) = \sum_{i=1}^{N+W} \sum_{j=1}^{N+W} f_{ef}^{i}(q) f_{ef}^{j}(q) \dfrac{sen(qd_{ij})}{qd_{ij}}, \\ p(r) = \sum_{i=1}^{N+W} \sum_{j=1}^{N+W} f_{ef}^{i}(q=0) f_{ef}^{j}(q=0) \delta(r - d_{ij}), \end{cases} \tag{4.8}$$

onde d_{ij} é interpretado como a distância entre o i-ésimo e o j-ésimo componente do sistema formado pelo conjunto dos N carbonos alfa da proteína e as W moléculas fictícias de H_2O. Em contraste com o modelo de Yang *et al*, no entanto, a atual representação para camada de hidratação contém águas arranjadas em uma rede CFC onde o número de moléculas de água por volume é ajustada para imitar o efeito de contraste de densidade eletrônica. Consecutivamente, esta aproximação leva a um cálculo mais rápido que o modelo original de Yang, pois usa menos moléculas de águas ao invés de considerar centenas de moléculas de H_2O fictícias dispostas em torno da proteína, obtidas em uma caixa de simulação pré-equilibrada por dinâmica molecular na densidade desejada. Assim, a abordagem da rede CFC é mais rápida, sem comprometer a precisão, conforme será visto na seção de resultados.

Em síntese, nosso modelo CG (*Coarse-Grained*), estabelecido pela Equação 4.8, calcula a função $p(r)$ do espaço real além do perfil de intensidade de espalhamento $I(q)$ do espaço recíproco o que faz com que a nossa abordagem seja mais flexível para muitas aplicações em especial àquela em que desejamos aplicá-la, como é o caso da obtenção de perfis de SAXS a partir de modelos construídos por alinhamentos tipo *threading*.

Os fatores de espalhamento efetivos f_{ef} no modelo CG são obtidos pela equação:

$$f_{ef}(q) = \left\langle \sum_{i=1}^{n} \sum_{j=1}^{n} f_i(q) f_j(q) \frac{sen(qd_{ij})}{qd_{ij}} \right\rangle^{1/2}, \tag{4.9}$$

onde $\langle \cdots \rangle$ denota a média sobre todos os resíduos de mesmo tipo encontrados em 200 estruturas

proteicas de alta resolução, não homólogas entre si e aleatoriamente selecionadas no PDB. O índice n representa o número de átomos de um determinado aminoácido e $f(q)$ é o fator de espalhamento atômico calculado pela Equação 4.4. Como um resultado desta simplificação, 20 funções $f_{ef}(q)$ são obtidas, cada uma associada a um tipo de resíduo. No caso das moléculas de H_2O, $n = 3$ e $d_{ij} = 0$ com $f(q)$ podendo ser os fatores de espalhamento do hidrogênio e do oxigênio.

Ajuste dos perfis teóricos de SAXS

Os perfis teóricos de SAXS, $I(q)$ e $p(r)$, são calculados a partir de um modelo tridimensional de proteína utilizando o modelo CG. A literatura [104] sugere uma densidade da camada de hidratação (ρ_{hidra}) cerca de 10% maior do que a densidade do solvente (ρ_{solv}). Para a água a 20^0C, $\rho_{solv} = 0,334e^-$ Å^{-3} e consequentemente, $\rho_{hidra} = 1,10 \times \rho_{solv} = 0,367e^-$ Å^{-3}. Assim, o contraste eletrônico definido anteriormente pela Equação 4.6, leva a $\delta\rho \approx 0,03\, e^-$ Å^{-3}.

Portanto, para ajustar uma curva de SAXS produzida pelo modelo CG sobre uma outra curva, experimental ou teórica, uma otimização é feita de modo a encontrar o melhor valor de $\delta\rho$ que minimize o acordo entre os perfis. Isso é feito com a busca de $\delta\rho$ no intervalo compreendido entre [0,00-0,05] e aferindo-se o acordo entre as curvas por uma métrica previamente estabelecida, sendo o χ^2 a métrica mais conhecida em estudos de SAXS:

$$\chi^2 = \frac{1}{n}\sum_{k=1}^{n}\left\{\frac{I_{exp}(q_k) - cI_{CG}(q_k)}{\sigma(q_k)}\right\}^2, \tag{4.10}$$

onde I_{exp} e I_{CG} representam a intensidade de espalhamento experimental e a produzida pelo modelo CG, respectivamente. A constante c é um fator de escala analiticamente calculado de forma a minimizar a expressão χ^2 como um todo, impondo $\partial\chi^2/\partial c = 0$. A função $\sigma(q)$ representa o erro experimental de $I_{exp}(q)$. No caso de uma curva teórica no lugar da experimental, $\sigma(q) = I(q) \times (q + 0,15) \times 0,30$ [51].

Outras formas funcionais além da apresentada pela Equação 4.10 também são possíveis e investigadas neste trabalho. Uma lista com nove métricas baseadas nos dados de SAXS são encontrados na Tabela 4.1. Como um exemplo, destaca-se a métrica baseada na $p(r)$, esquema

IX da mesma tabela:

$$corr\left(p_{exp}(r), p_{CG}(r)\right) = \frac{\sum_i^m \left(p_{exp}(r_i) - \langle p_{exp}(r)\rangle\right)\left(p_{CG}(r_i) - \langle p_{CG}(r)\rangle\right)}{\sqrt{\sum_i^m \left(p_{exp}(r_i) - \langle p_{exp}(r)\rangle\right)^2}\sqrt{\sum_i^m \left(p_{CG}(r_i) - \langle p_{CG}(r)\rangle\right)^2}}, \tag{4.11}$$

que tem por objetivo medir o grau de correlação entre as curvas $p_{exp}(r)$ e $p_{CG}(r)$.

Banco de dados de proteínas

Geralmente, um método computacional preditivo é calibrado e aferido através de um procedimento em larga escala denominado *benchmarking*. Este processo baseia-se em dois conjuntos independentes e aleatórios, neste caso, dois conjuntos de estruturas tridimensionais de proteínas. O primeiro, denominado conjunto de treino, serve para ajustar e calibrar os parâmetros do método e o segundo, conhecido como conjunto de teste é usado para quantificar o desempenho do procedimento como um todo mantendo fixos os parâmetros determinados através do primeiro conjunto.

Devido à escassez de dados experimentais requeridos, a predição em larga escala foi realizada computacionalmente. Os conjuntos de treino e de teste foram selecionados usando a ferramenta PISCES [107] da qual, uma lista com cerca de 7000 cadeias de proteínas foram obtidas, todas com identidade sequencial menor que 30% entre qualquer par, resolução cristalográfica melhor que 3Å e número de aminoácidos no intervalo de [50-400]. Mais especificamente, a partir de um *download* de 7466 estruturas e suas respectivas sequências de aminoácidos obtidas diretamente do banco de dados PDB, filtramos e adequamos as estruturas tridimensionais para que, de forma automática a partir de *scripts* escritos em linguagem Perl, pudéssemos: eliminar rotâmeros de resíduos redundantes presentes em diversas estruturas; renumerar a posição de aminoácidos da estrutura cristalográfica de acordo com a sequência depositada; analisar estruturas enoveladas com ao menos 5% de resíduos em estrutura secundária do tipo hélice-α ou fita-β. Ao final desta prévia catalogação, rodamos MUSTER para cada uma das 7466 cadeias monoméricas de proteínas para que fizéssemos a classificação da dificuldade de predição da estrutura de acordo com o $Z - score$ do MUSTER categorizando a proteína como um alvo fácil ($Z - score > 7,5$) ou difícil ($Z - score \leq 7,5$). De posse dos resultados, separamos aleatoriamente e em proporções semelhantes, dois conjuntos de proteínas:

- Conjunto treino: 341 cadeias monoméricas com 201 (59%) de alvos fáceis e 140 (41%) alvos difíceis;
- Conjunto teste: 412 cadeias monoméricas com 232 (56%) de alvos fáceis e 180 (44%) alvos difíceis.

Combinação de *threading* e SAXS

O propósito de se combinar dados de *threading* com dados experimentais de SAXS é aumentar as chances de seleção de bons moldes estruturais para um dado alvo. Diversas funções de escore F_{SAXS} foram testadas para se aferir o grau de similaridade entre as curvas de SAXS. Contudo, todas elas foram combinadas ao escore do programa MUSTER (F_{MUSTER}) por uma simples combinação linear na forma:

$$F_{SAXSTER}(i,j) = F_{MUSTER}(i,j) + wF_{SAXS}(i,j), \tag{4.12}$$

onde F_{MUSTER} refere-se ao $Z-score$ (Equação 4.2) do alinhamento produzido por MUSTER entre a sequência da proteína-alvo i e a estrutura da proteína-molde j. F_{SAXS} é uma métrica baseada em SAXS podendo ser aquela descrita na Equação 4.10 ou qualquer uma dentre as descritas na Tabela 4.1 e w é um peso que faz o balanço entre os dois escores, cujo valor é otimizado pelo conjunto treino para cada métrica F_{SAXS} utilizada. Como um exemplo, $w = 0,809$ para Esquema IX da Tabela 4.1.

Métrica estrutural

O *Template Modeling score* [68], ou simplesmente TM-score foi escolhido como métrica de similaridade estrutural entre duas proteínas. O TM-score é computado após os alinhamentos estruturais serem realizados e é dado por:

$$TM - score = max\left[\frac{1}{L_{alvo}}\sum_{i}^{L_{ali}}\frac{1}{1+\left(\frac{d_i}{d_0(L_{alvo})}\right)^2}\right], \tag{4.13}$$

onde $d_0(L_{alvo}) = 1,24\sqrt[3]{L_{alvo}-15} - 1,8$ e L_{alvo} e L_{ali} são os comprimentos da sequência alvo e do alinhamento inteiro, respectivamente. O termo d_i representa a distância euclidiana entre os

C_α dos resíduos i entre o alvo e o molde. Uma vez aplicado este escore entre a estrutura nativa e aqueles resíduos alinhados na estrutura da proteína-molde encontramos valores no intervalo [0,1], segundo a seguinte interpretação:

- $TM - score = 1,0$: estruturas idênticas;
- $TM - score \geq 0,5$: estruturas têm a mesma topologia;
- $TM - score < 0,17$: estruturas não compartilham características em comum.

4.3 – Resultados e discussão

4.3.1 – Teste para cinco proteínas com dados experimentais

A elaboração de um conjunto de treino ou de teste baseado em dados reais de SAXS requer a disponibilidade das respectivas estruturas experimentais de alta resolução cujos perfis teóricos para $I(q)$ e $p(r)$ ajustem as curvas experimentais de SAXS. Além desse requerimento, o método descrito neste capítulo é desenvolvido para proteínas monoméricas em solução, o que reduz significativamente os dados disponíveis na literatura. Diante disto, antes de nos concentrarmos nos resultados em larga escala, preparamos um conjunto de teste com 5 proteínas e respectivas curvas experimentais de SAXS coletadas no banco de dados BIOISIS (`http://www.bioisis.net`) [108] e também no arquivo do nosso grupo no Instituto de Química da UNICAMP.

A seguir, os resultados para estas 5 proteínas são apresentados e discutidos.

Validação do modelo CG

Uma das tarefas-chave de SAXSTER é ser capaz de ajustar uma curva experimental a um molde teórico do banco de dados. Consequentemente, se por alguma razão SAXSTER não conseguir ajustar uma curva experimental, esta inabilidade será refletida no desempenho de busca dos moldes no banco de dados.

Em geral, bons ajustes são alcançados com o programa de simulação de dados teóricos de SAXS tanto para a $I(q)$ quanto para a $p(r)$ conforme resultados exemplificados na Figura 4.4 para 5 proteínas monoméricas que possuem estrutura e dados de SAXS disponíveis. A PDDF ou $p(r)$

experimental foi obtida pelo programa GNOM [47], o qual realiza uma transformada de Fourier indireta a partir da curva experimental $I(q)$. Para o modelo CG, as curvas $I(q)$ e $p(r)$ foram obtidas pelas Equações 4.8. Todas as curvas $p(r)$ foram normalizadas para a unidade enquanto que as curvas $I(q)$ estão dispostas em escala relativa para melhor visualização. Os valores de χ^2 mensuram a qualidade dos ajustes que são aferidos pela Equação 4.10, com o valor ótimo do parâmetro $\delta\rho$ definido pela Equação 4.6.

Em geral, observa-se um bom acordo entre as curvas experimentais e as teóricas preditas pelo nosso modelo CG com uma média $\chi^2 = 0,624$ indicando que as diferenças entre estas curvas estão, na maioria dos casos, dentro dos erros experimentais, já que $\chi^2 \leq 1,00$.

Geralmente, o modelo CG ajusta melhor os dados de SAXS de proteínas que possuem estruturas determinadas experimentalmente do que àquelas que possuem estruturas de homólogas, ou geradas a partir de modelagem por homologia. Por exemplo, a Figura 4.4(C e C') mostra os perfis de SAXS da PF1528 de *Pyrococcus furiosus* com dados experimentais de SAXS retirados do banco de dados BIOISIS e estrutura tridimensional modelada por homologia [108]. Logo, a incerteza conformacional do grande N-terminal desta proteína pode ter influenciado em um maior impacto no desacordo entre os perfis teóricos e experimentais com $\chi^2 = 0,81$.

As estruturas da Figura 4.4 D e E possuem múltiplos domínios. A estrutura U2AF65, fator de *Splicing*, foi obtida pela junção de três domínios a partir de estruturas homólogas e a estrutura da proteína albumina do soro humano (HSA, *Human Serum Albumin*) é também, a estrutura de uma homóloga (PDB ID: 1AO6, cadeia A) em relação à proteína submetida ao SAXS, embora com identidade sequencial maior que 90%. A incerteza da orientação dos três domínios cristalográficos do primeiro caso, não impede o razoável acordo observado entre as $I(q)$ teórica e experimental, com $\chi^2 = 0,65$ mas pode ter contribuído para um desacordo mais acentuado no segundo exemplo, com $\chi^2 = 1,32$.

Lisozima é muitas vezes utilizada como um bom padrão de calibração e validação de simulações de SAXS [49, 105, 109, 106]. Em parte, devido à estrutura de alta resolução ser disponível no PDB e cuja curva teórica de SAXS ter um excelente ajuste aos dados experimentais. O ajuste do modelo CG para o monômero de Lisozima possui excelente acordo tanto com a $I(q)$ quanto com a $p(r)$ experimentais como pode ser visto na Figura 4.4(B e B'). O χ^2 calculado com nosso modelo CG é de $0,252$, enquanto que o valor calculado a partir de programas semelhantes como o CRYSOL [49] (versão 2.6) e FoXS [110] (versão 2010), os quais usam todos os átomos da

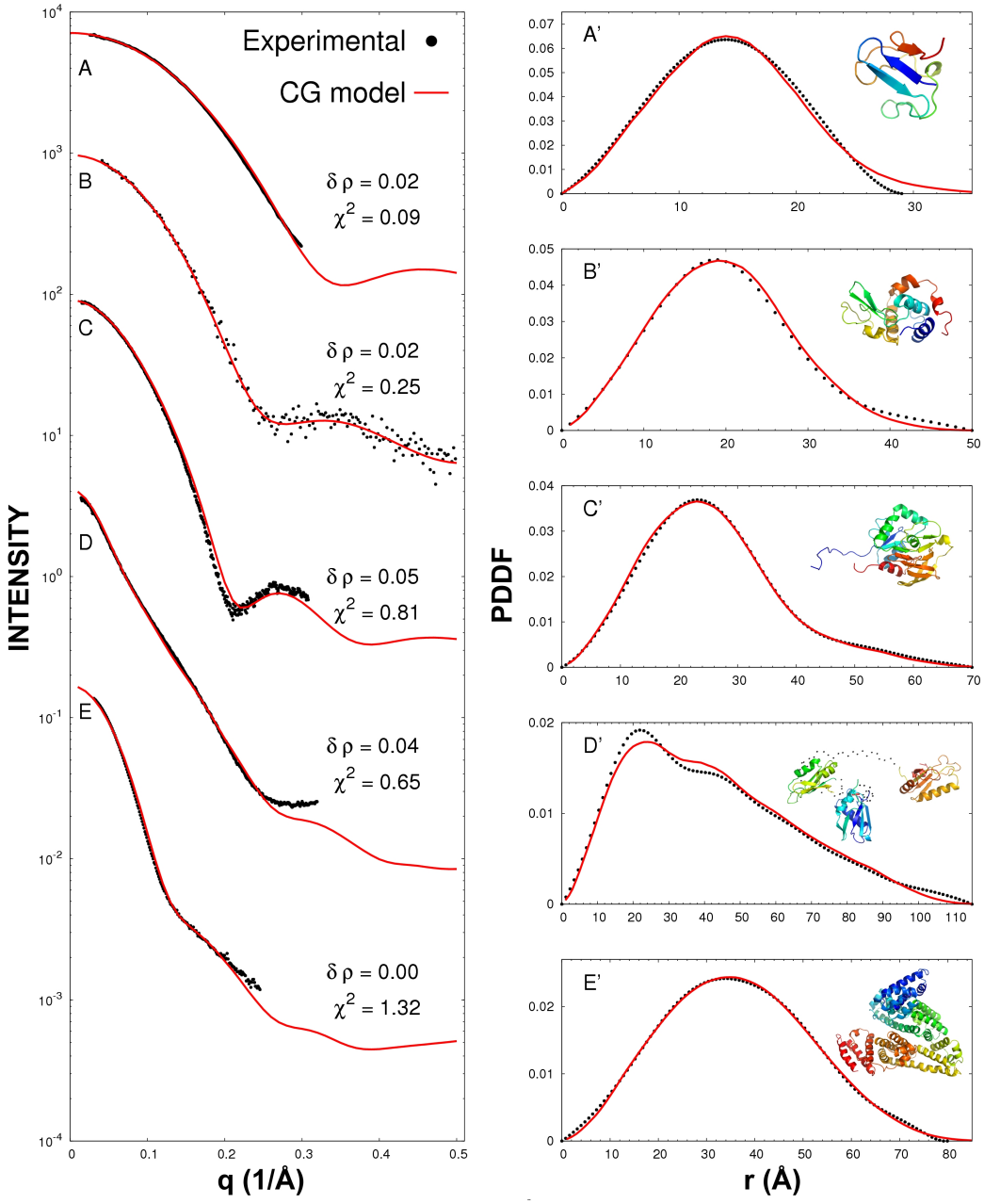


Figura 4.4: Comparação entre os perfis de SAXS teóricos produzidos pelo modelo CG e os experimentais em ambos os espaços recíproco (esquerda) e real (direita). (A e A') PF1282 de *P. furiosus*. Dados exp. retirados do BIOISIS (BIOISIS ID: 1RBDGP). (B e B') Lisozima. Dados exp. de SAXS retirados do pacote CRYSOL e modelo estrutural do PDB (PDB ID: 6LYZ). (C e C') PF1528 de *P. furiosus* (BIOISIS ID: 1AMIGP). (D e D') U2AF65 fator de *Splicing* (BIOISIS ID: 1U2FKP). (E e E') HSA. Dados de SAXS coletados pelo nosso grupo e não publicados. Estrutura de homóloga (PDB ID: 1AO6A).

estrutura proteica, fornecem um ajuste com $\chi^2 = 0,203$ e $\chi^2 = 0,202$, respectivamente. Ainda que os ajustes com CRYSOL e FoXS são ligeiramente melhores, estes dados confirmam que o modelo CG, o qual é baseado apenas no traçado dos C_α, e apenas um parâmetro livre ($\delta\rho$) pode fornecer um ajuste aos dados experimentais com suficiente acurácia, bem abaixo no erro típico das medidas de SAXS.

Reconstrução do envelope a partir do modelo CG e o PDB

Como primeira implementação do modelo CG na identificação de proteínas-molde, foi examinado o grau de habilidade que uma métrica baseada em SAXS possui no reconhecimento de forma tridimensional da proteína-alvo. Esta primeira abordagem é muito semelhante àquela descrita no Capítulo 3, quando foi tratado do reconhecimento de formas tridimensionais de objetos de mesma classe. Logo, os resultados desta seção ainda não se vale de nenhuma informação sobre alinhamentos *threading* nem se considera privilegiadamente eventuais proteínas homólogas ao alvo. Apenas são utilizadas as estruturas depositadas no PDB e uma dada métrica baseada em SAXS para que se possa, a partir do perfil experimental das 5 proteínas-alvo, criar uma classificação das estruturas do PDB que mais se assemelham estruturalmente a elas[2].

Para uma rápida ordenação, foi utilizada uma métrica baseada na correlação de Pearson entre as curvas $p(r)$ descrita pela Equação 4.11 que tende à unidade quando as formas das curvas $p(r)$ são semelhantes. Uma vez que todas as estruturas do PDB são classificadas, selecionamos a estrutura da proteína com maior afinidade com a $p(r)$ da proteína-alvo e em seguida, adicionamos diversas pequenas esferas virtuais no interior da estrutura tridimensional, de forma que esse conjunto de esferas preencham todo o volume proteico. Para representar a superfície molecular da proteína de interesse, utilizamos o programa NCSMASK [55] para criar um envelope formado pelo conjunto das esferas virtuais. Para fins de comparação, reconstruímos o envelope molecular usando o programa DAMMIF [53], que é uma abordagem mais sofisticada que utiliza apenas a curva experimental de SAXS e simulação de Monte Carlo para posicionar esferas virtuais no es-

[2]É importante salientar que abordagens semelhantes são encontradas na literatura, como é o caso do banco de dados DaRa `http://dara.embl-hamburg.de/` [111] que possui milhares de curvas teóricas pré-calculadas em um banco de dados local computadas a partir de diversas estruturas tridimensionais de proteínas com o mesmo intuito de ordená-las segundo o acordo com o perfil de SAXS inserido pelo usuário.

paço, de forma que a correspondente distribuição produza um perfil teórico de SAXS semelhante à curva experimental.

A Figura 4.5 mostra a comparação dos envelopes obtidos por DAMMIF e aqueles selecionados pelo modelo CG. Apesar da simplicidade do método empregado aqui, observa-se que a reconstrução tridimensional promovida pelo nosso método está em bom acordo com os envelopes obtidos por DAMMIF. Entretanto, não se pode esperar que este procedimento simples reconheça características estruturais de mais alta resolução das proteínas-alvo. Conforme já discutido no Capítulo 3, objetos com formas tridimensionais semelhantes não possuem necessariamente a mesma categorização, isto é, é possível que objetos de classes diferentes possuam perfis de SAXS parecidos. De fato, as proteínas-molde selecionadas através da maior correlação com a curva $p(r)$ do alvo, têm na média, um alinhamento estrutural com a respectiva estrutura da proteína-alvo de TM-score de $\approx 0,3$ o que está razoavelmente perto da similaridade aleatória. Já quando usamos apenas MUSTER neste conjunto de 5 proteínas, a média do TM-score sobe para $0,51$, mostrando que ao menos para este conjunto considerado, *threading* é mais eficiente que dados de SAXS no reconhecimento de enovelamentos.

Por outro lado, esses resultados encorajam o uso combinado de SAXS com *threading* à medida em que esta abordagem pode ser extremamente útil em casos onde alinhamentos incorretos obtidos por *threading* venham a produzir formas tridimensionais que não concordam com o perfil experimental obtido por SAXS. Logo, SAXS agiria como uma espécie de filtro para os alinhamentos *threading*.

4.3.2 – Resultados dos experimentos computacionais de larga escala

Correlação das métricas baseadas em SAXS e o TM-score

Existem diferentes métricas que podem ser usadas na comparação dos perfis de SAXS entre duas proteínas. Aqui, nós avaliamos o desempenho de nove diferentes funções (F_{SAXS}), que estão listadas na Tabela 4.1 frente ao conjunto de treino composto por 341 estruturas para que possamos conhecer qual a melhor abordagem que trará a melhor resposta quantitativa do reconhecimento de enovelamentos para que possamos avaliar o método SAXSTER proposto neste trabalho frente ao conjunto de teste.

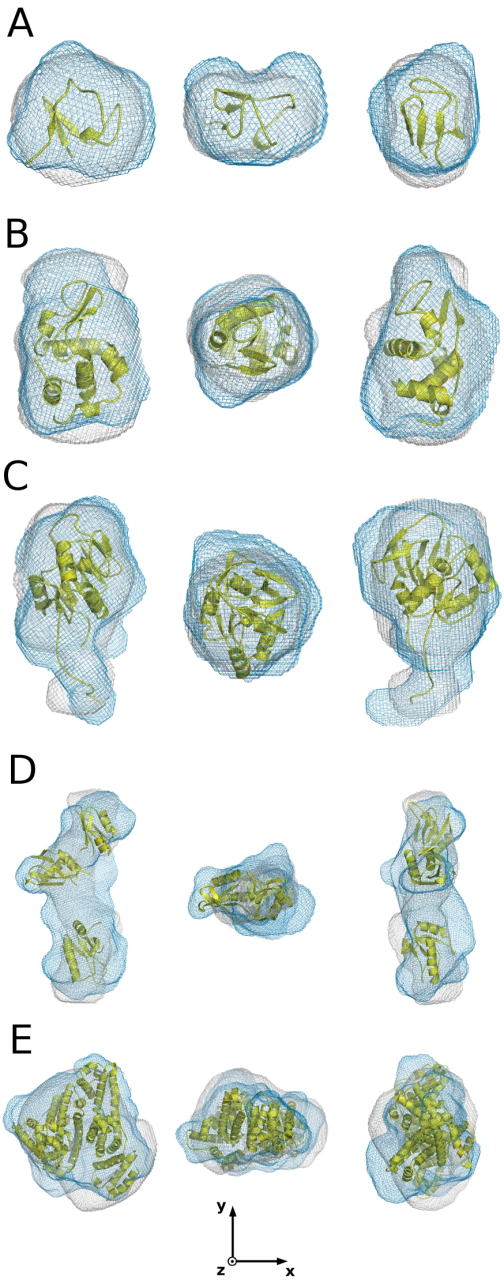


Figura 4.5: Envelopes de baixa resolução obtidos com DAMMIF (em cinza) e pelo método de classificação baseado no modelo CG (em azul). As estruturas das proteínas-alvo são mostradas no interior dos envelopes. Para cada caso, as figuras do centro e da direita são diferentes vistas a partir da figura à esquerda, por uma rotação de 90^0 no sentido anti-horário em relação ao eixo x e ao eixo y, respectivamente. (A) 1RBDGP; (B) 6LYZ; (C) 1AMIGP; (D) 1U2FKP; (E) HSA.

4.3 Resultados e discussão

A partir da sequência de aminoácidos de uma proteína-alvo, inicialmente rodamos MUSTER de forma a obter o alinhamento *threading* para cada uma das 40096 estruturas do PDB desconsiderando todas aquelas que possuem identidade sequencial maior que 30% com o alvo. Isto é feito para que a imparcialidade nos resultados seja mantida, e proteínas que possuem homólogas no PDB possam ser comparadas às proteínas que não possuem.

Em seguida, preenchemos os *gaps* de todos os alinhamentos realizando uma simulação de *random walks*, conforme descrição na Seção 4.2.3. Com isso, obtivemos estruturas tridimensionais composta apenas de C_α que refletem a mesma sequência de aminoácidos da proteína-alvo. O passo seguinte consiste no cálculo das curvas teóricas de SAXS para cada estrutura, agora denominadas modelos, através do modelo CG usando a Equação 4.8, podendo ser tanto $I(q)$ quanto $p(r)$ com o propósito de compará-las ao perfil de SAXS do alvo. Por fim, os modelos são ordenados pela métrica utilizada que, de maneira geral, é dada pela Equação 4.12.

Como um exemplo, a Figura 4.6 (A e B) mostra os resultados obtidos para uma proteína-alvo (PDB ID: 1IC2, cadeia A) utilizando o Esquema I e Esquema IX para F_{SAXS} da Tabela 4.1, respectivamente. Nesta figura, *Z-score* é a medida estatística da métrica F utilizada, podendo ser $F = F_{SAXS}$ ou $F = F_{MUSTER} + F_{SAXS}$, ou seja, $Z - score = (F - \langle F \rangle)/\sigma$, onde σ é o desvio padrão dos valores computados para F.

Neste exemplo, ambos esquemas utilizados para F_{SAXS} são aptos para o reconhecimento de bons moldes, isto é, moldes que possuem um alinhamento estrutural com o alvo de *TM-score* $> 0,5$ estão relacionados a um respectivo *Z-score* alto. Por outro lado, o Esquema I apresenta desvantagens em relação ao Esquema IX por apresentar modelos com *Z-score* muito próximos das primeiras posições do *ranking*, permitindo que diversos falsos positivos (*TM-score* $< 0,5$ e *Z-score* alto) se acumulem no entorno de verdadeiros positivos (*TM-score* $> 0,5$ e *Z-score* alto) podendo haver pouca margem para a distinção da qualidade entre os modelos, já que em casos reais, dispomos apenas do *Z-score* para fazer a avaliação.

Uma maneira de escolher a melhor função F_{SAXS} é avaliar o coeficiente de correlação de Pearson (PCC, *Pearson Correlation Coeficient*), entre TM-score e Z-score associado (Tabela 4.1). Por exemplo, a distribuição promovida pelo Esquema IX (Figura 4.6 B), claramente possui maior correlação nos dados (PCC = 0,35) que o Esquema I (PCC = 0,19). Quando adicionamos a métrica do programa MUSTER (F_{MUSTER}) ao escore F, ambas as distribuições (Figura 4.6 A' e B') passam a ter PCC $\approx$ 0,73 indicando que MUSTER por si, possui uma melhor correlação

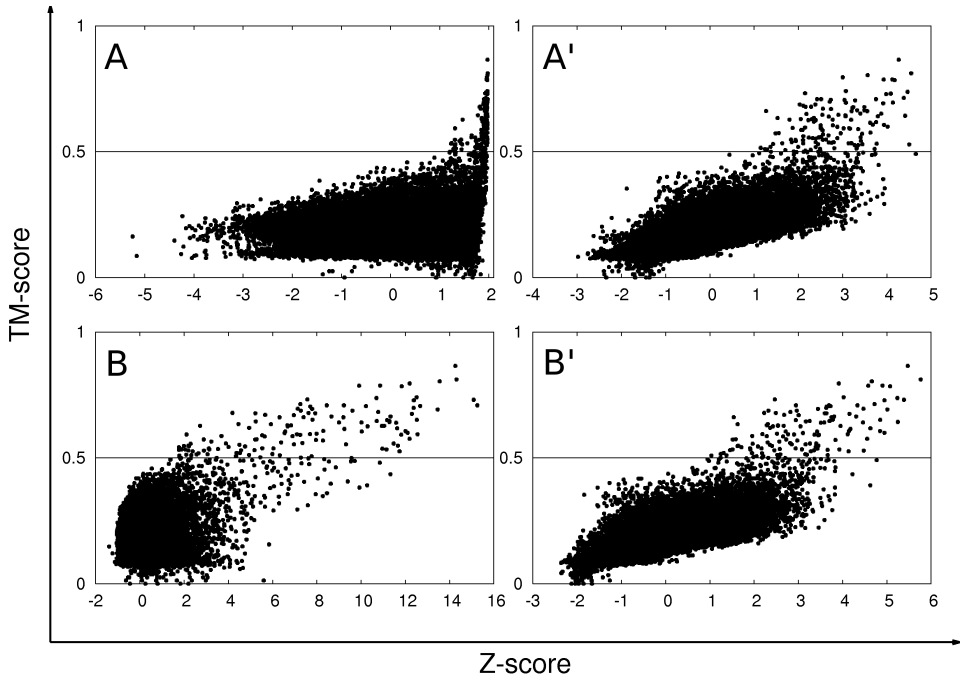


Figura 4.6: Exemplo de distribuição do TM-score vs Z-score para uma proteína-alvo (PDB ID: 1IC2) contra todas as estruturas do PDB que possuem identidade sequencial $\leq 30\%$ com o alvo. Gráficos à esquerda (A e B) mostram dados obtidos por duas métricas diferentes baseadas em SAXS e os da direita (A' e B') mostram o efeito da adição do escore do programa MUSTER aos respectivos escores de SAXS. (A) e (A') usam o Esquema I como parte do escore de SAXS e (B) e (B') usam o Esquema IX, ambos descritos na Tabela 4.1.

Tabela 4.1: Resultado da classificação de proteínas-molde para o conjunto de treino a partir de diversas funções-escore.

		<PCC>[a]		<TM-score>[b]	
Esquema	Função-escore: $F_{SAXS}(i,j)$	SAXS	SAXSTER	Top 1	Top 5
I[c]	$= \frac{1}{n}\sum_{k=1}^{n}\left\{\frac{I_i(q_k)-I_j(q_k)}{\sigma(q_k)}\right\}^2$	0,34	0,47	0,5414	0,6042
II	$= \frac{\sum_{k=1}^{n}\lvert I_i(q_k)-I_j(q_k)\rvert}{\sum_{k=1}^{n}\lvert I_i(q_k)\rvert}$	0,39	0,31	0,5431	0,6063
III	$= \frac{\sum_{k=1}^{n}\lvert log\left[I_i(q_k)\right]-log\left[I_j(q_k)\right]\rvert}{\sum_{k=1}^{n}\lvert log\left[I_i(q_k)\right]\rvert}$	0,36	0,31	0,5455	0,6078
IV	$= \sum_{k=1}^{n} q_k\left\{I_i(q_k)-I_j(q_k)\right\}^2$	0,38	0,36	0,5423	0,6037
V	$= \sum_{k=1}^{n} q_k\left\{log\left[I_i(q_k)\right]-log\left[I_j(q_k)\right]\right\}^2$	0,30	0,30	0,5463	0,6085
VI	$= \frac{\sum_{k=1}^{n} q_k^2\lvert I_i(q_k)-I_j(q_k)\rvert}{\sum_{k=1}^{n} q_k^2 I_i(q_k)}$	0,37	0,30	0,5468	0,6085
VII	$= \sum_{k=1}^{n}\left\{p_i(r_k)-p_j(r_k)\right\}^2$	0,38	0,32	0,5429	0,6060
VIII[d]	$= log\left\{1-corr\left(I_i(q),I_j(q)\right)\right\}$	0,36	0,62	0,5430	0,6083
IX	$= log\left\{1-corr\left(p_i(r),p_j(r)\right)\right\}$	0,35	0,61	0,5484	0,6095

[a] Média do coeficiente de correlação de Pearson entre TM-score e Z-score.

[b] Média do TM-score entre as proteínas-alvo do conjunto de treino e as respectivas proteínas-molde.

[c] $\sigma(q)$ é o erro experimental de $I(q)$. No caso de curvas teóricas, $\sigma(q) = I(q) \times (q + 0,15) \times 0,30$ [51].

[d] A função $corr(f,g)$ é o coeficiente de correlação de Pearson entre f e g definido pela Equação 4.11.

nos dados que qualquer métrica baseada em SAXS. Entretanto, a função escore F que usa o Esquema IX mais o escore do MUSTER ($F = F_{MUSTER} + F^{IX}_{SAXS}$) consegue melhorar o TM-score do primeiro modelo que é selecionado por *threading* apenas, aumentando o TM-score de 0,49 para 0,81.

Para resumir os resultados, a Tabela 4.1 mostra o PCC médio e o TM-score médio do melhor alinhamento (Top 1) e o melhor entre o cinco primeiros alinhamentos (top 5) para 341 proteínas do conjunto de teste. A coluna 3 apresenta os resultados dos escores baseados apenas nos dados de SAXS (F_{SAXS}) e as colunas 4-6 mostram os resultados combinados ao Z-score do MUSTER ($F_{SAXSTER} = Z - score_{MUSTER} + wF_{SAXS}$). Os pesos ótimos w desta combinação foram determinados maximizando a correlação $F_{SAXSTER}$ com o TM-score. Globalmente, a média de PCC para SAXS foi aproximadamente 0,36. Quando combinados ao escore de MUSTER, os escores baseados nos Esquemas VIII e IX tem um destacado PCC entre os demais. Embora o TM-score médio do primeiro modelo e o melhor entre os cinco melhores são todos maiores em relação aos modelos selecionados originalmente pelo MUSTER, o melhor desempenho vem do Esquema IX. Estes dados em conjunto, sugerem que apesar das diferenças sutis entre os escores baseados em SAXS, eles têm diferentes influências no reconhecimento de enovelamentos, e dependendo do caso, certas funções escore podem ser mais eficientes que outras.

O Esquema IX foi finalmente destacado como sendo ligeiramente melhor que os demais de acordo com aos resultados para as 341 proteínas do conjunto de treino. Esta forma funcional baseada na $p(r)$ tem três características peculiares que podem ter contribuído para o desempenho observado. Primeiro, ela não depende de nenhum fator de escala entre os perfis de SAXS, o que evita erros de posicionamento relativo entre as curvas; Segundo, sua forma logarítmica funcional realça diferenças numéricas sutis entre Z-score próximos. Terceiro, é baseada no espaço real e sofre menos influência das limitações criadas pelo modelo CG cuja derivação não leva em conta nenhuma correção de volume excluído, o qual é mais importante no espaço recíproco.

Teste de desempenho

Testamos o programa SAXSTER para 412 proteínas-alvo aleatoriamente selecionadas por PISCES, das quais 232 são consideradas de fácil predição estrutural e 180 são casos difíceis, de acordo com a categorização do MUSTER. A Tabela 4.2 compara o TM-score das proteínas-

molde selecionadas pelo Z-score original de MUSTER e também combinado com diversas métricas inspiradas em SAXS. No caso de SAXSTER, utilizamos a métrica baseada na $I(q)$ e na $p(r)$, segundo os Esquemas VIII e IX presentes na Tabela 4.1, respectivamente. Também testamos os três métodos (*random walk*, MODELLER e I-TASSER) de construção dos modelos a partir dos alinhamentos *threading*.

De modo geral, todos os métodos baseados em SAXS ajudaram a melhorar a predição do MUSTER uma vez que em todos esses casos a média do TM-score é aumentada. Esta melhoria é estatisticamente significativa com um p-valor no intervalo de [10^{-6} - 10^{-8}] de acordo com o teste t de Student para todos os casos. Estes dados mostram que o modelo CG funciona tão bem quanto CRYSOL e FoXS, que exploram modelos com todos os átomos. A vantagem do modelo CG é de poder operar sobre estruturas com carbonos alfa apenas, diferentemente destes outros programas[3]. Logo, podemos combinar a utilização do modelo CG com técnicas de preenchimento de *gaps* que utilizam apenas os C_α como *random walk* ou I-TASSER, tornando o tempo computacional empregado na construção dos modelos viável (minutos-horas) quando comparado ao tempo proibitivo despendido na construção de modelos completos com MODELLER (horas-dias).

Não existem diferenças significativas entre os modelos construídos por *random walk*, MODELLER e I-TASSER do ponto de vista do TM-score. Apesar da maior qualidade dos *loops* construídos com I-TASSER em relação aos métodos de *random walk* e MODELLER, esse tipo de modelagem não contribui com um alto ganho sobre os respectivos TM-score médio das proteínas consideradas fáceis (p-valor $< 10^{-3}$). A diferença torna-se indistinguível em casos de proteínas consideradas difíceis, provavelmente devido aos grandes *gaps* dos alinhamentos e a qualidade dos *loops* construídos por estes três métodos ser igualmente pobre.

Na Figura 4.7, comparamos lado a lado o desempenho de MUSTER e de SAXSTER a fim de se verificar a influência dos dados de SAXS na predição dos enovelamentos. Nota-se que, seja qual for a estratégia utilizada, existem muito mais pontos acima da reta diagonal que abaixo, demonstrando que há uma indiscutível melhora quando associamos os perfis de SAXS aos alinhamentos originais obtidos por MUSTER. Vale a pena mencionar que SAXSTER, apesar de melhorar a classificação, não altera o alinhamento produzido por MUSTER. Se MUSTER logra sucesso

[3]Há uma versão do programa FoXS que possui a opção explícita de ajuste da curva experimental a partir das coordenadas dos C_α de uma proteína, muito embora a qualidade do ajuste esteja muito aquém da opção de se utilizar todos os átomos, de acordo com nossos testes preliminares.

Tabela 4.2: TM-score médio das primeiras proteínas-molde (melhor entre as cinco primeiras) selecionadas por diferentes métodos.

Método	Todos os alvos	Fáceis ($n = 232$)	Difíceis ($n = 180$)
MUSTER	0,5299 (0,5952)	0,6330 (0,6885)	0,3970 (0,4750)
+ CRYSOL + MODELLER	0,5457 (0,6011)	0,6440 (0,6940)	0,4189 (0,4812)
+ FoXS + MODELLER	0,5456 (0,6018)	0,6428 (0,6936)	0,4204 (0,4836)
+ SAXSTER I(q) + *random walk*	0,5438 (0,6036)	0,6416 (0,6938)	0,4178 (0,4874)
+ SAXSTER I(q) + MODELLER	0,5461 (0,6032)	0,6414 (0,6923)	0,4233 (0,4884)
+ SAXSTER I(q) + I-TASSER	0,5486 (0,6036)	0,6446 (0,6939)	0,4250 (0,4872)
+ SAXSTER p(r) + *random walk*	0,5467 (0,6005)	0,6407 (0,6916)	0,4256 (0,4832)
+ SAXSTER p(r) + MODELLER	0,5449 (0,6052)	0,6411 (0,6950)	0,4209 (0,4895)
+ SAXSTER p(r) + I-TASSER	0,5479 (0,6045)	0,6436 (0,6954)	0,4245 (0,4874)

em admitir a melhor proteína-molde disponível na primeira posição do *ranking*, SAXSTER é incapaz de melhorar a classificação, o que explica a grande quantidade de pontos ao longo da linha diagonal na Figura 4.7.

Para todos os casos de sucesso, cerca de 93% das proteínas-molde selecionadas ocupam as primeiras 10 posições do *ranking* original de MUSTER. Em particular, os melhores moldes selecionados para os alvos díficeis foram identificados entre os primeiros 40 *hits*. Portanto, estes dados indicam que SAXSTER precisa somente focar nas primeiras posições ordenadas pelo MUSTER. Por esta razão, o algoritmo foi otimizado a partir do conjunto treino levando em conta apenas as 100 primeiras proteínas-molde identificadas por *threading*, resultando em um peso de $w = 0.809$ na função escore $F_{SAXSTER} = F_{MUSTER} + wF_{SAXS}$ do Esquema IX da Tabela 4.1.

Na Figura 4.8 (A-D), quatro exemplos típicos de sucesso do SAXSTER são mostrados. Os alvos 2FKCA e 2PJPA (Figura 4.8 A e B) são proteínas com múltiplos domínios. MUSTER seleciona uma proteína-molde que possui um alinhamento completamente errado com 2FKCA, e alinha corretamente apenas um dos domínios de 2PJPA. Estes equívocos provocam um desacordo entre os perfis de SAXS produzindo um alto valor de χ^2 entre as $I(q)$ de 2FKCA ($\chi^2 = 8,09$) e 2PJPA

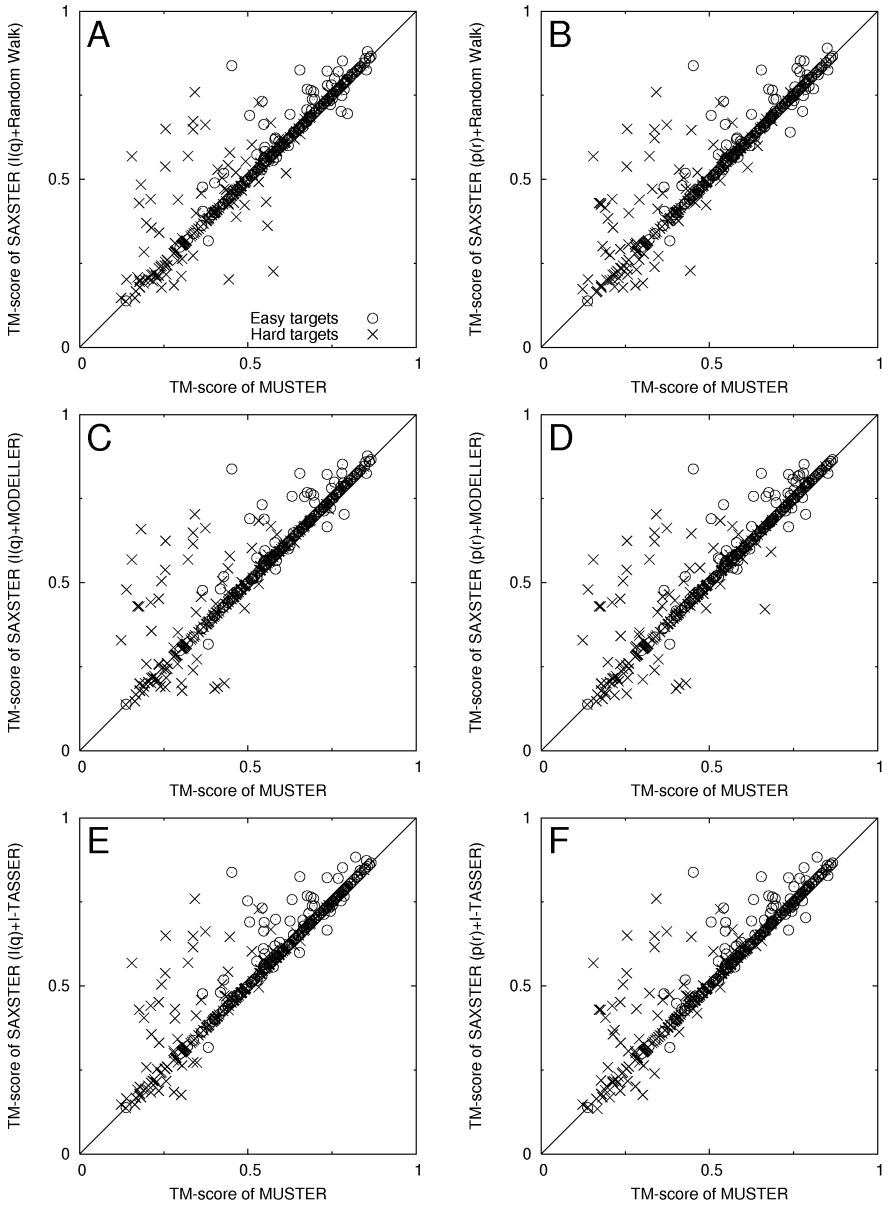


Figura 4.7: TM-score das primeiras proteínas-molde selecionadas por SAXSTER contra as selecionadas por MUSTER. As métricas são baseadas na $I(q)$ (coluna à esquerda) e na $p(r)$ (coluna à direita). Modelos foram construídos por *random walk* (A e B), MODELLER (C e D) e por I-TASSER (E e F).

($\chi^2 = 16,43$) e suas respectivas proteínas-molde selecionadas por MUSTER. SAXSTER ajuda filtrar os alinhamentos incorretos e identificar as proteínas-molde corretas melhorando o acordo com a curva $I(q)$ das proteínas-alvo, resultando em $\chi^2 = 0,62$ e $\chi^2 = 0,32$, para 2FKCA e 2PJPA

respectivamente. O escore combinado além de promover um maior acordo entre as curvas de SAXS, também melhora a topologia global da proteína-molde identificada. Consequentemente, o TM-score entre a proteína-molde e a proteína-alvo foi melhorado de 0,16 para 0,61 no caso do alvo 2FKCA, e de 0,48 para 0,81, no caso de 2PJPA.

Um exemplo de proteína com um único domínio é mostrada na Figura 4.8 D. Este alvo possui uma forma alongada composta por duas hélices-α, mas a proteína-molde selecionada por MUSTER é globular. SAXSTER consegue localizar a correta topologia fazendo com que o TM-score aumente de 0,32 para 0,49 e o χ^2 diminua de 17,73 para 0,91, o que demonstra mais uma vez que os dados de SAXS foram de fato o agente causador por trás da identificação do enovelamento correto.

Teste do SAXSTER para proteínas alongadas

Do ponto de vista da $p(r)$, uma proteína pode ser interpretada por seu histograma de distâncias intramolecular. Supomos que em nossa aplicação do SAXSTER, um escore baseado em SAXS deve ser menos sensível para proteínas de forma globular, uma vez que a maioria das proteínas do PDB possuem formas similares em baixa resolução, e portanto, apresentam histogramas de distâncias parecidos. Curiosamente, no Capítulo 3 vimos fato semelhante ocorrer, onde verificou-se que certos objetos tridimensionais são mais suscetíveis a serem casos de sucesso no reconhecimento de forma em relação a outros com padrão tridimensional mais frequentemente encontrado.

Para testar esta hipótese para o caso de proteínas, construímos um novo conjunto de teste composto apenas por proteínas alongadas e examinamos a habilidade do SAXSTER no reconhecimento das proteínas-molde.

Em uma situação real, a estrutura da proteína-alvo normalmente não é conhecida, logo precisamos de algum método para quantificar, mesmo que aproximadamente, o grau de esfericidade ou de alongamento que uma estrutura tridimensional possui. Para isso, primeiramente analisamos o raio de giro de 7466 proteínas não redundantes, isto é, identidade sequencial $< 30\%$ entre qualquer par em função do número de aminoácidos. A Figura 4.9 (a) mostra um ajuste da distribuição de raios de giro em função do número de aminoácidos na forma $R_G = 4,27L^{0,27}$. Este ajuste é muito próximo da função encontrada na literatura, $R_G = 3,0L^{0,33}$ [112, 113]. Portanto,

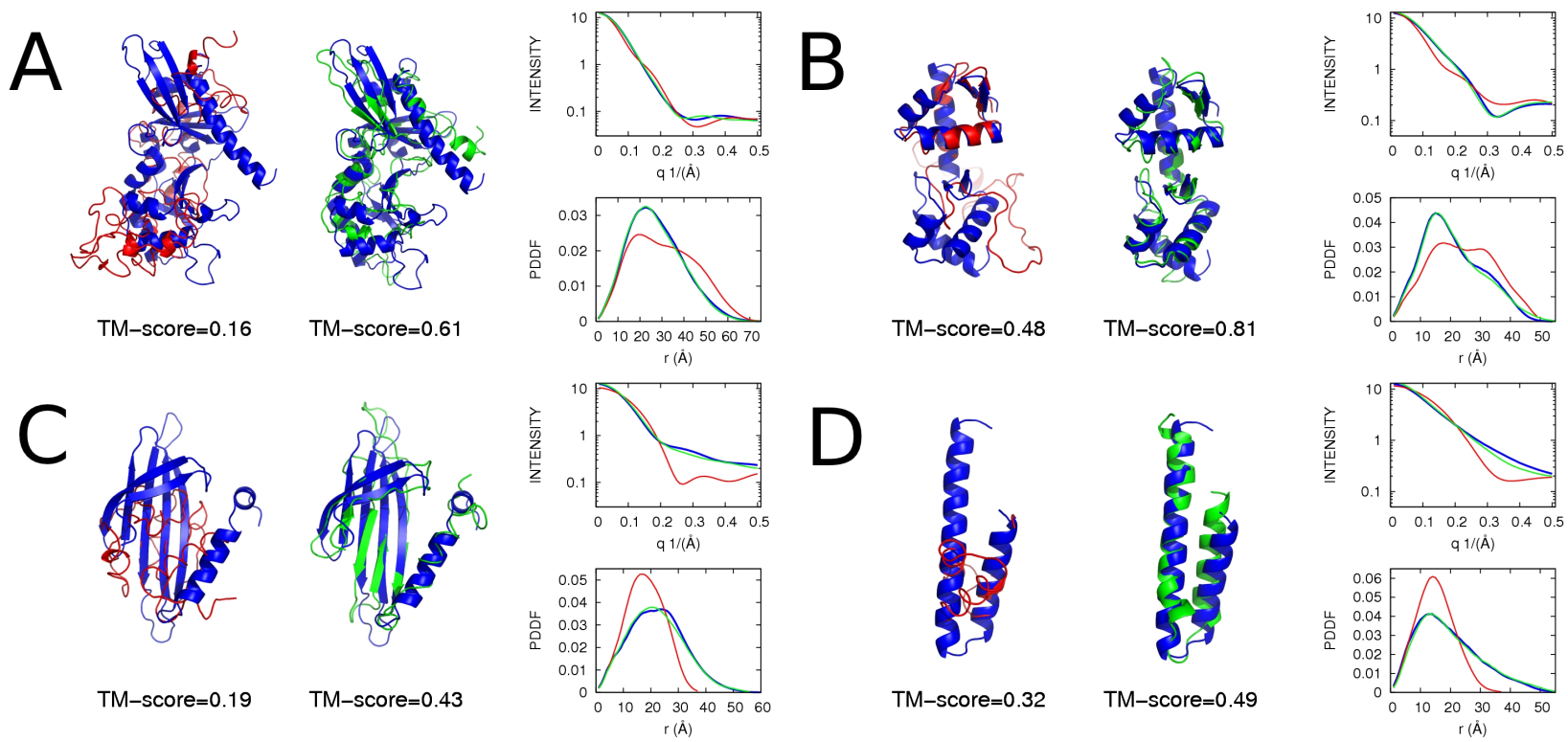
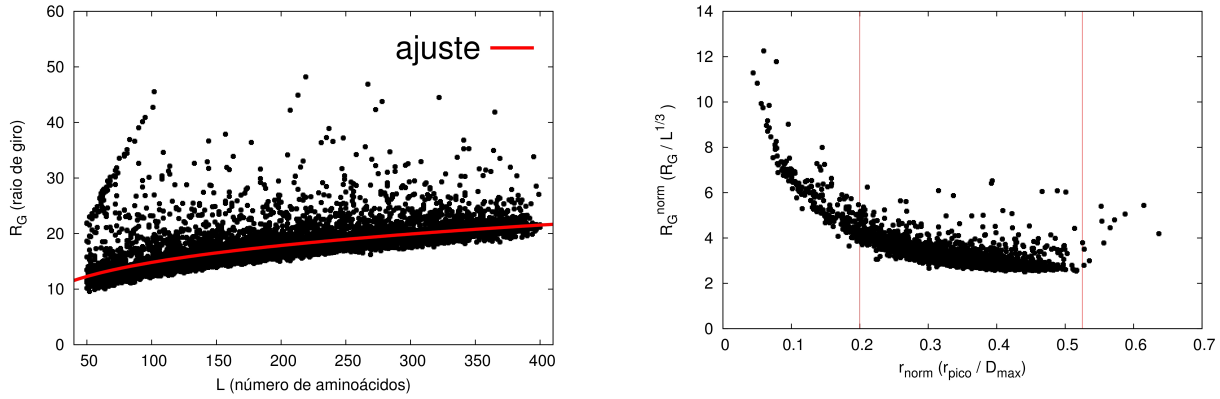


Figura 4.8: Exemplos representativos de proteínas-molde selecionadas por MUSTER e SAXSTER. As estruturas em azul representam as proteínas-alvo enquanto que as em vermelho e em verde representam as proteínas-molde selecionadas por MUSTER e SAXSTER, respectivamente. Os perfis $I(q)$ e $p(r)$ são mostrados para cada caso seguindo o mesmo código de cores. PDB ID: (A) 2FKCA, (B) 2PJPA, (C) 2W4YA, (D) 2RKLA.



(a) (b)

Figura 4.9: Análise do raio de giro de 7466 proteínas não redundantes. (a) Raio de giro em função do número de aminoácidos. (b) Raio de giro normalizado em função da posição do pico da $p(r)$ normalizado pela máxima distância intramolecular. A região onde a maioria das proteínas globulares se localizam está indicada entre retas verticais vermelhas: à direta localizada em $r_{norm} = 0,525$ (esfera perfeita) e à esquerda, arbitrariamente localizada em $r_{norm} = 0,20$.

para todos os fins práticos, podemos considerar o raio de giro $R_G \propto L^{1/3}$ para definir o raio de giro normalizado na forma $R_G^{norm} = R_G/L^{1/3}$.

Pela forma da $p(r)$ podemos inferir qualitativamente o tipo de envelope molecular da proteína alvo. Em particular, uma esfera perfeita tem a seguinte distribuição [8] :

$$p(r) = 12x^2(2 - 3x + x^3), \text{ com } x \equiv \frac{r}{D_{max}}, \tag{4.14}$$

onde D_{max} é a máxima distância intramolecular. Esta função possui um máximo global localizado em $x \approx 0,525$ e portanto é praticamente simétrica em relação ao seu centro. Para termos uma medida aproximada do grau de esfericidade ou de alongamento de uma proteína, medimos a posição do máximo global da $p(r)$ (r_{pico}) em relação à D_{max} definindo $r_{norm} = r_{pico}/D_{max}$. Em geral, proteínas oblatas (achatadas), prolatas (alongadas) e ocas possuem a seguinte relação: $r_{norm}^{prolata} < r_{norm}^{oblata} < r_{norm}^{esfera} < r_{norm}^{ocas}$.

A Figura 4.9 (b) mostra a distribuição do raio de giro normalizado em função da posição do máximo global normalizado da $p(r)$ para 7466 proteínas. Conforme previsto, a grande maioria das proteínas concentra-se em uma região com pouca variabilidade do raio de giro normalizado

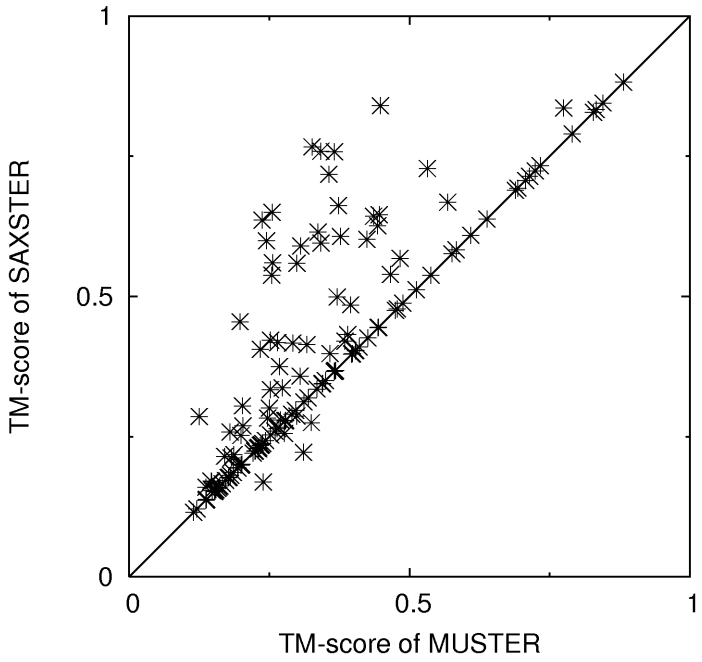


Figura 4.10: TM-score das primeiras proteínas-molde selecionadas por SAXSTER e por MUSTER para 141 casos de proteínas-alvo difíceis com perfis de $p(r)$ menos frequentes.

localizada no intervalo $0,20 \leq r_{norm} \leq 0,525$, onde a concentração de proteínas essencialmente globulares é relativamente maior. Em particular, a região $r_{norm} \leq 0,20$ abriga não só proteínas alongadas mas também proteínas com múltiplos domínios globulares dispostos longitudinalmente, como o caso da proteína da Figura 4.4 D'.

Deste grupo de proteínas, de formas tridimensionais menos frequentes, extraímos 141 casos difíceis ($Z_{MUSTER} \leq 7,5$) para aferir o desempenho de predição de SAXSTER. Os resultados são mostrados da Figura 4.10

Estes resultados mostram claramente que a melhora do reconhecimento das formas tridimensionais por SAXSTER é mais acentuada para proteínas com formas alongadas quando comparadas ao resultado com proteínas de topologias diversas. Dos 141 casos analisados, 57 têm TM-score maior que o identificado por MUSTER, sendo que em 16 deles o TM-score aumentou mais que 0,25, o que essencialmente significa converter proteínas que antes não eram preditas para proteínas com o correto enovelamento. Entretanto, 78 casos permaneceram inalterados devido aos melhores alinhamentos já terem sido identificados por MUSTER, não dando margem para SAXSTER melhorar a identificação. Globalmente, o TM-score médio de SAXSTER é aumentado

em 18% (de 0,3394 para 0,4013). Esta melhora é estatisticamente significativa, correspondendo a um p-valor $\leq 10^{-9}$ no teste t de Student.

4.3.3 – Servidor SAXSTER

Como um resultado adicional, podemos destacar o servidor SAXSTER que esta disponível publicamente na rede (Figura 4.11). SAXSTER foi concebido com a finalidade de se obter moldes estruturais para proteínas de interesse usando modelagem computacional e vínculos experimentais obtidos pela técnica de espalhamento de raios X a baixos ângulos.

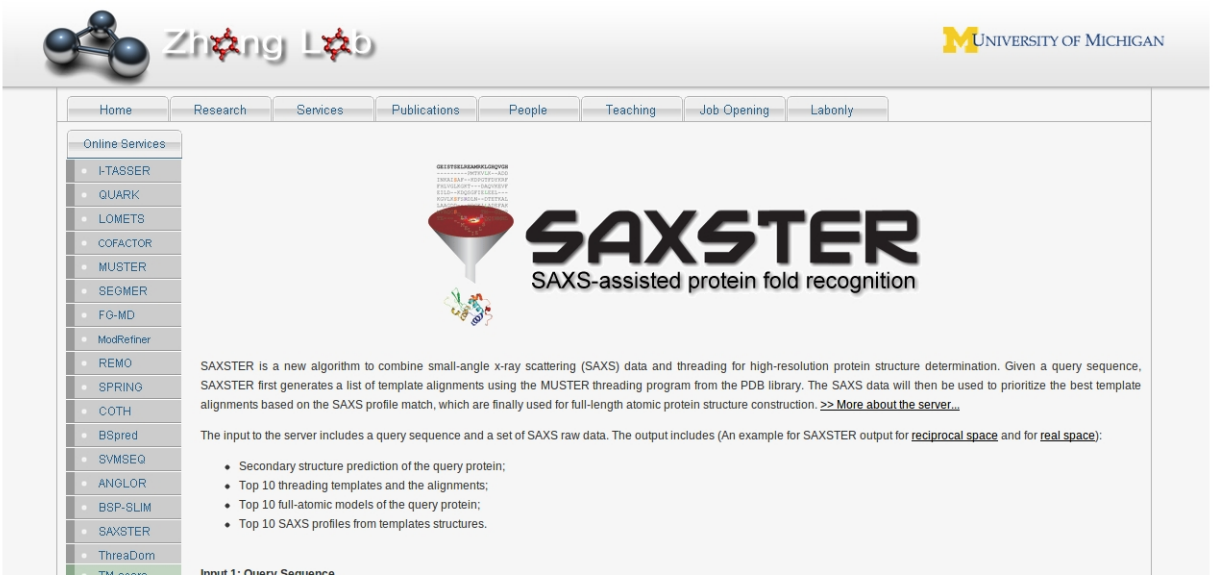


Figura 4.11: Programa disponível publicamente através da internet pelo endereço `http://zhanglab.ccmb.med.umich.edu/SAXSTER/`.

4.4 – Conclusão

Desenvolvemos um novo método chamado SAXSTER que utiliza dados de SAXS para melhorar a predição estrutural de proteínas. A estratégia extrai os perfis de SAXS a partir de alinhamentos *threading* e modelos estruturais baseados nos carbonos alfa realizando uma simulação computacional do tipo *Coarse-Grained* para obtenção das curvas de intensidade de espalhamento $I(q)$ ou da função de distribuição de distâncias $p(r)$. Em seguida, a curva $I(q)$ ou $p(r)$ de cada modelo é comparada com os dados de SAXS da proteína-alvo para priorizar as proteínas-molde que tenham um perfil de SAXS semelhante ao do alvo.

Testamos SAXSTER em 412 proteínas não redundantes. Constatamos que os dados de SAXS podem melhorar consistentemente o resultado global da classificação dos alinhamentos. Como os alinhamentos ótimos contém *gaps*, exploramos três métodos de modelagem: *random walks*, MODELLER e I-TASSER. Embora os resultados de reconhecimento das estruturas são pouco sensíveis aos métodos de reconstrução dos *loops*, todas as abordagens melhoraram a predição e foram estatisticamente significativas, com p-valores que variam de 10^{-6} a 10^{-8}. Embora estes sejam resultados encorajadores, deve-se notar que neste trabalho focamos apenas na reclassificação e seleção dos modelos *threading* sem qualquer alteração no alinhamento em si. Se os algoritmos construtores dos alinhamentos falham, SAXSTER torna-se incapaz de obter modelos estruturais corretos.

Adicionalmente, analisamos o desempenho de SAXSTER frente a proteínas alongadas a partir de um segundo conjunto de 141 proteínas com $p(r)$ acentuadamente assimétricas. De acordo com TM-score médio, concluimos que SAXSTER tende a melhorar o desempenho preditivo conforme a estrutura da proteína-alvo se afasta da forma globular.

Capítulo 5

Considerações Finais

Neste trabalho de tese, o problema da caracterização estrutural de proteínas foi abordado de maneira contextualizada e com um viés em modelagem computacional utilizando vínculos experimentais obtidos com a técnica de espalhamento de raios X a baixos ângulos (SAXS).

A discussão foi iniciada abordando o problema específico da caracterização estrutural da proteína da bactéria *Xylella fastidiosa* que é o agente causador de diversas doenças em plantas economicamente importantes para diversos países. O problema foi inicialmente abordado do ponto de vista experimental, por técnicas de espalhamento de raios X como SAXS e cristalografia. As análises de dados de SAXS, que são inerentemente de baixa resolução, proporcionaram um claro entendimento que existem limites de aplicação, mas por outro lado, possibilitaram a determinação do estado oligomérico associado a unidade funcional biológica da XfSurE confirmando especulações anteriores baseadas em dados cristalográficos. Quando associada às técnicas computacionais, a interpretação das análises por SAXS foi realçada, conforme resultados combinados com os modos normais de vibração da estrutura tetramérica. Neste caso, o vínculo experimental imposto pela curva $I(q)$ possibilitou que uma estrutura em solução fosse predita apenas com o uso de um único modo normal, estando relacionado às possíveis transições alostéricas de XfSurE.

A partir de um outro estudo descrito no Capítulo 3, foi possível verificar que embora o conteúdo informacional de uma curva unidimensional de SAXS tenha a capacidade de discriminar duas estruturas tridimensionais distintas, há uma dificuldade em fazê-lo quando os objetos são similares. Quanto mais próximas duas curvas $I(q)$, maior é a possibilidade do surgimento de falsos positivos. Se o objetivo for encontrar estruturas similares em um banco de dados diverso a partir

de um dado objeto-alvo, as chances de sucesso são proporcionais ao quão infrequente a forma deste objeto é para este banco de dados. Se a forma do objeto-alvo representar uma forma típica encontrada no banco, várias estruturas satisfarão o critério de busca com relativo sucesso.

Essa característica do reconhecimento de forma a partir dos dados de SAXS também explica os diversos mínimos locais encontrados na modelagem de corpo rígido para a proteína XfSurE. Naquele caso, pequenas distorções na estrutura produziram perfis de SAXS muito semelhantes, praticamente indistinguíveis, segundo a métrica utilizada. Logo, se o propósito for o reconhecimento do enovelamento de uma proteína usando dados de SAXS, informação adicional deve ser levada em conta. Isto é especialmente importante quando se deseja que a seleção de um modelo não somente se "ajuste" aos dados de SAXS, mas que ele também esteja relacionado evolutivamente com a proteína-alvo. No Capítulo 4, essa possibilidade foi explorada considerando a sequência de aminoácidos como "informação adicional". A abordagem combinada de SAXS e *threading* se mostrou melhor que o emprego da técnica *threading* tradicional, sem o uso de SAXS. Particularmente, foi demonstrado que o desempenho é melhor para proteínas que possuem formas menos frequentes no PDB, indo de encontro com os achados anteriores.

Estes resultados em conjunto ilustram abordagens que tem se tornado mais frequentes no campo da biologia computacional. Uma maior cooperação entre grupos experimentais e teóricos poderia alargar este campo de pesquisa e abrir novas frentes de trabalho. Existem excelentes iniciativas nesta direção, mas ainda são incipientes dada a potencialidade latente. Neste sentido, este trabalho de doutorado possui o mérito, mesmo que modesto, de ter ajudado a preecher uma lacuna pouco explorada na área da bioinformática estrutural.

Referências Bibliográficas

[1] A. Lehninger, D. L. Nelson, and M. M. Cox, *Lehninger Principles of Biochemistry*. W. H. Freeman, fifth edition ed., June 2008.

[2] A. K. Dunker, M. S. Cortese, P. Romero, L. M. Iakoucheva, and V. N. Uversky, "Flexible nets. The roles of intrinsic disorder in protein interaction networks.," *FEBS J*, vol. 272, pp. 5129–5148, Oct 2005.

[3] A. K. Dunker, C. J. Brown, J. D. Lawson, L. M. Iakoucheva, Z. Obradovic, and A. K. Dunker, "Intrinsic disorder and protein function," *Biochemistry*, vol. 41, pp. 6573–6582, 2002.

[4] L. Pauling and R. B. Corey, "The polypeptide-chain configuration in hemoglobin and other globular proteins.," *Proc Natl Acad Sci U S A*, vol. 37, pp. 282–285, May 1951.

[5] M. F. Perutz, "An optical method for finding the molecular orientation in different forms of crystalline haemoglobin; changes in dichroism accompanying oxygenation and reduction.," *Proc R Soc Lond B Biol Sci*, vol. 141, pp. 69–71, Mar 1953.

[6] M. G. Rossmann and D. M. Blow, "The detection of sub-units within the crystallographic asymmetric unit," *Acta Cryst.*, vol. 15, pp. 24–31, 1962.

[7] H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov, and P. Bourne, "The Protein Data Bank," *J. Mol. Biol.*, vol. 28, pp. 235–242, 2000.

[8] O. Glatter and O. Kratky, *Small Angle X-ray Scattering*. London: Academic Press Inc. Ltd., 1982.

[9] D. Svergun and M. H. J. Koch, "Small-angle scattering studies of biological macromolecules in solution," *Rep. Prog. Phys.*, vol. 66, pp. 1735–1782, 2003.

[10] C. D. Putnam, M. Hammel, G. L. Hura, and J. A. Tainer, "X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution.," *Q Rev Biophys*, vol. 40, pp. 191–285, Aug 2007.

[11] H. D. T. Mertens and D. I. Svergun, "Structural characterization of proteins and complexes using small-angle X-ray solution scattering.," *J Struct Biol*, vol. 172, pp. 128–141, Oct 2010.

[12] D. A. Jacques and J. Trewhella, "Small-angle scattering for structural biology–expanding the frontier while avoiding the pitfalls.," *Protein Sci*, vol. 19, pp. 642–657, Apr 2010.

[13] T. L. Blundell and L. Johnson, *Protein Crystallography*. Academic Press, 1976.

[14] R. Apweiler, A. Bairoch, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, D. A. Natale, C. O'Donovan, N. Redaschi, and L.-S. L. Yeh, "UniProt: the Universal Protein knowledgebase.," *Nucleic Acids Res*, vol. 32, pp. D115–D119, Jan 2004.

[15] D. J. Rigden, *From Protein Structure to Function with Bioinformatics*. Liverpool-UK: Springer, 2009.

[16] A. M. Saraiva, M. A. Reis, S. F. Tada, L. K. Rosselli-Murai, D. R. S. Schneider, A. C. Pelloso, M. A. S. Toledo, C. Giles, R. Aparicio, and A. P. de Souza, "Functional and small-angle X-ray scattering studies of a new stationary phase survival protein E (SurE) from Xylella fastidiosa-evidence of allosteric behaviour," *FEBS J.*, vol. 276, pp. 6751–6762, 2009.

[17] M. A. dos Reis, A. M. Saraiva, M. L. dos Santos, A. P. de Souza, and R. Aparicio, "Crystallization and preliminary X-ray analysis of stationary phase survival protein E (SurE) from Xylella fastidiosa in two crystal forms," *Acta Cryst.*, vol. F68, pp. 464–467, 2012.

[18] M. R. Lambais, M. H. Goldman, L. E. Camargo, and G. H. Goldman, "A genomic approach to the understanding of Xylella fastidiosa pathogenicity.," *Curr Opin Microbiol*, vol. 3, pp. 459–462, Oct 2000.

[19] M. R. Pooler and J. S. Hartung, "Specific PCR detection and identification of Xylella fastidiosa strains causing citrus variegated chlorosis.," *Curr Microbiol*, vol. 31, pp. 377–381, Dec 1995.

[20] A. H. Purcell and D. L. Hopkins, "Fastidious xylem-limited bacterial plant pathogens.," *Annu Rev Phytopathol*, vol. 34, pp. 131–151, 1996.

[21] A. J. Simpson and *et al*, "The genome sequence of the plant pathogen Xylella fastidiosa. The Xylella fastidiosa Consortium of the Organization for Nucleotide Sequencing and Analysis.," *Nature*, vol. 406, pp. 151–159, Jul 2000.

[22] C. Mura, J. E. Katz, S. G. Clarke, and D. Eisenberg, "Structure and function of an archaeal homolog of survival protein E (SurEalpha): an acid phosphatase with purine nucleotide specificity.," *J Mol Biol*, vol. 326, pp. 1559–1575, Mar 2003.

[23] J. Y. Lee, J. E. Kwak, J. Moon, S. H. Eom, E. C. Liong, J. D. Pedelacq, J. Berendzen, and S. W. Suh, "Crystal structure and functional analysis of the SurE protein identify a novel phosphatase family.," *Nat Struct Biol*, vol. 8, pp. 789–794, Sep 2001.

[24] R. G. Zhang, T. Skarina, J. E. Katz, S. Beasley, A. Khachatryan, S. Vyas, C. H. Arrowsmith, S. Clarke, A. Edwards, A. Joachimiak, and A. Savchenko, "Structure of Thermotoga maritima stationary phase survival protein SurE: a novel acid phosphatase.," *Structure*, vol. 9, pp. 1095–1106, Nov 2001.

[25] A. M. D. Gonçalves, A. T. Rêgo, M. Thomaz, F. J. Enguita, and M. A. Carrondo, "Expression, purification, crystallization and preliminary X-ray characterization of two crystal forms of stationary-phase survival E protein from Campylobacter jejuni.," *Acta Crystallogr Sect F Struct Biol Cryst Commun*, vol. 64, pp. 213–216, Mar 2008.

[26] A. Pappachan, H. S. Savithri, and M. R. N. Murthy, "Structural and functional studies on a mesophilic stationary phase survival protein (Sur E) from Salmonella typhimurium.," *FEBS J*, vol. 275, pp. 5855–5864, Dec 2008.

[27] W. Iwasaki and K. Miki, "Crystal structure of the stationary phase survival protein SurE with metal ion and AMP.," *J Mol Biol*, vol. 371, pp. 123–136, Aug 2007.

[28] M. Proudfoot, E. Kuznetsova, G. Brown, N. N. Rao, M. Kitagawa, H. Mori, A. Savchenko, and A. F. Yakunin, "General enzymatic screens identify three new nucleotidases in Escherichia coli. Biochemical characterization of SurE, YfbR, and YjjG.," *J Biol Chem*, vol. 279, pp. 54687–54694, Dec 2004.

[29] S. A. Hunsucker, B. S. Mitchell, and J. Spychala, "The 5'-nucleotidases as regulators of nucleotide and drug metabolism.," *Pharmacol Ther*, vol. 107, pp. 1–30, Jul 2005.

[30] D. Kern and E. R. P. Zuiderweg, "The role of dynamics in allosteric regulation.," *Curr Opin Struct Biol*, vol. 13, pp. 748–757, Dec 2003.

[31] K. Gunasekaran, B. Ma, and R. Nussinov, "Is allostery an intrinsic property of all dynamic proteins?," *Proteins*, vol. 57, pp. 433–443, Nov 2004.

[32] Q. Cui and M. Karplus, "Allostery and cooperativity revisited," *Protein Sci*, vol. 17, pp. 1295–1307, Aug 2008.

[33] J. Monod, J. Wyman, and J. P. Changeux, "On the nature of allosteric transitions: a plausible model," *J Mol Biol*, vol. 12, pp. 88–118, May 1965.

[34] D. I. Svergun, C. Barberato, M. H. Koch, L. Fetler, and P. Vachette, "Large differences are observed between the crystal and solution quaternary structures of allosteric aspartate transcarbamylase in the R state.," *Proteins*, vol. 27, pp. 110–117, Jan 1997.

[35] C. P. Macol, H. Tsuruta, B. Stec, and E. R. Kantrowitz, "Direct structural evidence for a concerted allosteric transition in Escherichia coli aspartate transcarbamoylase.," *Nat Struct Biol*, vol. 8, pp. 423–426, May 2001.

[36] R. Cabrera, H. Fischer, S. Trapani, A. F. Craievich, R. C. Garratt, V. Guixé, and J. Babul, "Domain motions and quaternary packing of phosphofructokinase-2 from escherichia coli studied by small angle x-ray scattering and homology modeling.," *J Biol Chem*, vol. 278, pp. 12913–12919, Apr 2003.

[37] T. Inobe, M. Arai, M. Nakao, K. Ito, K. Kamagata, T. Makio, Y. Amemiya, H. Kihara, and K. Kuwajima, "Equilibrium and kinetics of the allosteric transition of GroEL studied by solution X-ray scattering and fluorescence spectroscopy.," *J Mol Biol*, vol. 327, pp. 183–191, Mar 2003.

[38] M. Karplus and J. A. McCammon, "Molecular dynamics simulations of biomolecules.," *Nat Struct Biol*, vol. 9, pp. 646–652, Sep 2002.

[39] C. Tang, C. D. Schwieters, and G. M. Clore, "Open-to-closed transition in apo maltose-binding protein observed by paramagnetic NMR.," *Nature*, vol. 449, pp. 1078–1082, Oct 2007.

[40] R. Elber and A. West, "Atomically detailed simulation of the recovery stroke in myosin by milestoning.," *Proc Natl Acad Sci U S A*, vol. 107, pp. 5001–5005, Mar 2010.

[41] Tirion, "Large amplitude elastic motions in proteins from a single-parameter, atomic analysis.," *Phys Rev Lett*, vol. 77, pp. 1905–1908, Aug 1996.

[42] W. Zheng, B. R. Brooks, and D. Thirumalai, "Low-frequency normal modes that describe allosteric transitions in biological nanomachines are robust to sequence variations.," *Proc Natl Acad Sci U S A*, vol. 103, pp. 7664–7669, May 2006.

[43] E. C. Dykeman and O. F. Sankey, "Normal mode analysis and applications in biological physics.," *J Phys Condens Matter*, vol. 22, p. 423202, Oct 2010.

[44] A. M. Saraiva, *Caracterização Funcional e Estrutural da Nucleotidase SurE de Xylella fastidiosa*. Tese de doutorado, Universidade Estadual de Campinas - Instituto de Biologia, 2009.

[45] M. H. A. N. F. A P Hammersley, S O Svensson and D. Häusermann, "Two-dimensional detector software: From real detector to idealised image or two-theta scan," *High Pressure Research*, vol. 14, pp. 235–248, 1996.

[46] J. C. da Silva, *Estudos de macromoléculas biológicas parcialmente desestruturadas usando espalhamento de raios-X*. Tese de doutorado, Universidade Estadual de Campinas - Instituto de Física *Gleb Wataghin*, 2010.

[47] D. Svergun, "Determination of the regularization parameter in indirect-transform methods using perceptual criteria," *Journal of Applied Crystallography*, vol. 25, pp. 495–503, Aug 1992.

[48] E. Mylonas and D. Svergun, "Accuracy of molecular mass determination of proteins in solution by small-angle X-ray scattering," *J Appl Crystallogr*, vol. 40, pp. 245–249, 2007.

[49] K. M. H. J. Svergun D., Barberato C., "CRYSOL - A program to evaluate x-ray solution scattering of biological macromolecules from atomic coordinates," *Journal of Applied Crystallography*, vol. 28, pp. 768–773, Dec 1995.

[50] M. A. dos Reis, R. Aparicio, and Y. Zhang, "Improving protein template recognition by using small-angle x-ray scattering profiles.," *Biophys. J.*, vol. 101, pp. 2770–2781, Dec 2011.

[51] K. Stovgaard, C. Andreetta, J. Ferkinghoff-Borg, and T. Hamelryck, "Calculation of accurate small angle X-ray scattering curves from coarse-grained protein models.," *BMC Bioinformatics*, vol. 11, p. 429, 2010.

[52] S. DI, "Restoring low resolution structure of biological macromolecules from solution scattering using simulated annealing," *Biophysical Journal*, vol. 76, pp. 2879–2886, Jun 1999.

[53] D. Franke and D. Svergun, "DAMMIF, a program for rapid ab-initio shape determination in small-angle scattering," *J. Appl. Cryst.*, vol. 42, pp. 342–346, 2009.

[54] S. D. Volkov VV, "Uniqueness of ab initio shape determination in small-angle scattering," *J. Appl. Crystallogr.*, vol. 36, pp. 860–864, Jun 2003.

[55] "High-throughput structure determination. proceedings of the 2002 ccp4 (collaborative computational project in macromolecular crystallography) study weekend. january, 2002. york, united kingdom.," *Acta Crystallogr D Biol Crystallogr*, vol. 58, pp. 1897–1970, Nov 2002.

[56] A. J. C. Wilson, ed., *International Tables for Crystallography*. Dordrecht/Boston/London: Kluwer Academic Publishers, 1992.

[57] T. Bergfors, *Protein crystallization: techniques, strategies, and tips: a laboratory manual*. Intl Univ Line, 1999.

[58] E. F. Garman and R. L. Owen, “Cryocooling and radiation damage in macromolecular crystallography.,” *Acta Crystallogr D Biol Crystallogr*, vol. 62, pp. 32–47, Jan 2006.

[59] F. Long, A. A. Vagin, P. Young, and G. N. Murshudov, “BALBES: a molecular-replacement pipeline.,” *Acta Crystallogr D Biol Crystallogr*, vol. 64, pp. 125–132, Jan 2008.

[60] F. Tama and Y. H. Sanejouand, “Conformational change of proteins arising from normal mode calculations.,” *Protein Eng*, vol. 14, pp. 1–6, Jan 2001.

[61] K. Suhre and Y.-H. Sanejouand, “ElNemo: a normal mode web server for protein movement analysis and the generation of templates for molecular replacement.,” *Nucleic Acids Res*, vol. 32, pp. W610–W614, Jul 2004.

[62] D. S. Goodsell and A. J. Olson, “Structural symmetry and protein function,” *Annual Review of Biophysics and Biomolecular Structure*, vol. 29, pp. 105–153, 2000.

[63] M. B. Kozin and D. Svergun, “Automated matching of high- and low-resolution structural models,” *J Appl Crystallogr*, vol. 34, pp. 33–41, 2001.

[64] H. Hartmann, B. Lohkamp, N. Hellmann, and H. Decker, “The allosteric effector l-lactate induces a conformational change of 2x6-meric lobster hemocyanin in the oxy state as revealed by small angle x-ray scattering.,” *J Biol Chem*, vol. 276, pp. 19954–19958, Jun 2001.

[65] L. M. Rice, E. A. Montabana, and D. A. Agard, “The lattice as allosteric effector: structural studies of alphabeta- and gamma-tubulin clarify the role of GTP in microtubule assembly.,” *Proc Natl Acad Sci U S A*, vol. 105, pp. 5378–5383, Apr 2008.

[66] J. A. Rivas-Pardo, A. Herrera-Morande, V. Castro-Fernandez, F. J. Fernandez, M. C. Vega, and V. Guixé, “Crystal structure, SAXS and kinetic mechanism of hyperthermophilic ADP-dependent glucokinase from Thermococcus litoralis reveal a conserved mechanism for catalysis.,” *PLoS One*, vol. 8, no. 6, p. e66687, 2013.

[67] A. T. P. Machado, *Determinação e refinamento de estruturas tridimensionais da proteína SurE de Xylella fastidiosa*. Dissertação de mestrado, Universidade Estadual de Ponta Grossa, 2013.

[68] Y. Zhang and J. Skolnick, “Scoring function for automated assessment of protein structure template quality.,” *Proteins*, vol. 57, pp. 702–710, Dec 2004.

[69] E. Krissinel and K. Henrick, “Inference of macromolecular assemblies from crystalline state.,” *J Mol Biol*, vol. 372, pp. 774–797, Sep 2007.

[70] W. Zheng and M. Tekpinar, “Accurate flexible fitting of high-resolution protein structures to small-angle x-ray scattering data using a coarse-grained model with implicit hydration shell.,” *Biophys J*, vol. 101, pp. 2981–2991, Dec 2011.

[71] W. Wriggers, “Using Situs for the integration of multi-resolution structures.,” *Biophys Rev*, vol. 2, pp. 21–27, Feb 2010.

[72] W. Wriggers and P. Chacón, “Using Situs for the Registration of Protein Structures with Low-Resolution Bead Models from X-ray Solution Scattering,” *J. Appl. Cryst.*, vol. 34, pp. 773–776, 2001.

[73] R. Osada, T. Funkhouser, B. Chazelle, and D. Dobkin, “Shape distributions,” *ACM T. Graphic.*, vol. 21, pp. 807–832, 2002.

[74] R. Bergmann, A. Linusson, and I. Zamora, “SHOP: Scaffold HOPping by GRID-Based Similarity Searches,” *J. Med. Chem.*, vol. 50, pp. 2708–2717, 2007.

[75] D. Shortle, “Structure prediction: Folding proteins by pattern recognition,” *Current Biology*, vol. 7, pp. 151–154, 1997.

[76] P. Shilane, P. Min, M. Kazhdan, and T. Funkhouser, “The princeton shape benchmark: http://shape.cs.princeton.edu/benchmark/,” in *Shape Modeling International*, (Genova), 2004.

[77] C. D. Manning, P. Raghavan, and H. Schütze, *An Introduction to Information Retrieval*. Cambridge: Cambridge University Press, 2009.

[78] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, “SCOP: a structural classification of proteins database for the investigation of sequences and structures.,” *J Mol Biol*, vol. 247, pp. 536–540, Apr 1995.

[79] M. Y. Lobanov, N. S. Bogatyreva, and O. V. Galzitskaya, "Radius of gyration as an indicator of protein structure compactness," *Molecular Biology*, vol. 42, pp. 623–628, 2008.

[80] D. N. Ivankov, N. S. Bogatyreva, M. Y. Lobanov, and O. V. Galzitskaya, "Coupling between properties of the protein shape and the rate of protein folding," *Plos One*, vol. 4, pp. 1–10, 2009.

[81] J. Kopp, L. Bordoli, J. N. D. Battey, F. Kiefer, and T. Schwede, "Assessment of CASP7 predictions for template-based modeling targets.," *Proteins*, vol. 69 Suppl 8, pp. 38–56, 2007.

[82] D. Cozzetto, A. Kryshtafovych, K. Fidelis, J. Moult, B. Rost, and A. Tramontano, "Evaluation of template-based models in CASP8 with standard measures.," *Proteins*, vol. 77 Suppl 9, pp. 18–28, 2009.

[83] J. U. Bowie, R. Lüthy, and D. Eisenberg, "A method to identify protein sequences that fold into a known three-dimensional structure.," *Science*, vol. 253, pp. 164–170, Jul 1991.

[84] D. T. Jones, W. R. Taylor, and J. M. Thornton, "A new approach to protein fold recognition.," *Nature*, vol. 358, pp. 86–89, Jul 1992.

[85] F. DiMaio, T. C. Terwilliger, R. J. Read, A. Wlodawer, G. Oberdorfer, U. Wagner, E. Valkov, A. Alon, D. Fass, H. L. Axelrod, D. Das, S. M. Vorobiev, H. Iwaï, P. R. Pokkuluri, and D. Baker, "Improved molecular replacement by density- and energy-guided protein structure optimization.," *Nature*, vol. 473, pp. 540–543, May 2011.

[86] S. Raman, O. F. Lange, P. Rossi, M. Tyka, X. Wang, J. Aramini, G. Liu, T. A. Ramelot, A. Eletsky, T. Szyperski, M. A. Kennedy, J. Prestegard, G. T. Montelione, and D. Baker, "NMR structure determination for larger proteins using backbone-only data.," *Science*, vol. 327, pp. 1014–1018, Feb 2010.

[87] W. Li, Y. Zhang, and J. Skolnick, "Application of sparse NMR restraints to large-scale protein structure prediction.," *Biophys J*, vol. 87, pp. 1241–1248, Aug 2004.

[88] Y. Xu, D. Xu, O. H. Crawford, and J. R. Einstein, "A computational method for NMR-constrained protein threading.," *J Comput Biol*, vol. 7, no. 3-4, pp. 449–467, 2000.

[89] W. Zheng and S. Doniach, "Protein structure prediction constrained by solution X-ray scattering data and structural homology identification.," *J Mol Biol*, vol. 316, pp. 173–187, Feb 2002.

[90] W. Zheng and S. Doniach, "Fold recognition aided by constraints from small angle X-ray scattering data.," *Protein Eng Des Sel*, vol. 18, pp. 209–219, May 2005.

[91] S. Wu and Y. Zhang, "MUSTER: Improving protein sequence profile-profile alignments by using multiple sources of structure information," *Proteins: Structure, Function, and Bioinformatics*, vol. 72, pp. 547–556, 2008.

[92] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool.," *J Mol Biol*, vol. 215, pp. 403–410, Oct 1990.

[93] D. T. Jones, "Protein secondary structure prediction based on position-specific scoring matrices.," *J Mol Biol*, vol. 292, pp. 195–202, Sep 1999.

[94] D. Frishman and P. Argos, "Knowledge-based protein secondary structure assignment.," *Proteins*, vol. 23, pp. 566–579, Dec 1995.

[95] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers*, vol. 2, pp. 2577–2637, 1983.

[96] S. Wu and Y. Zhang, "ANGLOR: a composite machine-learning algorithm for protein backbone torsion angle prediction.," *PLoS One*, vol. 3, no. 10, p. e3400, 2008.

[97] P. J. Silva, "Assessing the reliability of sequence similarities detected through hydrophobic cluster analysis.," *Proteins*, vol. 70, pp. 1588–1594, Mar 2008.

[98] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins.," *J Mol Biol*, vol. 48, pp. 443–453, Mar 1970.

[99] A. Sali and T. L. Blundell, "Comparative protein modelling by satisfaction of spatial restraints.," *J Mol Biol*, vol. 234, pp. 779–815, Dec 1993.

[100] Y. Zhang, "Template-based modeling and free modeling by I-TASSER in CASP7.," *Proteins*, vol. 69 Suppl 8, pp. 108–117, 2007.

[101] A. Roy, A. Kucukural, and Y. Zhang, "I-TASSER: a unified platform for automated protein structure and function prediction.," *Nat Protoc*, vol. 5, pp. 725–738, Apr 2010.

[102] P. Debye *Ann. Physik*, vol. 46, pp. 809–823, 1915.

[103] R. D. B. Fraser, T. P. Macrae, and E. Suzuki, "An improved method for calculating the contribution of solvent to X-ray diffraction pattern of biological molecules," *J. Appl. Cryst.*, vol. 11, pp. 693–694, 1978.

[104] D. I. Svergun, S. Richard, M. H. Koch, Z. Sayers, S. Kuprin, and G. Zaccai, "Protein hydration in solution: experimental observation by x-ray and neutron scattering.," *Proc Natl Acad Sci U S A*, vol. 95, pp. 2267–2272, Mar 1998.

[105] M. V. Petoukhov and D. I. Svergun, "New methods for domain structure determination of proteins from solution scattering data," *J. Appl. Cryst.*, vol. 36, pp. 540–544, 2003.

[106] S. Yang, S. Park, L. Makowski, and B. Roux, "A rapid coarse residue-based computational method for X-ray solution scattering characterization of protein folds and multiple conformational states of large protein complexes.," *Biophys J*, vol. 96, pp. 4449–4463, Jun 2009.

[107] G. Wang and R. Dunbrack, "PISCES: a protein sequence culling server," *Bioinformatics*, vol. 19, pp. 1589–1591, 2003.

[108] G. L. Hura, A. L. Menon, M. Hammel, R. P. Rambo, F. L. Poole, S. E. Tsutakawa, F. E. Jenney, S. Classen, K. A. Frankel, R. C. Hopkins, S.-J. Yang, J. W. Scott, B. D. Dillard, M. W. W. Adams, and J. A. Tainer, "Robust, high-throughput solution structural analyses by small angle X-ray scattering (SAXS).," *Nat Methods*, vol. 6, pp. 606–612, Aug 2009.

[109] F. Förster, B. Webb, K. A. Krukenberg, H. Tsuruta, D. A. Agard, and A. Sali, "Integration of small-angle X-ray scattering data into structural modeling of proteins and their assemblies.," *J Mol Biol*, vol. 382, pp. 1089–1106, Oct 2008.

[110] D. Schneidman-Duhovny, M. Hammel, and A. Sali, "FoXS: a web server for rapid computation and fitting of SAXS profiles.," *Nucleic Acids Res*, vol. 38, pp. W540–W544, Jul 2010.

[111] A. V. Sokolova, V. V. Volkova, and D. I. Svergun, "Prototype of a database for rapid protein classification based on solution scattering data," *J. Appl. Cryst.*, vol. 36, pp. 865–868, 2003.

[112] E. S. Huang, S. Subbiah, and M. Levitt, "Recognizing native folds by the arrangement of hydrophobic and polar residues.," *J Mol Biol*, vol. 252, pp. 709–720, Oct 1995.

[113] K. T. Simons, I. Ruczinski, C. Kooperberg, B. A. Fox, C. Bystroff, and D. Baker, "Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins.," *Proteins*, vol. 34, pp. 82–95, Jan 1999.